What Can We Learn from Science's Steroid Era? A Proposal to Reinterpret Fifty Years of Statistical Significance Claims

Nicolás Idrobo[*] Arthur Lupia [†] Hwayong Shin [‡] Rocío Titiunik[§]

August 23, 2018

Preliminary and Incomplete—Do Not Cite or Circulate

[*] Ph.D. student, Department of Economics, University of Michigan.

[†] Hal R. Varian Collegiate Professor, Department of Political Science, University of Michigan.

[‡] Ph.D. student, Department of Political Science, University of Michigan.

[§] James Orin Murfin Professor, Department of Political Science, University of Michigan.

Abstract

In baseball, home runs are seen as important human achievements. When a game includes a home run, the event influences how the game is remembered. Moreover, baseball's history is regularly rewritten to highlight these accomplishments.

In a scientific journal article, statistically significant findings are seen as important human achievements. When an article includes a statistically significant finding, the event influences how the article is remembered. Moreover, scholarly literatures are regularly rewritten to highlight these accomplishments.

In the late 20th century, the number of home runs—and significance claims—increased at unprecedented rates. Initially, many observers interpreted these increases as reflecting improved human achievement. Later, we learned that other, hidden factors played a role.

In baseball, the hidden factor was anabolic steroids. The sport reacted by banning steroid users from the game and by keeping the era's leading home run hitters out of its hall of fame.

In science, the hidden factor was the combination of a massive increase in computing power, which allowed researchers to run many statistical analyses, and professional incentives that equated worthy science with "statistical significance." Given what we now know about the process by which many significance claims were produced, we contend that science can react constructively. To this end, we introduce a new framework that can help scholars and other readers more accurately interpret empirical claims from science's "steroid era."

The framework has four main characteristics. First, it acknowledges the need to carefully interpret past findings. Second, it expands the discussion of reliable science beyond replicability, which increases its relevance to non-experimental studies. Third, it distinguishes problems caused by model selection from problems caused by model misspecification. Fourth, it proposes a way to interpret significance claims produced under rapidly evolving levels of computing power.

1 Introduction

Scientific discoveries influence many decisions. These discoveries help people correct errors and use information more effectively. Their cumulative impact is worldwide. Science has changed how we live. For these reasons, science is a genuine human achievement.

What properties give science such power? An important factor is that science is built from processes that evaluate causal, correlational, and existential conjectures with respect to replicable logic and curated evidence. Science allows us to derive simplicity from complexity, while sometimes revealing that what appears simple is actually complex.

Science has evolved over time. One notable aspect of this evolution is the role of statistics. In the last fifty years, many scientists have taken a greater interest in statistics. This interest has transformed many disciplines. It has changed scholarly norms about what kinds of data and supporting materials are considered sufficient to validate a knowledge claim.

In the early part of this fifty-year period, statistical claims often took the form of descriptive correlations. Subsequent generations of scholars emphasized multivariate modeling. In this later era, many key findings took the form of statistically significant coefficients. Some statistically significant claims gained iconic status. Producers of these claims often became leaders in their fields.

At the same time that scientists became more interested in statistics, a related phenomenon was occurring. Throughout this fifty-year period, scientists gained access to an exponential increase in computing power. This change in access affected the types of statistical analyses that a scholar could produce.

In the early 1970's, for example, a researcher who wanted to produce results from a single multivariate model had to move boxes of punch cards from her office to the computer center. She then had to wait several hours for the results—during which time she would have had time to pray that relatively minor punch card errors had not caused the process to fail. The 1970s researcher could not afford the luxury of fitting many models. She had to specify variables and models a-priori, so that she could communicate those to the technician at the computer center who would execute the run.

By the late 1990's, the situation was quite different. Increasingly sophisticated and efficient software for data management and statistical analysis proliferated. An empirical researcher in the late 1990s could analyze many statistical models in a few seconds by pushing a few buttons on the personal computer in her office. She could see results very quickly. If the 1990's researcher saw a result contrary to her expectations, she had a number of tools at her disposal to explore other analyses that might produce a different outcome. She could quickly explore variations of the model that omitted some variables, added others, squared some, logged others, and interacted various variable combinations. The 1990s could also easily "clean" the data on the spot by rescaling outcomes, excluding outliers, dropping missing values. For the 1990s researcher, the final model was found via interactive trial-and-error, where many of the decisions in the process were taken after seeing auxiliary statistical inferences that revealed whether the candidate specification produced statistically significant claims and/or effects in the "right" direction. In sum, the 1990s researcher had many opportunities that the 1970s researcher did not.

The ability to more flexibly interact with data undoubtedly alleviated many of the burdens that made empirical research in the early 1970s slow and rigid. Close interaction with the data often reveals obvious patterns that directly inform sensible decisions at the analysis stage. This greater dynamism and flexibility, however, comes with hidden costs. Standard results taught in graduate schools hold when only one or few hypotheses are tested, and when the statistical model is chosen independently of the data that will be used for final analysis. When these practices are violated (e.g., if scholars use an interactive trial-and-error process to find specifications that produce certain types of outcomes), they may inadvertently sacrifice the properties of statistical inferences that made them a source of scientific legitimacy in the first place.

We assume that when these violations occurred, they were not the result of widespread intentional dishonesty or malevolence. Rather, they were an understandable reaction to the context of the time. Journal editors, article reviewers, and readers were part of a vast academic advancement ecosystem that assigned higher value to statistically significant claims than they did to null results or descriptive analysis. As a result, scholars had strong professional incentives to produce these types of claims. The combination of these incentives with the technological revolution resulted in the mass production of statistical inferences, in particular of statistically significant findings.

As the more statistical significance claims appeared, a scholarly vanguard raised important questions about how to interpret them. These concerns manifested in many ways. They included questions about robustness, reproducibility, and replication. They include the scholarly movement commonly known as "the causal revolution." One result of these efforts is a growing consensus that many published significance claims represent "false positives" (Ioannidis, 2005).

Which leads to the question—how should we think of claims made in the era of cheap and powerful statistical estimation programs? Can we now claim to "know" what scholarly leaders of previous eras claimed to know? Are this era's iconic findings worth teaching to students? Are they valid enough to convey to the public and to policymakers? Or should this work be ignored altogether?

In this paper, we offer a template for reinterpreting previous generations' empirical work. Our goal is to help readers draw more accurate understandings of what that era's work did—and did not—teach us. To clarify how we are approaching this problem, we first develop an analogy. We then use the analogy to motivate a framework for more accurately interpreting the previous era's claim. The analogy draws a parallel between the "steroid era" in baseball and the era in which scientists had increasing opportunity and motive to focus on statistically significant findings.

In baseball, the "steroid era" describes a period in the late 20th century where players hit an unprecedented number of home runs. This surge in baseball's most focal accomplishment increased interest in the sport and created a new generation of "stars."

Observers of the sport offered various theories of the increase in home runs. Some pointed to better training. Others inquired about changes in the physical properties of the ball (a.k.a., the "juiced ball" theory). As the surge continued, others suspected that a hidden activity was fueling the surge. There were growing suspicions that the accomplishments represented not true human accomplishment, but the products of hidden factors that violated conventional understandings of how home runs were produced. As it later turned out, the surge was caused by undisclosed steroids use.

Baseball has now revised its view of the "steroid era's" alleged accomplishments. It banned many of the era's stars from playing the game. It has kept others from receiving career honors, such as induction into the "hall of fame." It has instituted a series of severe penalties for people who engage in the practice today. The baseball establishment, and many fans, now treats formerly iconic accomplishments as unworthy of recognition or celebration. Baseball now sees the "steroid era" as a dark mark. Home runs, while now fewer in number, are again regarded as products of genuine human accomplishment.

We contend that the practice of informal model selection through interactive trial-and-error, and the mass production of statistical significance claims that resulted, play a role analogous to the role of anabolic steroids in baseball. In decades of published papers, these practices were not explicitly discussed or reported. Instead, the empirical findings were often described as the product of a single empirical model and single analysis. Because statistical inferences that emerge from interactive trial-and-error do not have the same properties as statistical inferences performed on an a-priori model, significance claims that emerge from this process need not have the same properties and meaning as significance claims that do not. One important consequence is that some or many of the era's findings may be false positives and will not be replicated, and thus the knowledge that is built on them is questionable.

We offer a framework for reinterpreting empirical claims that were produced in science's "steroid era". Our argument is related to prior arguments in the transparency and replication literature, but makes three distinct contributions.

First, we link the properties of statistical inferences to the particular era in which they were produced, showing that the interpretation of scientific findings likely depends on the technologies and professional incentives of each era. We explicitly connect the computing revolution, the easy availability of data, and the strong professional incentives that moved scientific research away from descriptive and towards inferential analyses, to the paradigm of statistical inference that must be used for interpretation. As a consequence, our argument implies that earlier scientific findings of certain types are more likely to be replicated than more modern findings of comparable types.

Second, our emphasis on the interactive process of trial-and-error distinguishes between a process of trying multiple tests where the number of tests ran is known, from a process of informal model selection where number of tests is not revealed to reviewers or readers of the claimed results. This distinction between multiple testing and model selection has profound consequences for our ability to interpret previous results and our ability to fix problems where they occur. One implication of our work is that there are some claims where we can use knowledge of researcher incentives and computing power to produce a more accurate interpretation of the finding they produce. At the same time, we can characterize a set of claims that are now impossible to interpret—and impossible to fix. Looking forward, more comprehensive disclosure of the number of tests runs and the theoretical foundations of model selection processes can help readers and reviewers interpret statistical claims more accurately.

Third, our argument focuses on all scientific studies, not only randomized experiments, as we distinguish model selection from model misspecification. This emphasis helps us make predictions about which findings are more likely to be replicated. It also clarifies conditions under which replication of a scientific finding reveals anything about whether the finding is true. As many prior studies on replication and transparency focus on randomized controlled trials, we believe our approach can help produce accurate interpretations of a broader set of scientific claims.

This version of the paper offers empirical evidence consistent with our argument. An analysis of 200 published papers using the American National Elections Studies (ANES) published between 1970 and 2017 shows that the number of statistically significant claims has increased dramatically, just as the number of non-inferential, descriptive analyses has dropped. In future versions of this manuscript, we will include a replication of 50 studies—10 studies per decade in the period 1970-2010—to establish whether, as we predict, the replicability of prior findings varies by era.

The paper proceeds as follows. In Section 2, we offer a theory of how changes in computing power and costs affected scholarly abilities and incentives to produce significance claims, distinguishing between the early and the modern era. In Section 3, we present a statistical example to discuss the properties of statistical inferences in various scenarios, which correspond to the eras discussed in Section 2. In Section 4, we present the analysis of ANES papers between 1970 and 2017 to evaluate our theory's predictions in terms of patterns of frequencies and types of particular statistical results. Section 5 includes a discussion of how our methods correspond to those taken by baseball's effort to restore its own integrity. Section 6 concludes. A technical appendix contains supporting information.

2 The Different Eras in the Production of Scientific Knowledge

Computing power has increased exponentially in the last five decades. Moore's law, according to which the number of transistors per microprocessor chip roughly doubles every two years, has been an empirical regularity since the 1970s. Intel's first microprocessor in 1971 had 2,300 transistors. By 2010, the typical transistor count was in the order of the hundreds of millions. Clock speeds followed roughly the same rates of increase until recently, and every one of the last fives decades brought a smaller and more portable class of computers (Waldrop, 2016).

This radical increase in computing power was accompanied by an exponential decrease in computer prices. This drop in prices made affordable, powerful personal computers and smart mobile devices an essential feature of modern society. It transformed most aspects of human life. Areas as disparate as travel, communications, defense systems, medical treatments, movies, and music, have been radically altered by this technological evolution.

The production of scientific knowledge was also affected. For example, these advances in technology made it possible for researchers to conduct increasingly complex, multifaceted statistical analyses. In turn, these advances corresponded with increasing interest in, and use of, statistical inference procedures in many academic disciplines.

2.1 The early era: punch cards, computer centers, and description

In the early 1970s, collecting and analyzing data was extremely costly. In this period, there were neither personal computers nor internet connections. A researcher seeking to do sophisticated statistical analysis saw significant obstacles at both the data collection and data analysis stages. An example of the costs of data collection is provided by Kurland and Molgaard (1981), who describe the process needed to access the medical records of the Mayo Clinic in 1970s. In order to gain access to the data, the investigator had to request the physical transportation of the medical records from central storage to the medical-statistics unit (using a dumbwaiter), provide a form listing the specific information needed from the records, and wait until trained "abstracters" recorded the requested data from each patient dossier.

In addition to the obstacles involved in accessing data, fitting a statistical model required making the data "computer-readable." This involved key punching the data onto computer cards, and feeding these cards into a rudimentary computer. Only then could a researcher proceed with the analysis. Högman and Ramgren (1970) describe a complicated system of colored punch cards specifically designed to produce blood-registry data in a computer-readable format. Their goal was to computerize the centralization of blood transfusion services in Sweeden. The cost of the computer routines was about $40,000 a year (equivalent to roughly $266,000 today adjusting for inflation), and this was to run the computer only once a week for three hours for the purposes of scanning the blood-donor registry. Other analyses such as accounting routines were produced once a month. A blood-type list was compiled twice a year. The computer center produced a punch card for each donor, which had to be returned to the blood bank for updating; the distance between the computer center and the blood banks could be up to 370 miles.

Given the obstacles and costs associated with statistical analyses, researchers had to define the variables to be included, the subgroups to be analyzed, the observations to discard, etc., before looking at the data. The technological conditions of this era were thus largely consistent with a paradigm of statistical inference where models are specified a priori and often independent of direct contact with the data. The analysis was performed in a computer center, physically removed from the researcher's office and from the data source. This way of conducting statistical analyses fit well with the publication standards of this era. In this era, publications often focused on descriptive evidence and did not require empirical studies to include statistical inferences. In fact, as we document in Section 4, in the sample we analyze, the proportion of tables devoted to descriptive (i.e., non-inferential) results has declined fourfold since the 1970s.

2.2 The modern era: terabytes, personal super-computers, and inferences

As decades progressed, this reality changed dramatically. Today, data availability and computer power are immense. Accessing data no longer requires boxes of punch cards, dumbwaiters, and abstracters. Datasets are now digital, stored in hard drives and shared among researchers via internet connections and shared cloud services synchronized in real time with personal computers. Even texts can now be stored electronically as matrices of words and processed as numerical outcomes. Moreover, powerful computing now allows researchers to perform hundreds or thousands of complex statistical analysis in seconds or minutes using millions of observations on their personal laptops, without leaving their offices. This would have sounded as science fiction to the technician feeding punch cards at the 1970s computer center.

The evolution of big computing opened exciting new possibilities for scientific progress. It multiplied opportunities for measurement, discovery, and exploration. These opportunities transformed how empirical scientific findings were produced. But the transformation was not equal in all sciences. In the biomedical sciences, the dimensionality of data grew rapidly due to advances in genomics and related fields. At the same time, external guidelines (such as from industry and the Federal Drug Administration) regulated many kinds of scholarship. Exploratory practices became more formal. Statistical procedures evolved in accordance with these guidelines and practices. In other sciences, data dimensionality was initially lower, external standards were less prominent, and exploratory practices were less formal. Perspectives on appropriate statistical practices evolved accordingly.

As an example, consider microarrays in genomics. In a microarray study, DNA or RNA abundance is measured for thousands of genes simultaneously. The goal is to identify which of those genes have expression levels associated with a particular disease (Dudoit et al., 2003). To achieve this goal, the researcher tests, for each gene, the null hypothesis that the gene's expression level is not associated with the outcome of interest. The typical setup thus involves testing hundreds or thousands of hypotheses, depending on how many genes are explored. This setup is the canonical example of what Efron (2010) calls "large-scale" inference.

It is well known that when researchers make hundreds or thousands of tests, the classical statistical setup fo testing a single a-priori null hypothesis is not appropriate. A central concern is making false discoveries—concluding that an association exists when in fact it does not. Imagine that a researcher tests 1,000 null hypotheses, and unbeknown to her, all of them are true. Nonetheless, if she uses a statistical test where the (marginal) probability of rejecting each individual null hypothesis is at most 5%, she will “discover” 50 exciting associations, none of which will be real.

As we discuss below, there are very well developed tools to control the probability of false discoveries or rejections in this case. These tools address the challenges posed by large-scale inferences, and allow researchers to explore in a statistically safe way. All of these methods, however, require the researcher to know the total number of tests performed, and use this number for adjustment. Without information on the total number of tests, controlling the probability of making false discoveries is impossible.

In medical and biological sciences, many researchers are well aware of the challenges of drawing sound conclusions from a multiplicity of tests. Best practices entail adjusting their inferences accordingly. Because in these sciences the exploration step is often explicit, the total number of tests performed (e.g. the total number of genes explored) is expected to be stated in advance, or at the very least used for adjustment ex-post.

For example, Food and Drug Administration (FDA) guidelines for clinical trials of human drugs address this problem explicitly. They discuss various adjustments for multiplicity to help researchers avoid drawing false conclusions about drug effectiveness. In these guidelines, it is assumed that the total number of tests performed is pre-specified before the trial even starts (and thus before the data to be used for analysis even exists).

Common practices in empirical social sciences are different. In political science, economics, and related disciplines, it is standard for researchers to engage in a much more informal exploratory process before settling into a final model. This is particularly true for researchers who work with nonexperimental or observational designs, for which pre-registration is difficult or impossible. It is also true for scholars who analyze experimental designs but do not present a pre-analysis plan. In both contexts, it is common for researchers to make many decisions, more than a few of which are made after seeing the results of previous analyses of the same data. These decisions include subsetting, removing outliers, including fixed effects, looking at subgroups or interactions, transforming the outcome, including covariates, and more. The result is a collection of analyses, including hypothesis tests, confidence intervals, etc., that are not planned or stated in advance.

This kind of exploratory analysis would fit directly into the large-scale inference framework if researchers kept track of the number of statistical tests performed. Quite often, however, the exploration stage is quite informal, and researchers do not report the total number of tests carried out from the beginning of a project until the work is complete. In fact, we suspect that many researchers would not even know how many tests were run. We do not believe this practice arises from an intention to cheat, but is rather the consequence of the combination of various factors governing the academic advancement ecosystem. Simmons et al. (2011) cite several factors behind the ubiquitousness of this practice. They include the human tendency for confirmatory bias and strong professional incentives to produce research findings that show nonzero effects.

In the last five decades, professional incentives have changed. Many scientific communities place most of an article's value on whether it shows "effects". Statistical significance claims became particularly influential. Hence, many ambitious scholars wrote papers that were centered around rejecting null hypotheses. Significance claims became increasingly valuable and academic advancement ecosystems evolved to reward their production.

The simultaneous evolution of these strong professional incentives and the computing revolution had direct consequences for the way in which scientific knowledge was produced and interpreted. In sciences where exploration was explicit and documented, inferences were adjusted in ways that facilitated accurate interpretation. In those sciences where exploration stayed informal, key properties of inferences more difficult to interpret—and may, in some cases, be unknowable.

When empirical research includes multiple tests and analyses during a process of interactive trial and error, when only one or few of these is reported, and when the total number of tests conducted is either unreported or unknown, the problem of multiple testing becomes a problem of model selection. Developing procedures for valid inference in these circumstances is very challenging. Depending on the procedure used for model selection, and how much information about selection step is shared or known, the challenges can be so severe as to invalidate statistical inferences altogether. We fear that a non-negligible proportion of findings in the social sciences fall in the latter category.

In the next section, we present an example to illustrate these challenges, and the distinction between the various concepts of multiple testing, model selection, replication and misspecification.

3 Statistical Inferences in the Early vs. Current Eras[1]

We follow the example in Leeb et al. (2015), and consider a model of n observations following a standard Normal distribution with mean $\mu \in \mathbb{R}^n$

$$y = \mu + \varepsilon, \tag{3.1}$$

where $\varepsilon \sim N(0, I_n)$. For simplicity, we assume that $\sigma^2 = 1$. We have p available covariates or explanatory variables, collected in the $n \times p$ matrix X.

We consider a family $\mathcal{M}$ of full column rank models, where in each model y is regressed on a subset of covariates in X. Each model is described by the set of indexes $M = \left\{j_1, \ldots, j_{|M|}\right\} \subseteq \{1, 2, \ldots, p\}$. Letting X_j refer to the jth column of X, $X_M = (X_{j_1}, \ldots, X_{j_{|M|}})$ denotes the columns of X that are included in model M (i.e., the columns of X whose indices lie in M), where $|M|$ is the size of M. The setup assumes that M is nonempty, and that all models considered are full rank.

We consider different scenarios, according to whether the researcher chooses the model prior to analyzing the data that will be used for estimation and inference, or she engages in a more exploratory analysis, where the relationships of interest will be, in part, dictated by the data. Each one of this scenarios corresponds to a stylized representation of the eras we discussed in Section 2.

Scenario 1: Single test with a priori model

In the first scenario, only one model is analyzed, and this model is selected before exploring the data that will be used to perform estimation and inference. This corresponds to the 1970s era, where models were often decided a-priori and interactive trial-and-error was prohibitively costly.

The researcher chooses the model M from the family of possible models $\mathcal{M}$:

$$y = X_M \beta_M + v_M \tag{3.2}$$

In 3.2, the parameter of interest is the population coefficient $\beta_M = (X_M' X_M)^{-1} X_M' \mu$.

The researches calculates the least squares estimator $\widehat{\beta}_M = (X_M' X_M)^{-1} X_M' y$.

[1] This section is still in progress. The notation, models and discussion will be modified

We let β_{jM} refer to the regression coefficient associated with covariate X_j in model M. In particular, we assume the researcher is interested in the partial average effect of X_1 on y; thus, the coefficient of interest is β_{1M}. Because in this scenario the researcher considers only one model, which is fixed and determined a-priori, we drop M from the notation.

The researcher tests the null hypothesis $H_0 : \beta_1 = 0$ using the test statistic

$$T = \widehat{\beta}_1 / s_1,$$

where s_1 is the standard error of $\widehat{\beta}_1$, $s_1 = \sqrt{[(X'_M X_M)^{-1}]_{1,1}}$, and we omit the subindex from T for simplicity. T follows a standard normal distribution,

Thus, using significant level $\alpha = 0.05$, the researcher rejects the null hypothesis if

$$|T| \geq Z_{1-0.05/2}$$

The two-sided p-value is

$$\widehat{p} = \mathbb{P}(T \geq t \mid H_0) = 1 - 2\Phi(|T|)$$

and the probability that the test statistic is greater than or equal to the observed value when the null hypothesis is true is at most 5%:

$$\mathbb{P}(|T| \geq Z_{1-0.05/2} \mid H_0) \leq 0.05$$

Thus, the probability of Type I error is controlled, and frequentist inferences can be interpreted in the usual way.

Scenario 2: Multiple tests with a priori model

This scenario is similar than above, except that that the researcher is interested in the effect of X_1 on multiple outcomes of interest, which we index by k, with $k = 1, \ldots, K$. We still assume that for each outcome y_k, we fit the same model M, that is, we regress each outcome on the same covariates X_M. Again, because M is fixed, we drop it from the subindex.

We now have a collection of K test statistics,

$$T_k = \widehat{\beta}_{1k} / s_{1k}, \quad k = 1, \ldots, K$$

,

corresponding to the K null hypotheses

$$H_{0k} : \beta_{1k} = 0, \quad k = 1, \ldots, K$$

The well known problem of multiple testing or multiple comparisons is that the probability of rejecting one or more null hypotheses among the K hypotheses increases very rapidly with K. For example, if the test statistics T_k are independent and $\alpha = 0.05$, the probability of making at least one false rejection is $1 - (1 - 0.05)^K$. This probability is approximately 0.40 when $K = 10$, 0.64 for $K = 20$, and is above 0.80 for any $K \geq 32$.

Thus, in any setting where the number of hypotheses tested is large, the naive procedure that ignores the multiplicity of tests and rejects H_k whenever $\widehat{p}_k \leq \alpha$ will lead to a large number of incorrect rejections or false discoveries.

This problem is easy to fix. There are multiple procedures that can be used to test large numbers of hypothesis while controlling the probability of making at least one false rejection, usually known as the Family-Wise Error Rate (FWER). One of the simplest methods is Bonferroni's, according to which one rejects those H_ks that have associated p-value below α/K. Bonferroni's procedure ensures that the FWER in testing K simultaneous hypotheses is at most α.

A very large literature in statistics has proposed various alternative methods to control the FWER, most of which seek to improve on the power of Bonferroni's procedure while still controlling the FWER (see, e.g., Lehmann and Romano, 2005). A seminal piece by Benjamini and Hochberg (1995) proposed to focus on controlling a different criterion, the False Discovery Rate (FDR).

For our discussion, the important point is that if the researcher is explicitly and formally engaging in multiple testing, there are multiple criteria that can be used to control the desired error and adjust inferences accordingly. The main characteristic of this scenario is that despite performing thousands of tests, the researchers keeps/reports the results for all of them, and draws conclusions that depend explicitly on the number of tests conducted.

Scenario 3: Multiple tests with model selection

A much more complex situation occurs when the researcher consider a multiplicity of tests but does so in a way where the total number of tests is not known

The scenario we consider now is one in which the researcher tries several specifications, but only reports the results from those misspecification where the effect of interest is statistically significant at 5%—that is, has p-value less than or equal to 0.05. We continue to assume that the researcher is mainly interested in β_1, and will pay close attention to the p-value associated with the test of the null hypothesis $H_0 : \beta_1 = 0$.

It is well known that selecting the model to be reported on the basis of whether the null hypothesis of interest is rejected invalidates the usual interpretation of hypothesis tests and confidence intervals. An extreme case occurs when the true value of β_1 is zero, and the researcher reports only those models where $\hat{p} <= 0.05$; in that case, the proportion of reported confidence intervals that cover the truth will be exactly zero.

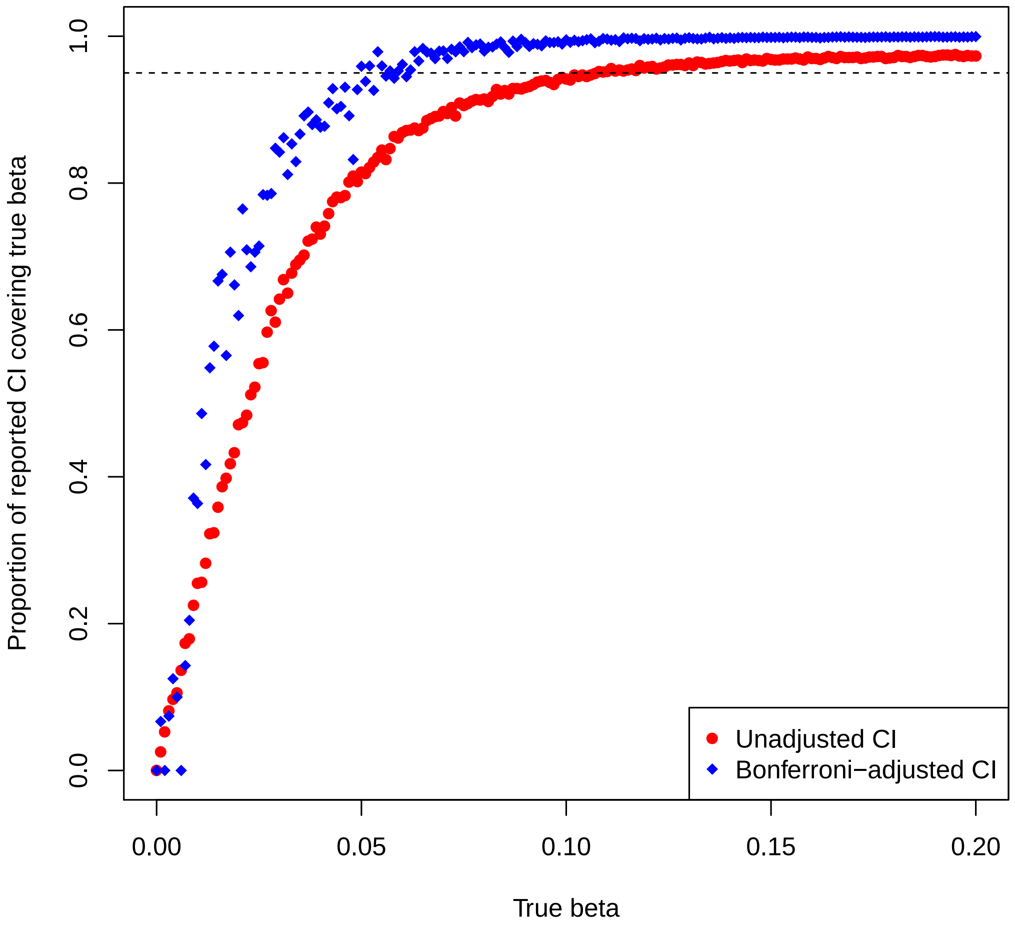
This extreme case is likely not the most common in practice. When researchers set out to investigate empirical relationships and test hypotheses, they do so informed by prior studies, theoretical knowledge, and informed intuitions. Thus, a more plausible scenario is one in which the parameter β_1 is nonzero, albeit its magnitude is unknown.

To explore how inferences are affected in this case, we performed simulations of model above, where we vary the true value of β_1. We let β_1 take the 201 distinct values in the set $[0, 0.01, 0.02, \ldots, 0.198, 0.199, 0.20]$. For each β_1, we generate 50 datasets according to 3.2, testing in each case the hypothesis that $\beta_1 = 0$, and reporting a 95% confidence interval only if the p-value is 5% or below. We then calculate the proportion of these reported confidence intervals that cover the true β_1—the conditional coverage. We repeat this procedure 500 times, and calculate the average conditional coverage across the 500 simulations. Our simulations are inspired by the examples and simulations discussed by Benjamini and Yekutieli (2005).

Figure 3.1 shows the results, plotting for each true value of β_1, average proportion of reported confidence intervals that cover it. The red dots show the result of the procedure just described (the blue dots show the analogous procedure when the p-value used in the selection rule and the confidence interval are is Bonferroni-corrected). The pattern is clear: reporting confidence intervals only for those parameters that are estimated to be statistically significant destroys coverage. Naturally, when the true β_1 is zero, the proportion of times it is covered by the confidence interval whenever this interval is reported—that is, the number of times the parameter is covered by the confidence interval divided by the number of times it is selected, is zero. This selective coverage increases as the true β_1 moves further away from zero, until it eventually reaches 95%.

The important point is that, even when each marginal confidence interval is level 95%, the conditional coverage after selection cannot be controlled (see Benjamini and Yekutieli, 2005, for a thorough discussion of this point). In the most extreme case, it can be as small as zero. And for a wide range of β_1s, it can be drastically below 95%. The actual coverage depends on the true value of β_1, which is of course unknown. Thus, a researcher that reports confidence intervals after selection cannot provide any guarantees about their conditional coverage.

Figure 3.1: Simulation of Coverage After Model Selection



The Role of Misspecification

In conversations about research integrity—or how best to interpret statistical significance claims—the focus is typically the phenomenon of informal model selection by which researchers' use of their degrees of freedom to arrive at the final model by interacting with the same data that they will use for the final analysis. These conversations typically focus on experimental designs, where the researcher randomly assigned a treatment and is interested in the effect of this treatment on one or more outcomes of interest. The focus on randomized controlled trials (RCTs) allows scholars to assume that the multiple statistical models that may be examined in the process of trial and error are all correctly specified. The typical scenario is one in which a researcher examines at the effect of the treatment on the outcome, examining a linear regression model with and without covariates. Because the covariates are orthogonal to the treatment in expectation, their inclusion or exclusion does not typically affect the parameter that is being estimated.

The research integrity literature's focus on RCTs has given the impression that the challenges to replication apply mostly to experimental designs, and that the discussion about replicability and interpretation of statistical inferences has as a prerequisite an analysis that is based on a specification that is known to be correct. *This has had the unintended consequence of mostly excluding non-experimental studies from the conversation on research integrity.*

However, nothing in our prior discussion depends on having a correctly specified model in the sense of having an estimator that is consistent for the population effect of interest. Consider again our example, where our interest was in β_M in model 3.2. By construction, this population parameter is $\beta_M = (X'_M X_M)^{-1} X'_M \mu$, that is, the orthogonal projection of μ as defined in 3.1 in the columns of X_M. Under the usual regularity assumptions, the least squares estimator $\widehat{\beta}_M = (X'_M X_M)^{-1} X'_M y$ is consistent for β_M. But it is only when $\mu = X_M \beta_M$ that the model is correctly specified and we can give β_M the causal interpretation of partial effect of X_M. That is, it's only when $\mu = X_M \beta_M$ that the estimator $\widehat{\beta}_M$ is consistent for

$$\frac{\partial \mathbb{E}(y|X_M)}{\partial X_M} = \beta_M \tag{3.3}$$

The relevant distinction is one between β_M as an estimator of the best linear predictor, and β_M as a causal effect. But the interpretation of the population parameter β_1 is unrelated to the interpretation of statistical inferences based on $\widehat{\beta}_1$.

4 The Evolution of Empirical Analyses in Political Science: 1970-2010

In this section, we illustrate the over-time trend of statistical claims in the field of political science by analyzing the findings in the papers that were published between 1970 to 2017.

Because specific traits of data (e.g. sample size, survey administration procedure) can influence the ability to produce statistically significant findings, we controlled the data type by analyzing the papers in the ANES (American National Election Studies) bibliography. Since ANES is one of the major data archives for political science, the set of papers using the ANES as the data source can reflect the overall trend of the empirical findings in the field.

Our premise is that interpreting prior results exactly as stated—or ignoring them altogether—is inconsistent with the most reasonable conclusions we can draw about the manner in which evolving research practices affected the content of statistical empirical claims. Thus we hypothesized that the cost of evaluating additional empirical models decreased over the decades in which political science journals began to regularly publish empirical claims. We expected that, over time, the costs of increasing the number of model specifications—(1) having greater number of tests (e.g. adding control variables in a regression model), (2) targeting threshold p-values (e.g. report greater number of tests that are significant at the 5% level)—have declined over time.

Regarding selecting the studies to be analyzed, we started from the ANES bibliography (1970-2017) that was retrieved on July 11, 2017 (initial number of observations = 7,028). To constrain the sample to have only journal articles, we deleted the studies of different publication type (e.g. book, newspaper article, dissertation thesis, conference paper, response/review paper) by dropping the observations with relevant terms (e.g. "annual meeting," "university press," "edited by," "reply") [2] With the resulting list of 2,312 observations, we randomly selected 40 observations by decade (1970-1979, 1980-1989, 1990-1999, 2000-2009, 2010-2017) for a total of 200 observations. We randomly selected additional 30 observations by decade as potential substitutes. If a paper in the selected list is not a research paper (e.g. response/review paper) that was not sorted out, or its content cannot be found on-line, then it was replaced by a paper from the substitute list.

Using the 200 observations that were randomly selected by decade from the ANES bibliography, we collected the following information to test our hypotheses. The specific rules that were used to extract the information are available in Appendix #.

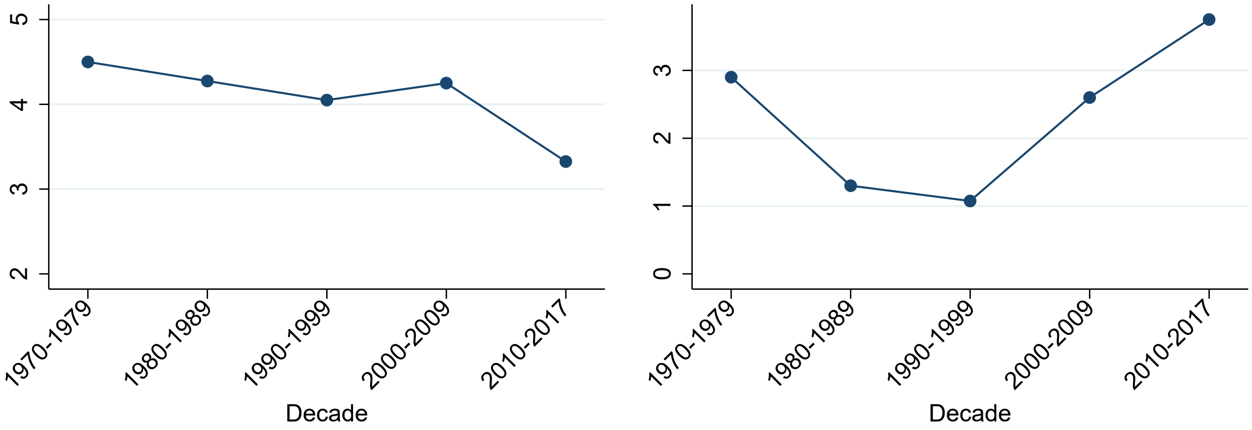
- Paper-level: number of tables, number of plots
- Table-level (each table within a paper): type (descriptive, inferential, neither descriptive nor inferential), number of tests, number of models
- Model-level (each model within a table): type of statistical model/analysis (e.g. OLS,

[2] The entire procedure is available in Appendix #.

logit, probit, t-test, likelihood ratio test etc.), number of covariates, number of significant coefficients at the 10%, number of significant coefficients at the 5%, number of significant coefficients at the 1%, one-tailed test

We use these 200 observations to produce descriptive statistics by decade, in order to establish raw patterns regarding the evolution of inferential analysis in the subset of published papers that use the ANES data. A first approach to this analysis is to understand if the number of tables in a paper has changed drastically over time, since papers have become more complex. Figure 4.1a shows that the average number of tables has remained more or less constant over time, only falling sharply in the 2010s. In the five decades of our analysis the average number of tables ranges between 3.33 and 4.5. This is probably a consequence of the space limitation imposed by journals, which has not changed by much over time. The number of plots in a paper has been more erratic. Figure 4.1b shows that the average number of plots was higher in the 1970s, then fell from an average of 2.9 plots per paper to 1.08 plots per paper in the 1990s, and then rose again to almost 3.75 plots per paper in the 2010s.

Figure 4.1: Number of Tables and Plots per Decade

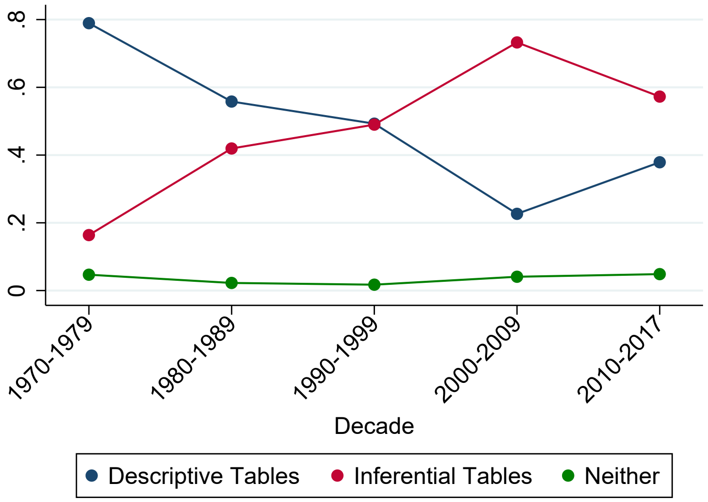


(a) Average Number of Tables

(b) Average Number of Plots

Our main interest lies in understanding the pattern between descriptive and inferential analysis, and whether descriptive analyses was progressively displaced by inferential ones. In order to do this, we classify each table as descriptive, inferential, or neither. A descriptive table is one that has some descriptive measures of the data (mean, median, minimum, etc.) but makes no inferential tests. A table classified as inferential is one that has some statistical test and reports a p-value or a confidence interval. Finally, a table classified as neither of those is one that does not contain any description or the data and also does not contain any type of statistical test. Figure 4.2 shows the average proportion of tables by type (descriptive, inferential or neither) over time. In the 1970s almost 79% of the tables of a paper were descriptive, while almost 16% were inferential. This pattern has been changing monotonically in time, and in the 2000s almost 73% of the tables in a paper are inferential while almost 23% of the tables are descriptive. The pattern changes from the 2000s to the 2010s, but we will argue later on this section that the latest decade suffers from some issues that do not make it easily comparable with the previous ones.

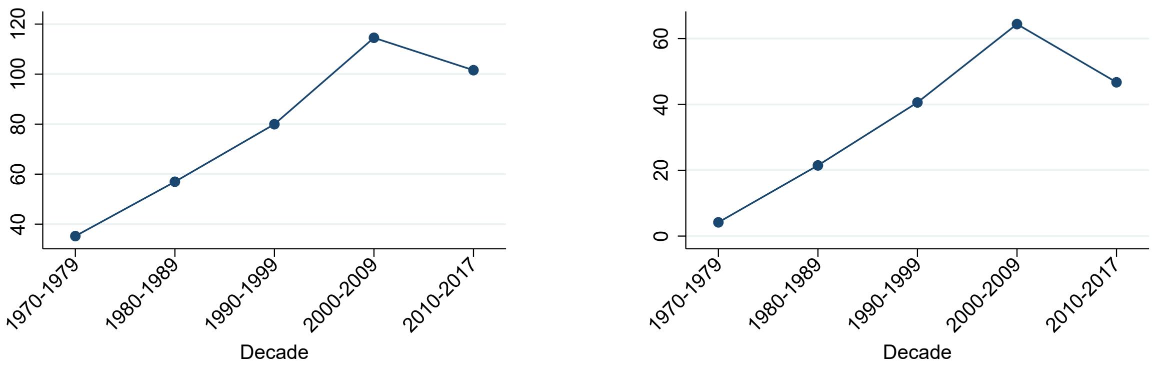
Figure 4.2: Proportion of Tables by Type



If it is true that the increase in computer power over time and the fall of its price changed the cost of doing inferential analysis, we should see specific patterns in the data. Figure 4.2 showed that inferential tables have been displacing descriptive ones, but is it also true that papers are doing more statistical tests? And if that is the case, is it true that more significant coefficients are being reported? Figure 4.3 sheds some light on these questions. The average number of tests has increased from almost 35 in the 1970s to almost 115 in the 2000s as shown by Figure 4.3a. That is, on average a paper in the ANES literature reported more than three times the number of tests in the 2000s in comparison to the 1970s. On the other hand, the average number of significant coefficients (at 10, 5 or 1%) has also increased monotonically as shown by Figure 4.3b.

In sum, the pattern shows that papers are reporting more inferential tables, more tests overall, and more significant coefficients over the decades. But it would be possible to argue that the number of significant coefficients has risen as a consequence of the rise of the number of tests. If that is true, we should see that the ratio of significant coefficients to the number of tests is more or less flat over time. That is not the case, as Figure 4.4 illustrates. In the 1970s, on average, almost 14% of the reported tests were significant at least at 10%, while

Figure 4.3: Number of Tests and Number of Significant Tests

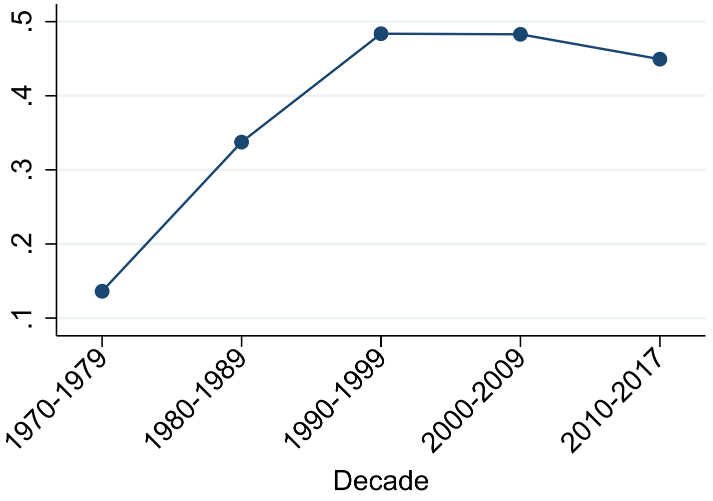


(a) Average Number of Tests

(b) Average Number of Significant Tests

for the 1990s and 2000s this number was almost 48%. This might be an indication that researchers were better at formulating the hypothesis they wanted to test, or that since it was easier to test hypothesis they were being able to test more hypotheses and show a larger number of tests with significant results to the audience.

Figure 4.4: Significant Tests to Total Number of Tests Ratio



The last period in the plots, 2010-2017, should be interpreted with caution for two reasons. First, it is posterior to the initiation of the transparency debate. It is possible that researchers in social sciences were already aware of the issues caused by multiple testing and incorrectly done model selection, and started to correct their practices. Second, there is a feature of the ANES bibliography that is worth noticing. The share of papers in our sample that was published in one of the three top journals in Political Science—American Political

Science Review, American Journal of Political Science and Journal of Politics—decreases dramatically in the 2010s. In the first three decades roughly 35-40% of the papers in the ANES bibliography were published in one of these three top journals. In the 2000s, this percentage drops to less than 30; finally, in the 2010-17 period, the percentage is almost 0%. This shows that the journal composition in our sample changes dramatically in 2010-2017, which may translate into a change in the kind of papers in this period and affect the comparability with other decades.

5 Discussion: What We Can Learn From The Steroid Era In Baseball

For over a century, Americans have spent many summer afternoons and evenings watching or listening to baseball games. Baseball games have an uneven rhythm.

The rhythm is partially serene. The game is played on a wide-open green field. The fields in professional baseball vary in size but cluster around a size of two and a half acres. From the vantage point of the fan in the stands, this two-and-half acre field—with its deep green grass punctuated with avenues of sand, with it's four white or dirt-stained bases, with it's elevated pitcher's mound and with its few chalk lines—is virtually empty for most of the game. When the game is in play, the entire population of two-and-a-half acre field is nine players from the defense, three or four umpires, one or two coaches for the offense, a batter from the side on offense, and the offensive side's player waiting-to-bat. For most of the game, most of the twenty or so people who are on the two-and-a-half acre field are standing still. They are waiting for something to happen.

The rhythm is also electric. Electricity is triggered by the crack of a wooden bat hitting a ball made of leather, yarn and rubber. The ball's rapid ascent into the sky. A moment of breathlessness. Everyone in the stadium asking themselves: *Where is it going to land?* Then, the destination becomes clear. The ball is going to...leave the field of play. It overtakes the distant wall. For a moment, the accomplishment subsumes the game itself. The large green silence to surrenders the electricity. The crowd roars (or groans). This is the home run.

Home runs constitute the highlights of many baseball games. Players who hit them become stars of the moment. For players who can hit a lot of home runs over longer periods of time, greater glory awaits. It is no longer the narrative of the game that is written around their work, it is the history of the game itself. Babe Ruth, Roger Maris, Hank Aaron. They are not written into the history as much as they become icons of that history. Their bodies

Table 4.1: Descriptive Statistics

Variable	Decade	Mean	Median	Std. Deviation	Min	Max	N
Number of Tables	1970	4.50	4.00	3.33	0	14.00	40
Number of Tables	1980	4.28	4.00	2.06	1	9.00	40
Number of Tables	1990	4.05	4.00	2.47	0	12.00	40
Number of Tables	2000	4.25	4.00	2.19	0	10.00	40
Number of Tables	2010	3.33	3.00	2.79	0	16.00	40
Number of Plots	1970	2.90	0.00	4.54	0	21.00	40
Number of Plots	1980	1.30	0.00	2.64	0	13.00	40
Number of Plots	1990	1.08	0.00	2.55	0	14.00	40
Number of Plots	2000	2.60	1.00	4.37	0	23.00	40
Number of Plots	2010	3.75	2.00	5.03	0	21.00	40
Proportion of Descriptive Tables	1970	0.79	1.00	0.33	0	1.00	36
Proportion of Descriptive Tables	1980	0.56	0.59	0.37	0	1.00	40
Proportion of Descriptive Tables	1990	0.49	0.50	0.31	0	1.00	37
Proportion of Descriptive Tables	2000	0.23	0.20	0.24	0	0.75	39
Proportion of Descriptive Tables	2010	0.38	0.33	0.37	0	1.00	36
Proportion of Inferential Tables	1970	0.16	0.00	0.32	0	1.00	36
Proportion of Inferential Tables	1980	0.42	0.39	0.38	0	1.00	40
Proportion of Inferential Tables	1990	0.49	0.50	0.32	0	1.00	37
Proportion of Inferential Tables	2000	0.73	0.80	0.30	0	1.00	39
Proportion of Inferential Tables	2010	0.57	0.67	0.37	0	1.00	36
Proportion of Non-Descriptive and Non-Inferential Tables	1970	0.05	0.00	0.11	0	0.50	36
Proportion of Non-Descriptive and Non-Inferential Tables	1980	0.02	0.00	0.09	0	0.50	40
Proportion of Non-Descriptive and Non-Inferential Tables	1990	0.02	0.00	0.06	0	0.25	37
Proportion of Non-Descriptive and Non-Inferential Tables	2000	0.04	0.00	0.11	0	0.50	39
Proportion of Non-Descriptive and Non-Inferential Tables	2010	0.05	0.00	0.13	0	0.50	36
Number of Tests	1970	35.22	54.00	33.93	0	120.00	40
Number of Tests	1980	56.95	54.00	48.35	0	208.00	40
Number of Tests	1990	79.97	63.00	70.85	0	302.00	40
Number of Tests	2000	114.53	86.50	92.82	0	422.00	40
Number of Tests	2010	101.57	78.00	98.30	0	546.00	40
Number of Significant Coefficients	1970	4.18	0.00	11.29	0	51.00	40
Number of Significant Coefficients	1980	21.48	14.50	30.09	0	149.00	40
Number of Significant Coefficients	1990	40.60	26.00	41.60	0	134.00	40
Number of Significant Coefficients	2000	64.40	37.50	84.31	0	464.00	40
Number of Significant Coefficients	2010	46.72	42.00	49.96	0	269.00	40
Significance Coefficients to Number of Tests Ratio	1970	0.14	0.00	0.26	0	1.00	24
Significance Coefficients to Number of Tests Ratio	1980	0.34	0.33	0.25	0	0.85	32
Significance Coefficients to Number of Tests Ratio	1990	0.48	0.50	0.24	0	0.98	33
Significance Coefficients to Number of Tests Ratio	2000	0.48	0.48	0.28	0	1.26	38
Significance Coefficients to Number of Tests Ratio	2010	0.45	0.52	0.27	0	1.00	33
Proportion of Papers in Top Polisci Journals	1970	0.40	0.00	0.50	0	1.00	40
Proportion of Papers in Top Polisci Journals	1980	0.38	0.00	0.49	0	1.00	40
Proportion of Papers in Top Polisci Journals	1990	0.35	0.00	0.48	0	1.00	40
Proportion of Papers in Top Polisci Journals	2000	0.28	0.00	0.45	0	1.00	40
Proportion of Papers in Top Polisci Journals	2010	0.03	0.00	0.16	0	1.00	40

of work are not written into the history as much as the history is rewritten around their bodies.

Players from competing teams, younger players, and children seek to emulate these stars and these moments. They share dreams of being these icons. They rethink their methods. They change how they prepare. All in pursuit of a dream. Of clearing the walls at key moments. Achieving fame, influence, and glory.

In the late 1980s, after 11 decades of professional baseball in America, something fascinating happened. Certain players started hitting home runs at a pace never before seen in the history of the game. The balls traveled faster and farther. They cleared the fences with greater frequency. It was very exciting. The surge altered baseball's ratio of electricity to serenity. The nation was enthralled as multiple players threatened the most iconic performance records of baseball's historic high priests. Down they went.

In 1998, Mark McGuire and Sammy Sosa spent the late summer converging on the record for most home runs in a season (61). Both players shattered the record. In that season, McGuire hit 70 home runs. Sosa hit 66 home runs, which is 26 more than he had ever hit in any past season.

The surge did not end there. Many players experienced sudden surges in their ability to hit home runs. Three years later, in fact, Barry Bonds broke McGuire's record, hitting 73 home runs in a season. Bonds would go on to pass baseball's grandest legends, Babe Ruth and Hank Aaron, to hit the most home runs in an American professional baseball career.

During this era, people asked many questions about how and why the surge was happening. Multiple theories abounded. There were uncontroversial theories built on the fact that players were training differently. More controversial explanations included the "juiced ball" theory—the idea that someone had intentionally or unintentionally changed the baseball's physical constitution to make it travel farther. A number of organizations conducted research on the juiced ball theory. Ultimately, "juiced ball theory" could not be supported.

What led to the sudden surge in performance? The sudden increase in some practitioners' abilities to hit home runs. Another explanation was circulating. The idea was that some players were engaging in a secret practice—or at least one that they were hiding from fans. This practice had nothing to do with a change in the field of play or a change in the ball. It was a change that, in the hands of someone sufficiently skilled in the art, could make players appear stronger and more vital than their bodies would otherwise allow.

Investigations began. Investigators acquired more data about past practices. Their attention began to focus on another explanation of the surge in home runs—anabolic steroids.

Anabolic steroids stimulate muscle building. They produce greater muscle mass per unit of exercise which can increase strength and hasten recovery from certain types of injury. They also have well known negative effects. These effect range from the relatively innocuous (increased acne) to deadly (kidney, liver, and heart malfunction).

The temptation to use steroids was great for some players. Leading home run hitters gained greater wealth and fame. During the peak of baseball's steroid era, McGuire, Sosa, and Bonds were among America's most famous athletes. A lingering concern about steroid use is not just that the lure of wealth and fame would lead other professional baseball players to begin the practice, but that the incentives would spread further—into minor league baseball, college baseball, high school baseball, and even little league. In these latter venues, the long-term effects of steroids were just as likely (and, in fact, increased for children at certain developmental stages)—while the monetary rewards for taking such risks were fare less likely.

Other players declined to use steroids. For example, in the years prior to the McGuire-Sosa-Bonds-led home run explosion, Ken Griffey Jr was one of baseball's most famous players and productive home run hitters. In the steroid era, his body did not undergo the massive and very visible transformations of his rivals. During that time, his accomplishments were overshadowed by the players who were employing a hidden factor to increase their numbers. While evidence has emerged tying other leading players to steroids, such evidence has not arisen regarding Griffey.

Today, Ken Griffey Jr is the only one of the players named in this article to be in baseball's Hall of Fame. The other players are now seen differently. They are seen as having played by a different set of rules. They are seen as having deceived others about how they were getting their results.

As conversations about scientific transparency and related topics evolve, it is only natural to ask questions about how scientific disciplines should interpret their legacies. These questions are not just of theoretical interest. Increasing numbers of studies are showing substantial negative implications accruing from the mass production and distribution of "false positives." In the field of cancer oncology, for example, faulty empirical practices produced false positives which in turn caused thousands of doctors and patients to place false hope in faulty practices and medicines—faults that would have been apparent if researchers were more forthcoming about how they produced results. If we believe that our work is socially valuable and consequential, then we should also be concerned about the effects of false positives from our own fields of research on the populations that we are attempting to serve.

We operate from the assumption that most, if not all, scholars who engaged in interactive trial-and-error to guide their empirical analysis did not do so out of malice. We appreciate that, over a series of decades, the combination of dramatic technological change and new professional incentives evolved in ways that made this practice natural and socially acceptable. Many fields of study had too few people who understood the statistical and inferential errors entailed in the production of such claims. Moreover, in the pre-Internet era, those few who did understand had limited ability to document the process, correspond with other experts about it, or propose remedies. So, we do not advocate following baseball's path in the treatment of prior generations. Becoming a scholar is very difficult. It is reasonable to assume that most people were doing the best with the opportunities and incentives that they had.

Today, however, we know more. We know that producing a statistical claim through an interactive process of trial-and-error in which the same data to be used for final analysis is used to perform exploratory statistical inferences and then representing the claim as if it were the product of a data-independent, a-prior model is likely to lead readers of the work to misinterpret what the finding means. So this practice has to stop.

Stopping this practice means changing incentives. Journal editors, for example, can better serve their readers by instituting review practices that elevate valid explanations and careful descriptive analyses over splashy significance claims. Researchers can preregister designs—or at least keep (and then share) detailed logs of all decisions made regarding data analyses—provided, of course, that the revelation of such decisions does not endanger human subjects or violate applicable contracts or laws.

Exploratory analysis is an important tool for scientific advancement. Not all scientific discovery come from a-priori theorizing; many crucially important scientific findings arose from empirical exploration. Performing exploratory analysis is not a problem; the problem arises when exploration is informal and unreported, and the results from exploration are presented as if they were the result of a-prior theorizing. In this case, statistical inferences do not have the usual properties, and there is no guarantee that the probability of making a false claim can be controlled. Researchers can contribute to the solution by clearly reporting which subset of the published findings are the result of exploratory trial-and-error, and which subset is the result of an a-priori analysis.

With such changes, we can help others more accurately interpret the meaning of our work. In so doing, our science can provide more value to more people. This would be a genuine human accomplishment.

6 Conclusion

Baseball, as other sports, creates a venue where we can evaluate human accomplishment. The game begins with a set of rules. It produces a clear result—a win, a loss, or a draw. As a result, sports offers a compact venue for comparative evaluations. It can be used to teach important life lessons about hard work, perseverance, and the benefits of learning to work with others.

These attributes of sports are why perceptions of cheating are a big deal. Steroid use, like other forms of cheating, were not illegal during the steroid era. By some ethical standards, there is nothing inherently wrong with pursuing every means possible to gain performance advantages over competitors. But cheating alters our ability to use the context of sports to comparatively evaluate human performance.

In many cases, cheating in sport is done for purposes of individual gain. More wins. More money. More fame. Because, however, cheating reduces our ability to make apples-to-apples comparisons of human achievement, instances of cheating are often treated not as crimes against the opposition, but crimes against the sport itself. When cheating is rampant, fans of the sport become less certain about the meaning of what they saw. Elements of uncertainty enter tales of work ethic and perseverance.

Science also has rules. The rules are part of what make science influential. To say that a scholar was rigorous, for example, implies that they rigorously worked with respect to a particular subfield's or paradigm's set of rules. The existence and evolution of such rules become the basis of how scholarly communities can claim to know what they know.

Scholars that engage in interactive trial-and-error can certainly experience private gains relative to scholars who try and report a single model whatever the result. They are more likely than others to be published. They become more likely than others to be cited. But at what cost? If scholars are sufficiently far removed from policymakers and members of the public who are affected by their false positives, they will not see the costs that this practice imposes. When this practice is widespread and statistically significant claims are rewarded by professional advancement, then the social costs are multiplied.

To be sure, many important things were discovered in political science's steroid era. Our work implies, however, that we may be wrong about which discoveries have value as we move forward. Our responsibility, at this point is to be more open and transparent about how we produce empirical claims and to change academic ecosystems so that private gain and public service are not mutually inconsistent.

Around the world, millions of people count on science to improve the quality of their lives. Science is not a game. The stakes are real. Our generation has an opportunity to make science better by more accurately interpreting the past and by incentivizing better practice. We appreciate the efforts of everyone who is working to make this happen.

Bibliography

Benjamini, Y., and Hochberg, Y. (1995), “Controlling the false discovery rate: a practical and powerful approach to multiple testing,” *Journal of the royal statistical society. Series B (Methodological)*, 289–300.

Benjamini, Y., and Yekutieli, D. (2005), “False discovery rate–adjusted multiple confidence intervals for selected parameters,” *Journal of the American Statistical Association*, 100, 71–81.

Dudoit, S., Shaffer, J. P., and Boldrick, J. C. (2003), “Multiple hypothesis testing in microarray experiments,” *Statistical Science*, 71–103.

Efron, B. (2010), *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, Vol. 1, Cambridge University Press.

Högman, C. F., and Ramgren, O. (1970), “Computer system for blood transfusion service,” *Transfusion*, 10, 121–132.

Ioannidis, J. P. (2005), “Why most published research findings are false,” *PLoS medicine*, 2, e124.

Kurland, L. T., and Molgaard, C. A. (1981), “The patient record in epidemiology,” *Scientific American*, 245, 54–63.

Leeb, H., Pötscher, B. M., and Ewald, K. (2015), “On various confidence intervals post-model-selection,” *Statistical Science*, 30, 216–227.

Lehmann, E. L., and Romano, J. P. (2005), *Testing statistical hypotheses*, Springer Science & Business Media.

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

Waldrop, M. M. (2016), “More than moore,” *Nature*, 530, 144–148.

Appendix A Constructing the Sample from the ANES Bibliography

The initial number of observations in the ANES bibliography is 7,028 (retrieved on July 11, 2017 from the ANES website). The following procedure is executed in order to construct the final sample:

1. Transform all strings to lower case, in order to make all string manipulations easier and uniform.

2. Extract the year when the document was published. If this cannot be done, the observation is dropped.

3. Generate a decade variable from the year of publication following the formula:

$$\text{decade} = \left\lfloor \frac{\text{year}}{10} \right\rfloor \cdot 10$$

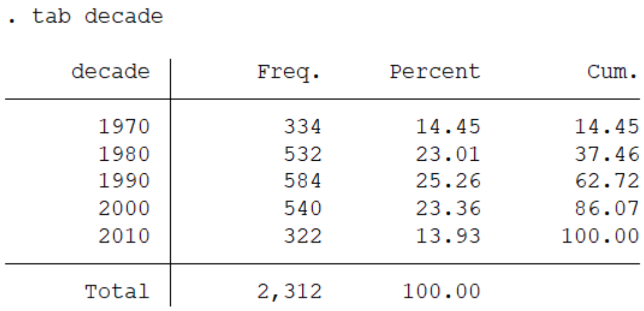
   where $\lfloor . \rfloor$ denotes the integer floor function. So the year 1952 becomes decade 1950, the year 1969 becomes decade 1960, etc. We only keep observations from the decades 1970, 1980, 1990 and 2000.

4. We drop all observations that have any of these combination of words in the title or journal field: “washington post”, “new york times”, “annual meeting”, “working paper”, “in the economist”, “chicago tribune”, “dissertation thesis”, “los angeles times”, “university press”, “edited by”, “annual” and “meeting”, “annual” and “conference”, “washington times”, “reply”, “press”, “books”, “news”, “in the”, “symposium”, “prepared” and “conference”, “thesis”, “.com”, “pp.”.

5. We also drop observations that do not have the symbols “:” and “-”, since these are used in all journal articles in the ANES bibliography.

6. We drop all observations with an empty title.

After following these steps the remaining number of observations is 2,312, meaning that 4,716 are dropped. From the remaining observations we have the following distribution by decade:



```
. tab decade

     decade |      Freq.     Percent        Cum.
------------+-----------------------------------
       1970 |        334       14.45       14.45
       1980 |        532       23.01       37.46
       1990 |        584       25.26       62.72
       2000 |        540       23.36       86.07
       2010 |        322       13.93      100.00
------------+-----------------------------------
      Total |      2,312      100.00
```

After this, we generate random uniform numbers in order to keep 40 observations by decade for a total of 200 observations.

Appendix B Rules Used for Collecting Data from the ANES Bibliography

For each variable, we followed these rules to classify and retrieve information:

- **Number of tables and number of plots:**
  - Count the number of distinct/independent tables and plots in the text of a paper.
  - Plots are defined as: symbolic representation of a phenomenon, regardless of involving numbers/data or not (e.g. graphs with horizontal and vertical axes, flow charts that summarize a theory)
  - If multiple plots are presented under a single header, ignore the author's grouping/numbering scheme, but count all disticnt plots (e.g. If three plots are presented under the header "Figure 1," then add 3 to the number of plots variable.)
  - If a table or a plot is presented in the footnote, then do not count in the respective variable.
  - If a table or a plot is presented in appendix, then do not count in the respective variable.
- **Descriptive and inferential tables:**
  1. **Descriptive table**
     - Descriptive table is defined as: a table that involves data but provides no statistical inference. It can contain summary statistics such as means and

percentages. Even if the author makes inferences (e.g. use terms such as “effect”) or presents statistical estimates (e.g. regression coefficients), if the table does not provide information that allows statistical inference (e.g. p-value, confidence interval, asterisks (stars that denote statistical significance), test statistics (z-score, t-score, chi-square)), then the table is categorized as descriptive.

2. **Inferential table**
   - Inferential table is defined as: a table that involves data and provides statistical inference. If a table contains information that allows statistical inference (e.g. p-value, confidence interval, asterisks, test statistics with degrees of freedom), from which confidence interval can be built and statistical hypotheses can be tested, then the table is categorized as inferential.
3. **Non-descriptive and non-inferential table**
   - A table can be neither descriptive nor inferential.
   - Examples of a table that is neither descriptive nor inferential include the following, but not limited to:
     - a table with a flow chart
     - a table with a list of terms
     - a table with a hypothetical scenario or example

- **Number of tests in a table:**
  - Count the number of distinct/independent tests presented in each table.
  - Examples of how to count the number of tests include the following, but not limited to:
    - If an inferential table presents regression coefficients, then add the number of coefficients with inferential information (e.g. p-value/CI/star/test-stat (t-score, beta hat/se)) to the count variable.
    - If an inferential table presents t-test statistics, then add the number of t-test statistics with inferential information.
    - If a table provides “ns” (not significant) where test-statistics or p-value should be presented, then count that entry as a test.

- **Type of model:**

  - Record the type of analysis used by each model in a table.
  - Examples the type of models recorded include the following, but not limited to: OLS, Logit, Multinomial Logit, Ordered Logit, Probit, Multinomial Probit, Ordered Probit, Two-stage Least Squares Regression, Negative Binomial Regression, Analysis of Covariance, Factor analysis, t-test of the difference in means, Chi-square test of independence, Log-likelihood ratio test, Likelihood ratio test, Tau test (Kendall rank correlation coefficient), (Pearson) correlation, Principal Component Analysis, Goodman and Kruskal's gamma (measure of rank correlation), LISREL (linear structural relations), Scheffe Test, F-test.
  - When a paper does not explicitly mention the type of analysis for a model, then we record that the type of the analysis is unknown.

- **Number of covariates in the model:**
  - For regression-type analysis, this variable counts the number of regression coefficients reported.
  - For factor analysis, the number of covariates is the number of manifest variables/items in each model.
  - Count the number of covariates as 1 for non-regression type analysis in an inferential table. Examples include but not limited to: t-test of the difference in means, chi-square test of independence, log-likelihood ratio test, likelihood ratio test, tau test (Kendall rank correlation coefficient), (Pearson) correlation, principal component analysis, Goodman and Kruskal's gamma (measure of rank correlation), LISREL (linear structural relations), Scheffe Test, F-test.

- **Three variables to capture the number of significant coefficients at 10, 5 and 1%:**
  - Count the number of significant coefficients at the 10%, 5%, 1% level, respectively, for each model.
  - If asterisks are used to indicate the statistical significance in a table, then record the number of significant coefficients at the levels that the author reports. For example, if an author reports statistical significance only at the 5% level, then record the number of significant coefficients at 5% and leave blank the other two variables (1 and 10%).

  - If the statistical inference from a table is discussed in the text of a paper, but not in the table (header, entries, notes, footnotes of a table etc.), then count as zero significant coefficients.
  - If a table reports statistical significance for both one-tailed and two-tailed tests, then record the number of significant coefficients with two-tailed tests.

- **One-tail p-value variable:**
  - If a table provides one-tailed test, then record a 1. If a table does not mention whether it uses two-tailed or one-tailed test, then record a 0. If a table mentions that it uses two-tailed test, then record a 0.