Brooklyn Law Review

Volume 87 | Issue 2 Article 2

2-16-2022

Credibility in Empirical Legal Analysis

Hillel J. Bavli

Credibility in Empirical Legal Analysis

Hillel J. Bavli†

INTRODUCTION

Empirical analysis is a central component of modern legal scholarship and litigation. But it is not trusted.[1] It is well known that data can be manipulated to generate any results that a researcher seeks to find. As Ronald Coase has famously stated, "if you torture the data enough, nature will always confess."[2] In scholarship, empirical claims are frequently understood more as a "cacophony of subjective opinions on the meaning of disparate findings" than as objective scientific results.[3] In the courts, experts are widely recognized as "hired guns" who will utilize data to arrive at whatever conclusion most favors the party that provides their paycheck.[4] Indeed, the state of distrust surrounding empirical analysis in law is well founded.

Empirical analysis is not inherently untrustworthy. But the way in which it is conducted in law is. The unreliability of empirical analysis in law cannot be attributed to a single cause. It is due to a wide range of methodological and institutional factors. There is, however, a single problem that underlies the lion's share of the quality failure in empirical legal analysis: *data fishing*.[5]

---

† Assistant Professor of Law, SMU Dedman School of Law; Affiliated Faculty, Harvard Institute for Quantitative Social Science. The author wishes to thank William Hubbard, Christopher Robertson, Edward Cheng, Pamela Metzger, Andrew Davies, Iavor Bojinov, Bernard Chao, Donald Rubin, Daniel Heitjan, Barry Goldstein, and Eric Ruben for their helpful comments, and SMU Dedman School of Law and the WWB Law Professor's Fund for their generous financial support.

1 *See* Kathryn Zeiler, *The Future of Empirical Legal Scholarship: Where Might We Go from Here*, 66 J. LEGAL EDUC. 78, 78–80 (2016); Gregory Mitchell, *Empirical Legal Scholarship as Scientific Dialogue*, 83 N.C. L. REV. 167, 169–70 (2004); Lee Epstein & Gary King, *The Rules of Inference*, 69 U. CHI. L. REV. 1, 6, 15, 54–55 (2002); *see also* John P.A. Ioannidis et al., *The Power of Bias in Economics Research*, 127 ECON. J. F236, F236 (2017).

2 R.H. COASE, HOW SHOULD ECONOMISTS CHOOSE? (1982), *reprinted in* RONALD H. COASE, ESSAYS ON ECONOMICS AND ECONOMISTS 15, 27 (1994).

3 Mitchell, *supra* note 1, at 176–79.

4 *See* Christopher Tarver Robertson, *Blind Expertise*, 85 N.Y.U. L. REV. 174, 184–89 (2010).

5 *See infra* Part I.

Data fishing, also known as data dredging or *p*-hacking, is a well-recognized problem in the hard and social sciences that involves using data to search for and selectively (and misleadingly) report results that are statistically significant or otherwise favorable to the researcher.[6] Data fishing allows a researcher to manipulate data and the researcher's analysis to find patterns that do not in fact exist, or to find particular results that will support the researcher's claims, even when the data are not supportive of, or are contrary to, these claims. Notwithstanding its invalidity, this practice is extremely prevalent in legal scholarship and litigation—with damaging consequences.[7]

It is estimated that, when pharmaceutical giant Merck allegedly manipulated its data regarding the side effects associated with its new pain drug Vioxx, this extreme case of data fishing led to the premature deaths of thousands of Vioxx users.[8] In law, data fishing can similarly destroy lives and cause a range of other harms. A scholarly article with misleading empirical results can lead to harmful policies that affect millions of people. Testimony based on misleading statistical analysis can lead to false convictions or false acquittals, or incorrect findings of liability or no liability in major class actions.[9] More broadly, data fishing causes large-scale harm by misleading lawmakers, factfinders, and other consumers of empirical research.

Data fishing is possible because statistical studies are not rigid formulaic structures: they require a significant level of researcher input. Like other forms of research, they demand thoughtful and logical design based on the particulars of a study. A researcher, for example, must decide how to define key variables, how to handle outlying data, what tests to use and what standards to apply for finding that a test result is indicative of a significant finding, how to model the data, what sample size to use, and many other factors. Data fishing allows researchers to manipulate these discretionary inputs to arrive at results that are favorable to them.

---

[6] *See* Joseph P. Simmons et al., *False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant*, 22 PSYCH. SCI. 1359, 1359 (2011) ("[I]t is common (and accepted practice) for researchers to explore various analytical alternatives, to search for a combination that yields 'statistical significance,' and to then report only what 'worked.'").

[7] *See infra* Sections I.D, I.E.

[8] *See* STEPHEN T. ZILIAK & DEIRDRE N. MCCLOSKEY, THE CULT OF STATISTICAL SIGNIFICANCE: HOW THE STANDARD ERROR COSTS US JOBS, JUSTICE, AND LIVES 28–31 (2008); Peter Jüni et al., *Risk of Cardiovascular Events and Rofecoxib: Cumulative Meta-analysis*, 364 LANCET 2021, 2021–27 (2004); *see also Merck Manipulated the Science About the Drug Vioxx*, UNION OF CONCERNED SCIENTISTS (Oct. 12, 2017), https://www.ucsusa.org/resources/merck-manipulated-science-about-drug-vioxx [https://perma.cc/J88M-MZKH].

[9] *See infra* Parts I, IV.

An important factor underlying data fishing is a concept called *motivational bias*. Motivational bias is the tendency for a researcher to favor one result over another due to the interests of the researcher.[10] An organization may sponsor research to show that its activities are effective; a political group may sponsor research to support its political agenda; a scholar may conduct research to obtain a well-placed publication; and a litigation party may sponsor research to support its argument in court. Thus, when data fishing is possible, motivational bias can combine with researcher discretion to yield spurious results and false claims.[11]

There are two categories of problems that lead to the invalidity of statistical studies that rely on data fishing: false positives and false impressions. Both of these categories of problems cause the reader to be misled to believe that the researcher has found significant and replicable results when in fact the researcher has not.[12] These are not minor problems—they frequently altogether invalidate the researcher's findings. As stated in a recent report by the National Academies of the Sciences, Engineering, and Medicine (NASEM), "when exploratory research is interpreted as if it were confirmatory research, there can be no legitimate statistically significant result."[13] At best, therefore, data fishing promotes distrust of empirical research and statistical claims; at worst, it propagates false information and causes poor decision-making, including the possibility of incorrect verdicts and destructive policy.

To be clear, I do not mean to suggest that data fishing necessarily results from purposeful deception or other ill intention. To the contrary, the data-fishing norm is strong, and the practice is often committed by well-intentioned researchers who may be completely unaware of the harms of data fishing or that their analysis even constitutes data fishing.[14] Indeed, a

---

[10] Gilberto Montibeller & Detlof von Winterfeldt, *Cognitive and Motivational Biases in Decision and Risk Analysis*, 35 RISK ANALYSIS 1230, 1230 (2015) (describing motivational biases as "conscious or subconscious distortions of judgments and decisions because of self-interest, social pressures, or organizational context"); *see* Simmons et al., *supra* note 6, at 1359–60 (noting that "[a] large literature documents that people are self-serving in their interpretation of ambiguous information and remarkably adept at reaching justifiable conclusions that mesh with their desires," and suggesting that this bias causes methodological decisions that lead to statistical significance).

[11] *See* Mitchell, *supra* note 1, at 180 (highlighting personal, economic, and political factors that bias empirical research).

[12] *See infra* Section II.C.

[13] NAT'L ACADS. OF SCIS., ENG'G, & MED., REPRODUCIBILITY AND REPLICABILITY IN SCIENCE 96 (2019) [hereinafter NASEM REPORT], https://doi.org/10.17226/25303 [https://perma.cc/7R4M-RY2F].

[14] Additionally, researchers are often unaware of their own motivational biases or how such biases may permeate their work. Even if a researcher is fully aware of the

primary cause of the problem is a lack of awareness and understanding among researchers. Motivational bias and researcher discretion are often inherent components of empirical research. But data fishing need not be. It can be largely eliminated with attentiveness to the issue and simple safeguards instituted by readers and researchers.

My aim in this article is to facilitate the elimination of data fishing in legal scholarship and litigation in two ways. First, I explain in clear and simple terms what data fishing is, why it is harmful, and why it should be eliminated in empirical legal research. Second, I draw on established methods in statistics and other fields to develop a concrete framework I call DASS—an acronym for Design, Analyze, Scrutinize, and Substantiate—for researchers (including empirical scholars and expert witnesses) to use to safeguard against data fishing, and for consumers of empirical research (including scholars, courts, policymakers, and members of the public) to use to evaluate the reliability of a researcher's statistical claims.

In summary, DASS requires (1) designing a study—essentially, contemplating and specifying its methodological features—prior to analyzing the study's outcome data; (2) analyzing the outcome data pursuant to the study's design; (3) scrutinizing the study to ensure that it is not misleading to readers with respect to the robustness of the study's results; and (4) substantiating these steps, including by attesting to the researcher's adherence to DASS in the study's report and by establishing evidence of its elements.

Central to DASS is the idea that it is the researcher's burden to proactively take steps to safeguard against data fishing and to persuade readers of these steps in the foreground of the study's report together with other aspects of the study's methodology. This underlying feature of DASS provides important advantages over common practices in the natural and social sciences.[15] It creates proper incentives for researchers and ensures that readers obtain information necessary to properly evaluate a study's statistical claims.

DASS thus represents an advancement over current practices in a number of respects. First, it combines key protections against data fishing in statistics and packages them in a

---

harms of data fishing, the researcher, perhaps influenced by the prevalence of the practice, may nevertheless have good intentions—e.g., to uncover truth or to convince readers of a point that the researcher believes to be true. Separately, in litigation, data fishing is often expected and accepted: it is often understood as a natural consequence of the adversarial system, reflecting the expectation that a litigant will search for and offer evidence that most favors their position.

[15] *See infra* Part II.

framework that is substantively and logistically suitable for a broad range of research contexts, including legal scholarship and litigation. Second, it improves on common practices in the natural and social sciences, which themselves are not sufficiently effective. These practices—which primarily constitute journal requirements for some form of design "preregistration" or open access to data[16]—reflect substantial progress toward credible statistical inference in a number of nonlegal fields.[17] However, they are applied infrequently and inconsistently, and, even when used, they by and large constitute a hodgepodge of requirements enforced by individual journals rather than a coherent standard.[18] DASS overlaps with these practices in various respects (most prominently, regarding the value it places on methodological prespecification), but it is distinct: it is a concrete framework for safeguarding against data fishing that is intended for use by both researchers and readers, and at its center is the idea that it is the researcher's responsibility not only to be proactive in safeguarding against data fishing but also to evidence her safeguards directly to readers and thereby signal credibility to them directly. Arguably, this is particularly important in legal scholarship, which generally does not involve a peer-review selection process for publications.[19]

In Part I of this article, I explain the problem of data fishing in clear and simple terms. I demonstrate the problem with straightforward examples and show why data fishing causes false positives and false impressions, and why it altogether invalidates statistical studies that rely on it. I then discuss implications of data fishing for legal scholarship and expert evidence in litigation.

In Part II, I explain the elements of DASS, including what they entail and how they safeguard against data fishing and negate

---

[16] *See infra* notes 74–77 and accompanying text. Other practices, such as requiring general ethics statements and robustness checks, also exist in some contexts.

[17] *See generally* Joshua D. Angrist & Jörn-Steffen Pischke, *The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics*, J. ECON. PERSPS., Spring 2010, at 3 (arguing that a focus on research design has been central to increased credibility in empirical economics); Daniel E. Ho & Donald B. Rubin, *Credible Causal Inference for Empirical Legal Studies*, 7 ANN. REV. L. & SOC. SCI. 17 (2011) (discussing advances in causal inference in empirical legal studies and emphasizing the importance of good research design).

[18] *See infra* Section I.D.

[19] Natural-science and social-science journals, unlike the vast majority of scholarly journals in law, generally involve a peer-review selection process for the publication of articles. *See, e.g.*, *Publication Process*, NEW ENG. J. MED., https://www.nejm.org/media-center/publication-process [https://perma.cc/SE2X-DNPF] (summarizing peer-review process for medical journal); *Journal Policies*, Q.J. ECON., https://academic.oup.com/qje/pages/policies [https://perma.cc/3YVU-DGK4] (summarizing peer-review process for economics journal). This difference highlights a significant vulnerability for empirical legal scholarship with respect to the risk of data fishing. Pressure on authors to present research at conferences and workshops may help, but not necessarily. In any event, it is far from sufficient.

its harmful consequences. Additionally, I discuss practical considerations regarding implementation and explain how certain details of the framework flex to accommodate a wide variety of research settings and conditions.

In Part III, I examine a significant concern: the importance of using data exploration to inform study design. Specifically, while DASS requires safeguarding against data fishing by designing a study prior to analyzing its data, a study's design will often benefit immensely from the use of exploratory analysis to develop and refine the study's methodology. To address this issue, I again draw on methods in statistics and other fields—in particular, pilot studies and a procedure called "cross validation"—to explain a simple approach, consistent with DASS, for fulfilling the researcher's need for exploratory analysis without compromising the validity of the study. This approach, which involves partitioning a dataset into data for exploration and data for testing (or, in the experimental context, obtaining pilot data for exploration prior to beginning the main study), can be used in conjunction with DASS to allow for such exploration while still safeguarding against data fishing and its harmful effects.

In Part IV, I examine in greater detail the implications of data fishing for litigation, and for expert evidence in particular. I argue that data fishing plays a substantial role in the "hired-gun" and battle-of-the-experts problems in evidence law. I explain that, in cases in which expert evidence involves empirical analysis, data fishing allows an expert to search for and use a methodology that is most favorable to the sponsoring litigant's position. Consequently, the mere opportunity for data fishing may give rise to a prisoner's dilemma situation in which opposing experts engage in data fishing and ultimately a disingenuous battle over methodology, and in which the jury trusts neither expert and often selects a winner based on criteria other than the merits of the experts' arguments. This situation harms accuracy and degrades the public's faith in the courts. I show, however, how courts can apply the DASS framework to address this problem by preventing data fishing.

Finally, I conclude by highlighting the importance of DASS's substantiation element and considering concretely, in light of this article's analysis, what it means for a researcher to attest to her adherence to DASS.

I. DATA FISHING: THE CENTRAL PROBLEM IN EMPIRICAL LEGAL RESEARCH

In recent years, there has been enormous growth in empirical legal studies, and empirical research now pervades many areas of legal scholarship and litigation.[20] At the same time, *confidence* in empirical legal research is low.[21] Worse, this distrust is justified. In this Part, I explain the central role of data fishing in the trust crisis in empirical legal research. I begin by describing a number of preliminary statistical concepts and by demonstrating the problem using a dataset and a recent study. I then explain concretely why data fishing leads to results that are statistically and substantively invalid. Finally, I discuss the prevalence of data fishing in empirical research in legal scholarship and litigation.

A. Statistical Inference: Preliminary Concepts

Statistical inference involves learning about a group of objects—a *population*—by examining a subgroup of that population—a *sample*.[22] For example, if I want to learn the average age of students in my one-hundred-student class, I may randomly select 10 students and use the average age of the 10 students to better understand the average age of the entire class of students. The entire class of students constitutes the population, and the random selection of 10 students constitutes the sample. I can define an *estimand*—the thing I want to estimate—as the mean of the ages of all students in my class, the *population mean*. I can define an *estimator*—the thing I will use to estimate the estimand—as the mean of the ages of all

---

[20] *See* Daniel E. Ho & Larry Kramer, *Introduction: The Empirical Revolution in Law*, 65 STAN. L. REV. 1195, 1195–1202 (2013) (discussing the "enormous shift in interest" and work in empirical legal studies); Theodore Eisenberg, *The Origins, Nature, and Promise of Empirical Legal Studies and a Response to Concerns*, 2011 U. ILL. L. REV. 1713, 1713 (2011) (discussing growth of empirical legal studies); Gregory C. Sisk, *The Quantitative Moment and the Qualitative Opportunity: Legal Studies of Judicial Decision Making*, 93 CORNELL L. REV. 873, 874–76 (2008) (commenting on growth of empirical legal research); Tracey E. George, *An Empirical Study of Empirical Legal Scholarship: The Top Law Schools*, 81 IND. L.J. 141, 142 (2006) (highlighting "dramatic[ ] . . . expan[sion]" of empirical legal studies "in law reviews, at conferences, and among leading law faculties" (footnotes omitted)); *see also* Zeiler, *supra* note 1, at 78–80 (discussing importance of empirical scholarship).

[21] *See supra* note 1.

[22] *See generally* MORRIS H. DEGROOT & MARK J. SCHERVISH, PROBABILITY AND STATISTICS 376–94 (Pearson Educ. 4th ed. 2012) (explaining statistical concepts related to estimation); DAVID COPE, FUNDAMENTALS OF STATISTICAL ANALYSIS 27–31 (2005) (explaining statistical concepts related to sampling); GEORGE CASELLA & ROGER L. BERGER, STATISTICAL INFERENCE 311–72 (2d ed. 2002) (explaining statistical concepts related to estimation).

students in my sample. Once I have computed this *sample mean*, it will serve as my *estimate* of the population mean.[23]

In this context, *hypothesis testing* is a procedure for examining whether a claim regarding a population is supported by the evidence, the sample data.[24] Assume, for example, that, in the context of an argument regarding trends away from traditional routes to law school, I want to evaluate the claim that the average age of law students in my course is different from 24. I could construct a hypothesis test that tests the *null hypothesis* that the mean age of students in my course is 24. I test this against an *alternative hypothesis* that the mean age of students in my course is different from 24. Assume that my sample mean is 25.5. Is this difference from 24 sufficient to reject the null hypothesis that the average age of students in my course (rather than just in my sample) is 24? To decide this, I could define a level of *statistical significance*, which, here, represents the difference between the sample mean and the hypothesized population mean that would provide good evidence that the population mean is not as hypothesized.[25]

Similarly, we could use hypothesis testing to test the claim that the average age of men in my course differs from the average age of women in my course. We could take a random sample of 10 men and 10 women and compare the difference in means in the two groups to the null hypothesis of "zero difference" to determine whether there is a statistically significant difference as to justify the conclusion that the average age of men and average age of women in my course differ. A difference of one or two years may be insignificant if the ages of students in the course vary substantially—in which case, the difference may simply reflect randomness associated with the samples. Or, the difference may be significant if there is little variability in the ages of students in the course.

Commonly, researchers use a *level of significance* of 0.05, or 5%.[26] This means that the researcher would reject the null hypothesis if the observed difference (e.g., from the hypothesized mean in the first example above, or between the male and female groups in the second example above) is sufficiently large such

---

[23] *See generally* COPE, *supra* note 22, at 28–31, 48–55 (explaining statistical concepts related to estimation).

[24] *See id.* at 36–41; DEGROOT & SCHERVISH, *supra* note 22, at 530–623.

[25] *See* COPE, *supra* note 22, at 40 ("A *statistically significant difference* between a population mean and the mean of a random sample is a difference large enough to justify the claim that the sample was taken from a population with a mean different from the mean of the given population.").

[26] *Id.*

that we would observe a difference of such size due to randomness (from sampling) only 5% of the time, assuming that the null hypothesis (e.g., the mean age equaling 24 in the first example, or the mean difference between men and women equaling zero in the second example) is true.[27]

A *p-value* is defined as the probability of observing a value at least as extreme as the observed value, assuming the truth of the null hypothesis.[28] Thus, the researcher compares the p-value to the level of significance to determine whether to reject the null hypothesis in favor of the alternative hypothesis.[29]

Finally, a *type I error* is defined as rejecting the null hypothesis when it is in fact true, and a *type II error* is defined as not rejecting the null hypothesis when it is in fact not true.[30]

B. What Is Data Fishing?

Data fishing is the practice of searching numerous research methodologies—including different models, design components, analytical methods, and hypotheses—and selectively reporting only those that produce significant or otherwise favorable results.[31] As suggested in the Introduction, a researcher will generally encounter numerous choices in conducting a study aimed at answering a particular research question.[32] These choices may relate to, e.g., whether to include a particular variable in a regression model, how to define a causal effect, how to estimate a causal effect, what sampling methods to use, how to categorize the data, what testing procedures to use, what level of significance to set, how to handle outliers, how to group the data, what multiple-comparisons adjustment to use, and many other aspects of a study's

---

[27] *Id.*

[28] *See* PAUL R. ROSENBAUM, OBSERVATION & EXPERIMENT: AN INTRODUCTION TO CAUSAL INFERENCE 40–43, 350 (2017).

[29] *See* DEGROOT & SCHERVISH, *supra* note 22, at 539 ("In general, the *p-value* is the smallest level [of significance] α_0 such that we would reject the null-hypothesis at level α_0 with the observed data."); *see also* ROSENBAUM, *supra* note 28, at 43; CASELLA & BERGER, *supra* note 22, at 397.

[30] *See* COPE, *supra* note 22, at 41–42.

[31] *See* Megan L. Head et al., *The Extent and Consequences of P-Hacking in Science*, PLOS BIOLOGY 1 (Mar. 2015), https://doi.org/10.1371/journal.pbio.1002106 [https://perma.cc/T7CR-6KBB] ("Inflation bias, also known as 'p-hacking' or 'selective reporting,' . . . . occurs when researchers try out several statistical analyses and/or data eligibility specifications and then selectively report those that produce significant results.").

[32] Simmons et al. refer to these options as *researcher degrees of freedom*. *See* Simmons et al., *supra* note 6, at 1359 ("In the course of collecting and analyzing data, researchers have many decisions to make: Should more data be collected? Should some observations be excluded? Which conditions should be combined and which ones compared? Which control variables should be considered? Should specific measures be combined or transformed or both?").

methodology. Many of these methodological choices involve multiple reasonable options; therefore, there are countless *combinations* of reasonable options.

Assume, then, that there are $C_1 \ldots C_n$ reasonable methodological combinations from which a researcher can choose. Numerous combinations, $C_1 \ldots C_m$, can be expected to generate significant results. Many of these, however, $C_1 \ldots C_f$, may be false positives. A researcher engaged in data fishing (knowingly or naively) may search for the combinations, $C_1 \ldots C_m$, that produce significant results and selectively report them in isolation of the many tests that produce nonsignificant results. By doing so, the researcher misrepresents the results as being reliable with respect to some stated standard when in fact they are not.[33]

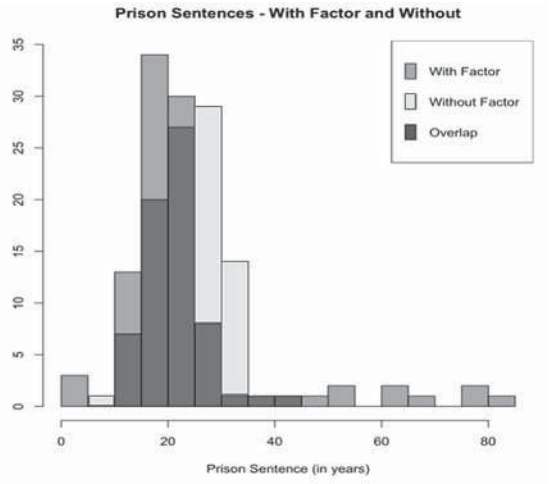
To illustrate, assume that a researcher wishes to show that a certain factor, "Factor X," reduces prison sentences for a particular category of serious felonies. She uses a dataset that contains prison-sentence data from two groups, one in which Factor X exists, and one in which it does not. The data is summarized in ***Table 1*** and ***Figure 1***.

Table 1. *Summary statistics for simulated prison-sentence data.*

Group	Sample Size	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
Factor	100	2	17	21	24	25	82
No Factor	100	8	20	24	24	29	42

[33] Note that a researcher may engage in data fishing without explicitly testing multiple combinations. The researcher does not need to conduct multiple tests explicitly to use the outcome data to develop her methodology. Indeed, "given a particular data set, it is not so difficult to look at the data and construct completely reasonable rules for data exclusion, coding, and data analysis that can lead to statistical significance—thus, the researcher needs only perform one test, but that test is conditional on the data." Andrew Gelman & Eric Loken, The Garden of Forking Paths: Why Multiple Comparisons Can Be a Problem, Even When There Is No "Fishing Expedition" or "P-Hacking" and the Research Hypothesis Was Posited Ahead of Time 3 (Nov. 14, 2013) (unpublished manuscript), http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf [https://perma.cc/YER7-ZJS7].

Figure 1. *Overlapping histograms for simulated prison-sentence data.*



Let us consider how the researcher might use data fishing to demonstrate to a reader that Factor X causes a reduction in prison sentences. First, she may perform a hypothesis test to determine whether the mean of the group with Factor X is significantly less than the mean of the group without Factor X (the "control" group). Performing a standard test called a t-test[34] on the data depicted in ***Figure 1*** results in a p-value of 0.965, which indicates no statistically significant difference under any standard level of significance. Disappointed, the researcher eyeballs the data and determines that the difference between the group means would be substantially wider if the few data points near the rightmost part of the histogram are not included in the analysis. The researcher, therefore, decides that these data points should be considered "outlier" sentences that should appropriately be "winsorized"—that is, reduced for purposes of the data analysis to be closer to the main body of the Factor X data.[35] In particular, the researcher winsorizes the data at the 97th percentile, thus reducing all data points above the 97th percentile down to the 97th percentile.

It is important to note that winsorization is not in and of itself an improper method. In fact, it is a standard method for accounting for outliers—outliers that may otherwise give rise to spurious results.[36] The problem, however, is that the researcher

---

[34] DEGROOT & SCHERVISH, *supra* note 22, at 576–85; COPE, *supra* note 22, at 36–41, 89–90.

[35] Winsorization is a statistical technique used to address outliers. *See* John W. Tukey, *The Future of Data Analysis*, 33 ANNALS MATHEMATICAL STAT. 1, 17–19 (1962).

[36] *See, e.g.*, *infra* notes 37–39 and accompanying text.

winsorizes for the purpose of achieving a significant result, not necessarily because it is appropriate for the circumstances at hand.

Once the data is winsorized, the researcher then performs a new hypothesis test comparing the means of the winsorized data. Performing a t-test on the same data, except now after winsorization, results in a p-value of 0.43. This p-value again indicates no statistically significant difference under any standard level of significance (e.g., 0.05, or even 0.1).

Frustrated, the researcher decides to winsorize further—first at the 95th percentile and then at the 90th percentile—and to rerun her hypothesis test. Using these winsorization schemes, she obtains p-values of 0.0419 and 0.0005, respectively. Both of these p-values are less than 0.05, indicating statistical significance at the 0.05 level. But, after realizing that 0.0419 will fall short of significance once she adjusts the p-value to account for disclosed multiple comparisons (a standard adjustment to prevent false positives associated with multiple tests), she chooses between (a) using a 0.1 level of significance (which is somewhat less standard than a 0.05 level) and winsorizing at the 95th percentile or (b) maintaining a 0.05 level of significance and winsorizing at the 90th percentile. She decides based on which option seems more plausible methodologically. For example, she may decide on the latter option, arguing that data falling above the 90th percentile constitute outliers. Or, she may choose the former option and argue that a 0.1 level of significance is appropriate. Either way, her argument for what is appropriate is heavily influenced by her interest in achieving a significant result.

Note that, in this example, I have focused on only two methodological factors—winsorization and levels of significance. In reality, however, the researcher may experiment with a very broad range of methodological factors. As in the example above—but on a far more extreme scale—she can then select one of the perhaps few methodological combinations that yield a significant result, present it in isolation of the other methodological combinations, and use it to support her claim—for example, that Factor X reduces prison sentences.

Let us consider a recent experiment in which a coauthor and I sought to test, among other things, the effect of certain evidence on the magnitude of awards for pain and suffering and punitive damages.[37] In designing the experiment, we considered various ways to define *magnitude*. For example, we could use the mean, the

[37] Hillel J. Bavli & Reagan Mozer, *The Effects of Comparable-Case Guidance on Awards for Pain and Suffering and Punitive Damages: Evidence from a Randomized Controlled Trial*, 37 YALE L. & POL'Y REV. 405 (2019) [hereinafter *The Effects of CCG*].

median, or the mean of the log-transformed data ("log mean"). As demonstrated in ***Table 2***, which is an excerpt from a table in *The Effects of Comparable-Case Guidance on Awards for Pain and Suffering and Punitive Damages* (*The Effects of CCG*) (although modified for clarity), the effect of the evidence on *magnitude* is highly sensitive to the way in which *magnitude* is defined.

Table 2. *Effect of certain evidence on magnitude under alternative definitions.*[38]

	Punitive Damages				Pain and Suffering			
	Log mean	*Median*	*Mean*	*Mean (winsor 90th)*	*Log mean*	*Median*	*Mean*	*Mean (winsor 90th)*
Treatment vs. Control	+	0	–	0	+	0	0	+

For example, comparing treatment (here, exposure to certain evidence) to control (no exposure to the evidence) while defining the magnitude of punitive damages awards using the mean and winsorizing at the 95th percentile (the default in that experiment) yields a negative effect of the treatment, whereas using the median yields no effect. Further, using the mean and winsorizing at the 90th percentile (rather than the 95th percentile) also yields no effect, whereas using the log mean (a standard measure for dollar values) and winsorizing at the 95th percentile yields a *positive* effect. Finally, comparing treatment to control using the pain and suffering data yields no negative effect for any of the associated combinations.[39]

It is important to realize that each of these combinations may be reasonable. A researcher can likely justify using any of them. Thus, it would be easy for a researcher to engage in data

[38] *Id.* at 445.

[39] A common method of data fishing involves experimenting with different possible ways of dividing a dataset and conducting numerous comparisons among the resulting subgroups. As Andrew Gelman and Eric Loken explain using an example involving research regarding response differences between Democrats and Republicans for purposes of proving the importance of context for understanding mathematical concepts,

> there is a huge number of possible comparisons that can be performed—all consistent with the data. For example, the pattern could be found (with statistical significance) among men and not among women—explicable under the theory that men are more ideological than women. Or the pattern could be found among women but not among men—explicable under the theory that women are more sensitive to context, compared to men. Or the pattern could be statistically significant for neither group, but the difference could be significant (still fitting the theory . . .). Or the effect might only appear among men who are being asked the questions by female interviewers.

Gelman & Loken, *supra* note 33, at 3.

fishing—that is, to test each combination to determine which of them provides the most favorable results, and then to use that combination and selectively report the favorable results in isolation of the others.

In *The Effects of CCG*, we presented the information in ***Table 2*** to explain and demonstrate the sensitivity of the results to the measure used to define *magnitude*, and we concluded that the effect at issue "is a question for future research."[40] If, however, a researcher sought to use this data to support a claim that the evidence reduces magnitude, increases magnitude, or has no effect on magnitude, data fishing would allow the researcher to cherry-pick the measure of magnitude that supports the claim and present that measure, and its accompanying result, in isolation of the other measures and associated results. For the reasons explained in the following Section, this practice would lead to incorrect results and false claims.

C. Why Data Fishing Produces Invalid Results

Why data fishing produces invalid results is not obvious. After all, if an explorer believes that a treasure exists in a sunken ship in a certain area of the ocean, and he funds an expedition to search for and recover it, his exploration would not invalidate the find. So why should similar exploration in a dataset invalidate a statistical find? The operative distinction between treasure hunting and data fishing is that data fishing produces results—and ultimately misleading results—by exploiting randomness due to sampling.

Data fishing gives rise to a number of concerns. In this article, I focus on two major concerns in particular. First and foremost, data fishing causes false positives—that is, false findings that something is true when in fact it is not. As I explain below, this occurs because data fishing involves undisclosed "multiple comparisons" and "overfitting." Second, and also important, data fishing causes false impressions—specifically, incorrect perceptions that a researcher's results are more robust and replicable than they in fact are. Let us examine each of these concerns.

1. False Positives

Data fishing causes false positives due to undisclosed and unaddressed "multiple comparisons" and due to "overfitting."

---

[40] *The Effects of CCG*, *supra* note 37, at 455.

When a researcher chooses a 0.05 level of significance, she is accepting a degree of error. At this level of significance, the researcher rejects the null hypothesis if there is sufficient evidence that only 5% of the time, she will reject the null hypothesis when it is in fact true. This means that, at this level of significance, one out of twenty hypothesis tests (involving a true null hypothesis) will result in a false positive, where the pattern detected is due to sampling randomness rather than a true characteristic of the population. At significance levels of 0.01 or 0.1, one out of one hundred or one out of ten hypothesis tests (involving a true null hypothesis) will result in a false positive, respectively. Thus, as a researcher performs more hypothesis tests, the likelihood of obtaining a false positive increases. This problem is known as the multiple-comparisons problem.

As a consequence of the multiple-comparisons problem, data fishing increases a study's rate of false positives. This is because data fishing involves searching the data for a methodology that will yield favorable results. If a researcher applies, for example, a 0.05 level of significance but engages in data fishing by searching for and selectively reporting significant results, then the researcher will have a false-positive rate higher, and often substantially higher, than 0.05. For instance, if a researcher performs twenty hypothesis tests at a 0.05 level of significance, the likelihood of at least one false positive is approximately 64%.[41] Similarly, if she conducts one hundred hypothesis tests, and in these tests the null hypotheses are in fact true, then, on average, the researcher will obtain five significant results based on chance alone. If the researcher is simply cherry-picking significant results and reporting them as significant with a 0.05 level of significance, it is possible that all or most of these results are false positives. In any event, the researcher is reporting a likelihood of type I error of 0.05 when, in fact, it is substantially higher. Similarly, if the researcher is seeking a particular result and tests numerous methodologies until one "works," this result has a high likelihood of being a false positive.

There are well-established methods in statistics for addressing the multiple-comparisons problem. These methods adjust a hypothesis test to account for the increased risk of false positives associated with multiple comparisons. However, data fishing generally involves *undisclosed* exploration for significant results, and therefore *undisclosed* multiple comparisons.[42] As such, it is rare that that a researcher will address this problem by

---

[41] STEPHEN B. HULLEY ET AL., DESIGNING CLINICAL RESEARCH 59 (3d ed. 2007).

[42] *See supra* Section I.B.

applying a multiple-comparisons adjustment. Consequently, the rate of false positives is likely to be far higher than that reported. Thus, data fishing involves multiple comparisons and consequently a high rate of false-positive results.

A closely related way in which data fishing causes false positives is due to a problem known as "overfitting." In general, overfitting occurs when a statistical model incorrectly interprets randomness due to sampling as a pattern or true characteristic of the population of interest.[43] Overfitting is a problem because

> our standard goal in statistical modeling is to develop a model that can capably generalize to new observations similar, but not identical, to the ones we have sampled. We generally do not care very much about how well we can predict scores for the observations in our existing sample, since we already *know* what those scores are. In this sense, the prediction error that we compute when we fit a model on a particular dataset is only a proxy for the quantity we truly care about, which is the error term that we would obtain if we were to apply our . . . model to an entirely new set of observations sampled from the same population.[44]

When a researcher explores data and attempts to find patterns by examining different combinations of methodological features, she essentially makes assumptions about the data, or places constraints on it. In essence, she develops a model to explain it. This occurs both with data fishing and with innocent exploration. For example, when the researcher decides to omit certain outliers, she makes assumptions regarding the role that the outliers play in the data. Implicitly, she has a theory for why the outliers can be omitted when performing her analysis. This model, in a sense, constrains the data. The more constraints the researcher puts on the data, the better she will be able to explain the patterns in the data through her model, up to the point that her model is so complex that it simply explains individually each point of the data. Such a model is useless for inference, however, because it is not generalizable to other datasets—it is so constrained, so complex, that it can only explain that particular dataset. This is the idea of overfitting.

Figure 2 illustrates this concept in the regression context by comparing an appropriate linear regression model for the data (the straight line) to an overfit model (the curved line). The overfit model perfectly fits the data in this sample; but applying it to a new sample from the same population would lead to a high level of error relative to the straight line. It is too specific to this dataset: it interprets

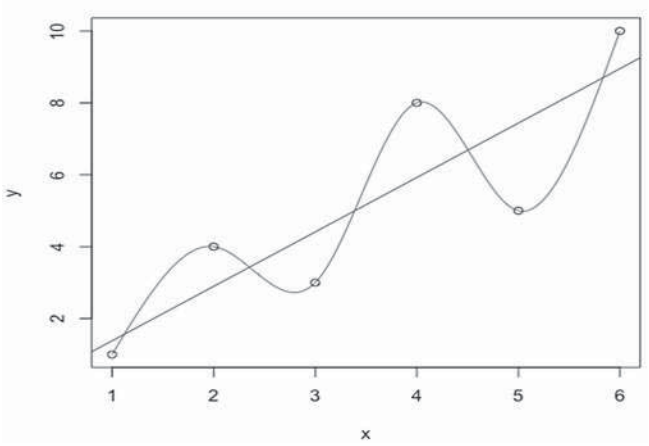
---

[43] *See* Tal Yarkoni & Jacob Westfall, *Choosing Prediction Over Explanation in Psychology: Lessons from Machine Learning*, 12 PERSPS. ON PSYCH. SCI. 1100, 1102 (2017) ("The tendency for statistical models to mistakenly fit sample-specific noise as if it were signal is commonly referred to as *overfitting*.").

[44] *Id.*

randomness (from sampling) as characteristics of the population and incorporates this randomness into the model as true effects. But a second sample would involve different data points (due to sampling variation) and this model would be inappropriate as applied to the new data.

Figure 2. *Comparison between an appropriate model (straight line) for the data and an overfit model (curved line).*



Data fishing can therefore lead to, or can be described as a form of, overfitting.[45] In this context, overfitting (a concept borrowed from the machine-learning literature) means that the researcher imposes too many specifications on a hypothesis test and mistakenly interprets a significant result as a true pattern associated with the population when it is in fact just due to randomness associated with the particular sample examined.

When a researcher searches for a combination of methodological features that will generate a significant result, she effectively builds a study's methodology to fit the sample data (in the sense of finding patterns in the specific sample of data examined). This causes a high likelihood that the detected pattern (indicated by statistical significance) is specific to the sample data rather than a true signal—that is, rather than a characteristic of the population. In other words, the study's methodology is likely to detect noise due to sampling rather than a pattern that is replicable in other samples.[46]

As Yarkoni and Westfall put it, "*p*-hacking can be usefully conceptualized as a special case of overfitting. Specifically, it can be

[45] *See id.* at 1104.
[46] *See id.* at 1102–04.

thought of as a form of *procedural overfitting* that takes place prior to (or in parallel with) model estimation—for example, during data cleaning, model selection, or choosing which analyses to report."[47] They explain:

> Every pattern that could be observed in a given dataset reflects some (generally unknown) combination of signal and error. The more flexible a statistical model or human investigator is willing to be—that is, the wider the range of patterns they are willing to "see" in the data—the greater the risk of hallucinating a pattern that is not there at all. . . . [A] procedurally overfitted or *p*-hacked analysis will often tell an interesting story that appears to fit the data exceptionally well in an initial sample but cannot be corroborated in future samples. . . . [T]he culprit is unrestrained flexibility—in this case, in the data analysis and interpretation of results . . . .[48]

In summary, data fishing leads to major problems involving undisclosed multiple comparisons and overfitting, and thereby causes false-positive results. Indeed, even a low level of flexibility in selecting combinations of methodological factors based on outcome data can produce false positives at a rate that is unacceptable and altogether invalidates a study's results. In a 2011 study, for example, Simmons et al. showed that even just a few methodological options (in Yarkoni and Westfall's words, just "a moderate amount of flexibility in analysis choice"[49]) "would lead to a stunning 61% false-positive rate!"[50] They emphasize that, with just "four common researcher degrees of freedom," "[a] researcher is more likely than not to falsely detect a significant effect."[51] Furthermore, the authors explain that, "[a]s high as these estimates are, they may actually be conservative."[52] After all, the authors "did not consider many other degrees of freedom that researchers commonly use."[53] These include:

> testing and choosing among more than two dependent variables (and the various ways to combine them), testing and choosing among more than one covariate (and the various ways to combine them), excluding subsets of participants or trials, flexibility in deciding whether early data were part of a pilot study or part of the experiment proper, and so on.[54]

---

[47] *Id.* at 1104. Although I draw on this idea of "procedural overfitting," for simplicity, throughout the current article, I use the more general term "overfitting." *Id.*

[48] *Id.*

[49] *Id.* at 1103–04.

[50] Simmons et al., *supra* note 6, at 1361.

[51] *Id.* (using the following points of flexibility: "flexibility in analyzing two dependent variables"; flexibility in "collect[ing] 10 additional observations per condition"; "flexibility in controlling for gender or for an interaction between gender and the independent variable"; and flexibility in "dropping (or not dropping) one of three conditions").

[52] *Id.*

[53] *Id.*

[54] *Id.*

Other studies have found similar results.[55]

2. False Impressions

In addition to causing false-positive results, data fishing misleads readers to attach more importance to the findings than they deserve. In particular, even if a result is not a false positive—i.e., it is in fact a characteristic of the population—it may lack robustness even to minor changes in methodology.

Distinguish between two concepts: replicability across different samples of data and replicability across different reasonable methodologies. The former concept, which, for purposes of this article, I call *sample replicability*, refers to the consistency of results when a researcher applies the same methodology as in an initial study but on a new sample of data. If a result is a false positive, it is a feature of the sample examined but is not sample replicable. The latter concept, which, for purposes of this article, I call *method replicability*, refers to the consistency of results when a new researcher applies a similar but not identical methodology as in the initial study using the same or a new sample of data.[56] Sample replicability relates to the robustness of the results to different samples while method replicability relates to the robustness of the results to different methodologies.[57] Both types of

---

[55] *See* Yarkoni & Westfall, *supra* note 43, at 1104 (citing Michael J. Strube, *SNOOP: A Program for Demonstrating Consequences of Premature and Repeated Null Hypothesis Testing*, 38 BEHAV. RSCH. METHODS 24 (2006)).

[56] My use of the language "similar but not identical" is intended to be consistent with the NASEM Report's use of the language "sufficiently similar conditions" in its description of assessing replicability. *See* NASEM REPORT, *supra* note 13, at 73 (citing Florian Cova et al., *Estimating the Reproducibility of Experimental Philosophy*, 12 REV. PHIL. & PSYCH. 9 (2018), in relation to the meaning of "sufficiently similar").

[57] There has been considerable confusion, controversy, and inconsistency surrounding the terms repeatability, replicability, and reproducibility. *See* NASEM REPORT, *supra* note 13, at 42–44 (describing inconsistency and confusion regarding the terms reproducibility and replicability); Hans E. Plesser, *Reproducibility vs. Replicability: A Brief History of a Confused Terminology*, 11 FRONTIERS NEUROINFORMATICS 1, 1–3 (2018). NASEM addresses this confusion and inconsistency in the NASEM Report, which was funded by the National Science Foundation and sponsored by the Alfred P. Sloan Foundation. The NASEM Report adopts the following terminology: "*Reproducibility* means obtaining consistent computational results using the same input data, computational steps, methods, code, and conditions of analysis. *Replicability* means obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data." News Release, National Academies of Sciences, Engineering, & Medicine, New Report Examines Reproducibility and Replicability in Science, Recommends Ways to Improve Transparency and Rigor in Research (Apr. 7, 2019), https://www.nationalacademies.org/news/2019/05/new-report-examines-reproducibility-and-replicability-in-science-recommends-ways-to-improve-transparency-and-rigor-in-research [https://perma.cc/LEM6-PGJX]; *see also* NASEM REPORT, *supra* note 13, at 6.

The concepts defined in the text of the current article—sample replicability and method replicability—are specific types of replicability within the definition of that concept adopted by the NASEM Report. *See* NASEM REPORT, *supra* note 13, at 85 ("In some cases,

replicability are fundamental to a good study. As emphasized above in the discussion of false positives, results that are not sample replicable are not useful. In this discussion of false impressions, our focus is on method replicability: the importance of a statistical result relies on a reasonable degree of robustness to alternative methodologies, and data fishing misleads the reader to believe that a result is more robust than it actually is.

Consider again the results in ***Table 2***. Assume that a researcher wishes to prove that applying some intervention, or "treatment," causes a reduction in damages. The researcher has discretion over a number of design options: she may focus on awards for pain and suffering or punitive damages; she may measure central tendency using the log mean, median, or mean; and she may winsorize at the 90th percentile or the 95th percentile (the default in that study). In actuality, the researcher would have discretion over numerous other factors and would have more options for each factor. For simplicity, however, assume that these are the researcher's only points of flexibility. Assume also that the researcher engages in data fishing. She performs hypothesis tests for each combination of these possibilities and draws up, in her private notes, the table presented in ***Table 2***. She then identifies the one combination—punitive damages, mean, and winsorization at the 95th percentile—that achieves her sought-after result, a significant negative effect. Finally, she reports this effect to readers in isolation of the other tests and uses it to support her claim that this treatment causes a reduction in damages.

Because this result is the product of data fishing, it has a relatively high likelihood of being a false positive. Let us assume, however, that it is not a false positive—given the methodology, it is a true characteristic of the population of interest. If this exact methodology (using the mean, winsorizing at the 95th percentile, and using the punitive damages data) is performed using a new sample, the researcher is likely to obtain the same result.

On the other hand, regardless of whether the result is a false positive, it is misleading in that it is extremely sensitive to

---

non-replicability arises from the inherent characteristics of the systems under study. In others, decisions made by a researcher or researchers in study execution that reasonably differ from the original study such as judgment calls on data cleaning or selection of parameter values within a model may also result in non-replication. Other sources of non-replicability arise from conscious or unconscious bias in reporting, mistakes and errors (including misuse of statistical methods), and problems in study design, execution, or interpretation in either the original study or the replication attempt."). By using these terms, I do not mean to add to the plethora of terms in the statistics literature or to the confusion surrounding the concepts of reproducibility and replicability. My intention is only to clarify an important conceptual distinction relevant to this article's discussion.

changes in methodology. It is not method replicable. In particular, the reader interprets the result as more robust than it is, because she justifiably assumes that the researcher designed the study neutrally based on what was most appropriate for the study and not based on a search for which methodology yields a result that most favors the researcher's claim. The reader is not aware, and understandably does not assume, that the researcher selected, *post hoc* and for the purpose of supporting her claim, the one methodological combination (out of eight) that yields a favorable result. The reader's assumptions are justified by the fact that, as shown in Section I.B, a researcher can usually arrive at whatever result she seeks if she can cherry-pick a methodology based on which one yields the most favorable result. In other words, the reader's assumptions are justified by the implicit representation that the study is confirmatory rather than exploratory.

One way of understanding the reader's assumptions and the ensuing problem is as follows: When a researcher states that she arrived at a significant result winsorizing at the 95th percentile and measuring central tendency using the mean—not an unreasonable combination in and of itself—the reader reasonably assumes some degree of robustness to design modification (e.g., measuring central tendency using the median instead of the mean). This assumption arises from the reasonable expectation that the researcher made these choices neutrally, based on her interest in using an appropriate design rather than an interest in achieving a specific result. The researcher's use of data fishing in a sense exploits this assumption and misleads the reader to overvalue the researcher's results.

Again, a researcher may engage in data fishing for a wide variety of reasons. Although it is easy to imagine a villainous researcher, motivated by greed or self-advancement, sitting at a desk and aggressively testing methodology after methodology to find and selectively report significant results (or better, programming a computer to test all such methodologies), it is likely that most data fishing occurs without the researcher's knowledge or understanding of the practice or its effects, or otherwise without ill intention. In any event, by causing false positives and false impressions, data fishing can be very harmful.

D. The Prevalence of Data Fishing in Empirical Legal Scholarship

There is a dearth of formal empirical research regarding the prevalence of data fishing in law.[58] Nevertheless, the prevalence of this practice can easily be inferred based on (1) strong empirical evidence of the prevalence of data fishing in the natural and social sciences; and (2) empirical evidence of the "deeply flawed" "state of empirical legal scholarship,"[59] as well as the absence of attentiveness to data fishing and certain other methodological issues, or of safeguards against data fishing that are common in other fields. Additionally, as explained below, the precise prevalence of data fishing in law is not critical. Rather, the mere *opportunity* to engage in this practice is often sufficient to cause substantial harm.

First, there is substantial evidence from other fields, including the natural and social sciences, that data fishing is extremely common, and there is little reason to believe that it is not as prevalent in legal scholarship as in other areas of the social sciences.[60] To the contrary, legal scholarship is behind many other fields in its attentiveness to data fishing and methodology in general.[61]

In a 2012 study, John et al. surveyed over two thousand psychologists regarding their engagement in "questionable research practices."[62] Among those psychologists, 63.4% admitted to "failing to report all of a study's dependent measures," 55.9% of participants in the study admitted to "[d]eciding whether to collect more data after looking to see whether the results were significant," 45.8% admitted to "selectively reporting [in a paper] studies that 'worked'," 38.2% admitted to "[d]eciding whether to exclude data after looking at the impact of doing so on the results," and 27% admitted to "reporting [in a paper] an unexpected finding as having been predicted from the start."[63] Over 90% of participants "admitted to having engaged in at least one [questionable research practice]."[64] Further, because the study involved self-admission,

---

[58] An empirical study by Lee Epstein and Gary King found "deep[]" methodological problems "[e]verywhere" in empirical legal scholarship, but, while the authors discuss the invalidity of results produced by data fishing, they do not explicitly address the prevalence of this practice in empirical legal scholarship. *See* Epstein & King, *supra* note 1, at 6, 15, 54–55.

[59] *See id.* at 6; *infra* notes 72–73 and accompanying text.

[60] *See* NASEM REPORT, *supra* note 13, at 76–85.

[61] *See* Epstein & King, *supra* note 1, at 6.

[62] Leslie K. John et al., *Measuring the Prevalence of Questionable Research Practices with Incentives for Truth Telling*, 23 PSYCH. SCI. 524 (2012).

[63] *Id.* at 525.

[64] *Id.* at 527.

these prevalence estimates are likely to be grossly underestimated.[65] Indeed, prevalence estimates based on other measures used in the study are substantially higher.[66] According to the authors:

> One would infer from the [data] that nearly 1 in 10 research psychologists has introduced false data into the scientific record and that the majority of research psychologists have engaged in practices such as selective reporting of studies, not reporting all dependent measures, collecting more data after determining whether the results were significant, reporting unexpected findings as having been predicted, and excluding data post hoc.[67]

In another recent survey study (involving 807 ecologists and evolutionary biologists), Fraser et al. found that "64% reported cherry picking statistically significant results in at least one publication; 42% reported *p* hacking by collecting more data after first checking the statistical significance of results, and 51% acknowledged reporting an unexpected finding as though it had been hypothesized from the start."[68]

In short, the prevalence of data fishing has been repeatedly tested and confirmed in wide-ranging fields across the natural and social sciences.[69]

Second, there is good reason to infer that empirical legal research, which frequently draws on other fields in the social sciences, such as economics and psychology, also suffers from very high rates of data fishing.[70] In fact, based on evidence that

---

[65] *Id.* at 526.

[66] *Id.* at 526–27.

[67] *Id.* (cross-references omitted).

[68] Hannah Fraser et al., *Questionable Research Practices in Ecology and Evolution*, PLOS ONE 9 (July 16, 2018), https://doi.org/10.1371/journal.pone.0200303 [https://perma.cc/FYC2-H8PY].

[69] *See id.*; John et al., *supra* note 62, at 524 (citing studies, highlighting prevalence of "questionable research practices," and reporting results of survey suggesting that "some questionable practices may constitute the prevailing research norm"); John P.A. Ioannidis, *Why Most Discovered True Associations Are Inflated*, 19 EPIDEMIOLOGY 640, 640–43 (2008) (highlighting that "[s]elective analyses and outcome reporting have been extensively demonstrated in clinical-trials research comparing protocols against reported results" and tabulating "articles suggesting that early studies give (on average) inflated estimates of effect" across wide-ranging research fields); Yarkoni & Westfall, *supra* note 43, at 1100–02, 1103–04 (highlighting prevalence of data fishing and citing studies); Andrew Gelman, *Too Good to Be True*, SLATE (July 24, 2013), https://slate.com/technology/2013/07/statistics-and-psychology-multiple-comparisons-give-spurious-results.html [https://perma.cc/R6UL-PPGK] (emphasizing that data fishing "happens all the time," is "standard practice[ ]," and is "considered acceptable"). *But see* NASEM REPORT, *supra* note 13, at 97 (highlighting "methodological shortcomings" in "quantitative assessment[s]" of "the prevalence of such inappropriate statistical practices as *p*-hacking, cherry picking, and hypothesizing after results are known").

[70] While there is an absence of formal statistical examinations of the prevalence of data fishing in empirical legal scholarship, the issue, as it pertains to law, has caught the attention of various scholars. *See, e.g.*, Mitchell, *supra* note 1, at 167–79 ("consider[ing] how the scientific status of empirical legal scholarship might be enhanced" and recommending

empirical legal scholarship often fails to adhere to well-accepted methodological practices in the natural and social sciences, and based on the absence of certain safeguards against data fishing that are common in other fields, there is cause to believe that data fishing is even more prevalent in law than in other fields.[71]

In 2001, Epstein and King conducted an empirical review of legal scholarship and found "serious problems of inference and methodology abound everywhere [that they found] empirical research in the law reviews and in articles written by members of the legal community."[72] Based on this study, they concluded that "the current state of empirical legal scholarship is deeply flawed" and that there is "little awareness of, much less compliance with, the rules of inference that guide empirical research in the social and natural sciences."[73] As discussed above, however, data fishing is prevalent even in fields in the natural and social sciences that are particularly attentive to methodological challenges; it is therefore likely that this practice is all the more prevalent in legal scholarship, in which there is arguably less awareness of, and at least less attentiveness to, methodological issues relative to other fields.

Furthermore, in empirical legal scholarship, there is a noticeable absence of safeguards against data fishing (and certain other undesirable methodological practices) common in other fields. In some research fields, such as biomedicine, journals require, or at least encourage, various procedures and documentation aimed at protecting against data fishing and other

---

"stringent disclosure requirements designed to foster critical review and replication of empirical legal research"); Ho & Rubin, *supra* note 17, at 17–19 (explaining "advances toward credible causal inference that have wide application for empirical legal studies," and highlighting the principle that "[r]esearch design trumps methods of analysis"); Mark Klock, *Finding Random Coincidences While Searching for the Holy Writ of Truth: Specification Searches in Law and Public Policy or Cum Hoc Ergo Propter Hoc?*, 2001 WIS. L. REV. 1007 (2001) (discussing "specification searches" in law and "provid[ing] examples and case studies"). *See generally* Epstein & King, *supra* note 1 (addressing the poor quality of empirical research in legal scholarship).

[71] Epstein & King, *supra* note 1, at 5–6.

[72] *Id.* at 15. It is important to emphasize that there has been substantial growth in attention to empirical legal research, and research methodology in particular, in recent years. This includes the founding of the Society for Empirical Legal Studies (SELS), the annual Conference on Empirical Legal Studies, and the Journal of Empirical Legal Studies—a peer-review journal devoted to empirical studies related to law. *See About SELS*, *Society for Empirical Legal Studies*, CORNELL L. SCH., https://community.lawschool.cornell.edu/sels/about-sels/ [https://perma.cc/XV53-JHN7]; *About the Journal*, *Journal of Empirical Legal Studies*, WILEY ONLINE LIBR., https://onlinelibrary.wiley.com/journal/17401461 [https://perma.cc/USP2-KN6S]. The SELS community and a number of others have made significant progress and have elevated the credibility of empirical legal research; however, there is still a long way to go, and even the total scholarship arising from these few research communities constitutes only a relatively small proportion of empirical legal scholarship.

[73] Epstein & King, *supra* note 1, at 6 (also remarking that "[t]he sustained, self-conscious attention to the methodology of empirical analysis so present in the journals in traditional academic fields . . . is virtually nonexistent in the nation's law reviews").

harmful research practices. These include signing ethics statements, submitting protocols and analysis plans with the manuscript, agreeing to share data, "preregistering" design and analysis plans at an early stage of a researcher's study, and others.[74] These requirements and practices do not exist in law, or they are at least very rare.[75] To the contrary, empirical research in law falls behind many other fields in its attentiveness to data fishing and other methodological issues.[76] This is reflected in Epstein and King's conclusion that empirical legal scholarship often fails to follow "rules of inference that guide empirical research" in other fields.[77] Arguably, this problem is only exacerbated by the absence of peer review in many or most article-selection processes in legal scholarship.[78]

---

[74] *See, e.g.*, *New Manuscripts, Statistical Reporting Guidelines*, NEW ENG. J. MED., https://www.nejm.org/author-center/new-manuscripts [https://perma.cc/B83B-PAP4] (recommending the following procedures, among others, and presumably accounting for them during the submission and review process: submission of "final protocols and statistical analysis plans (SAPs) . . . with the manuscript, as well as a table of amendments made to the protocol and SAP indicating the date of the change and its content"; ensuring that primary outcome analysis matches prespecified analysis in protocol (and providing justification when a deviation occurs); submission of "a signed and dated version" of the study's "prespecified SAP with a description of hypotheses to be tested" for an observational study if the study included such a plan; use of prespecified multiple-comparisons adjustment methods; depositing of SAPs in an online repository ("encourage[d]"); use of retesting and robustness checks ("encouraged")); David Harrington et al., *New Guidelines for Statistical Reporting in the* Journal, 381 NEW ENG. J. MED. 285 (2019) (commenting on new guidelines and requirements for reporting statistical results to better account for "error that can result from uncritical interpretation of multiple inferences," including the possibility of unreported comparisons). *See generally* Chris Allen & David MA Mehler, *Open Science Challenges, Benefits and Tips in Early Career and Beyond* (Oct. 17, 2018), http://psyarxiv.com/3czyt [https://perma.cc/WSU9-YT54] (discussing benefits and costs of "open science" methods for improving the reliability of empirical research); Epstein & King, *supra* note 1, at 46 (describing requirements for publishing in a "leading empirical methods journal," including the requirement that authors "indicate in their first footnote in which public archive readers can find the data, programs, recodes, or other information necessary to replicate the numerical results in their article"). Note that certain types of clinical trials outside of law must, *by law*, be preregistered. *See* 42 C.F.R. § 11.22.

[75] It is true that research involving human subjects, in law, as in other fields, often requires that a researcher submit a detailed design for approval by an institutional review board (IRB) prior to obtaining or generating data. *See* HULLEY ET AL., *supra* note 41, at 227. These procedures may reduce the researcher's methodological flexibility and at least motivate the researcher to consider and record various methodological decisions prior to observing any outcome data. But IRB approval is frequently not required for empirical studies in legal scholarship, *see id.* at 227–28, and, although empirical research in law can involve human subjects, human-subjects research is more common in other fields, such as medicine and psychology. In any event, although IRB approval requires and motivates some methodological prespecification, it rarely requires precise details of a researcher's intended statistical analysis—at least not to the level of detail that would prevent data fishing. *See id.* at 227.

[76] *See* Epstein & King, *supra* note 1, at 6; *see also* Zeiler, *supra* note 1, at 78, 78–86.

[77] Epstein & King, *supra* note 1, at 6.

[78] It has been suggested that the prevalence of methodological problems may be greater in *student-edited* law reviews—the central forum for legal scholarship—than in peer-review journals central to scholarship in other fields. *See* Zeiler, *supra* note 1, at 78–79. Epstein and King highlight the "astonish[ment]" that scholars in fields outside of

Finally, regardless of the precise prevalence of data fishing in legal scholarship, there is at least an enormous *opportunity* for researchers to engage in and report results based on data fishing. This opportunity is itself sufficient to cause far-reaching harm.

For example, a reader's awareness that data fishing is possible and that the researcher may have an interest in engaging in data fishing is sufficient to cause the reader to distrust a study's results and the researcher's claims, even if the researcher's methodology seems perfectly sound. Moreover, this environment—in which researchers have the opportunity to engage in data fishing without detection by readers—further incentivizes researchers to engage in this practice. A researcher may assume that other researchers are engaging in data fishing and that not engaging in it will put the researcher at a disadvantage—for example, in asserting arguments against other researchers or simply in advancing her career. Additionally, a researcher may believe that data fishing would create certain advantages, even if other researchers are not engaged in it. The researcher also knows that data fishing is relatively unverifiable, and there is no expectation for her to show that she has not engaged in it. In this environment, a prisoner's dilemma may emerge, causing researchers to engage in data fishing and readers to distrust empirical scholarship.

To be sure, the prisoner's dilemma model perhaps oversimplifies what is in fact occurring in empirical legal scholarship. For example, readers are not entirely blind to data fishing. They are also not entirely blind to a researcher's engagement in good statistical practices, thus allowing for researchers to develop reputations for good empirical scholarship and to have the incentive to develop such reputations. The point is, however, that knowing the precise prevalence of data fishing in legal scholarship is not necessary for our purposes. The *opportunity* for researchers to engage in data fishing alone is sufficient to cause very substantial harm. Therefore, even without empirical certainty regarding the prevalence of data fishing in law, eliminating the practice is crucial. As I discuss below, this can be accomplished by implementing a few simple steps.

E. The Prevalence of Data Fishing in Litigation

Data fishing is also prevalent in the courtroom. In litigation, and expert evidence in particular, data fishing is often more overt. It exacerbates the "hired-gun" problem—in which experts arrive at

---

law feel when they discover that acceptance decisions at top law journals are made by students. Epstein & King, *supra* note 1, at 48.

opinions based on who hires them rather than based on the truth—and leads to an environment in which opposing experts battle over methodology in an attempt to produce favorable testimony for their respective litigant sponsors and in which the factfinder justifiably distrusts expert testimony and often chooses a winner based on inappropriate criteria.[79] As in legal scholarship, the precise scale and scope of data fishing in litigation is unknown. But, in this context also, the prevalence of data fishing can easily be inferred. Furthermore, the opportunity for data fishing itself is often sufficient to cause far-reaching harm.

There are, in theory, safeguards in litigation against data fishing, but they rarely stop it from occurring. Federal Rule of Evidence 702 requires that an expert's testimony be "the product of reliable principles and methods" and that "the expert has reliably applied the principles and methods to the facts of the case."[80] The standard established in *Daubert v. Merrell Dow Pharmaceuticals, Inc.* requires that the trial court fulfill a "gatekeeping role" to "ensur[e] that an expert's testimony both rests on a reliable foundation and is relevant to the task at hand."[81] The court must ensure that "the testimony is based on sufficient facts or data," that it "is the product of reliable principles and methods," and that "the expert has reliably applied the principles and methods to the facts of the case."[82] And the court must exclude testimony that is not "based on scientifically valid principles."[83]

Additionally, a party may challenge an expert's data-fishing practices on cross examination or may present to the jury evidence of the invalidity of an expert's results.[84] Arguably, therefore, "the Federal Rules of Civil Procedure and of Evidence work together to enable a well-prepared party to punish an adversary for its expert's use of data [fishing]."[85]

The problem is that these tools are generally not effective in eliminating data fishing in the courtroom, except perhaps (and only sometimes) in its most extreme form. Some courts have excluded evidence based on data fishing, but most have not,

---

[79] *See* Robertson, *supra* note 4, at 184–92; Jonah B. Gelbach, *Expert Mining and Required Disclosure*, 81 U. CHI. L. REV. 131, 135–44 (2014); Richard A. Posner, *An Economic Approach to the Law of Evidence*, 51 STAN. L. REV. 1477, 1535–36 (1999).

[80] FED. R. EVID. 702.

[81] Daubert v. Merrell Dow Pharms., Inc., 509 U.S. 579, 597 (1993).

[82] FED. R. EVID. 702.

[83] *Daubert*, 509 U.S. at 597.

[84] *See* Gelbach, *supra* note 79, at 135–36.

[85] *Id.* at 131, 136 (arguing that "[v]arious aspects of evidence and civil-procedure law disincentivize data mining by expert witnesses in federal civil litigation," but that *expert* mining—"hiring multiple experts, asking each to provide an expert report on the same issue, and then put[ting] on the stand only the one who provides the most favorable report"—has the same effect as data mining).

and it is difficult to say whether a jury discounts an expert's testimony based on a data-fishing challenge, or if they even understand the concept of data fishing in the first instance.[86] In any event, it is rare for an expert to provide affirmative evidence showing that she safeguarded against, or at least did not engage in, data fishing.

In general, experts at least have the opportunity to engage in data fishing by exploring data for results that will best support their sponsors' positions and selectively presenting their findings as the products of confirmatory analysis. As a result, many cases devolve into a "battle of the experts" in which opposing experts debate methodology after potentially engaging in such exploration to find the best methodological combinations to support their respective sponsors.[87] For example, in a discrimination case, opposing experts often disagree on whether to include a particular variable in a regression model. Commonly, such disagreements do not arise from, e.g., differences in methodological schools; rather, they often arise because the experts (or nontestifying experts hired by the litigants) have previously explored the data and determined which models most favor their respective sponsors' legal positions.[88]

One way in which a litigant may be able to engage in data fishing while bypassing scrutiny under Rule 702 and *Daubert*, and on cross-examination, is by hiring a nontestifying expert and taking care to stay within the protection of the attorney-work-product privilege. While testifying experts may be required to disclose analysis on which they relied in forming their opinions, analyses by nontestifying experts are generally protected from disclosure.[89] Therefore, a litigant may develop

---

[86] *See, e.g.*, *In re* Roundup Prods. Liab. Litig., 390 F. Supp. 3d 1102, 1137 (N.D. Cal. 2018) (rejecting defendant's argument that expert evidence should be inadmissible based on the expert's "engage[ment] in 'p-hacking' [i.e., data fishing], manipulation of data to obtain statistically significant results"); Karlo v. Pittsburgh Glass Works, LLC, 849 F.3d 61, 82 (3d Cir. 2017) (rejecting the district court's determination that an expert engaged in a "sort of subgrouping 'analysis'" that constituted "data-snooping [i.e., data fishing], plain and simple" without any "generally accepted statistical procedures . . . to correct his results for the likelihood of a false indication of significance," and concluding that "the [d]istrict [c]ourt applied an incorrectly rigorous standard for reliability" (quoting Karlo v. Pittsburgh Glass Works, LLC, No. 2:10-cv-1283, 2015 WL 4232600, at *13 (W.D. Pa. July 13, 2015))). *But see* Ohio Pub. Emps. Ret. Sys. v. Fed. Home Loan Mortg. Co., No. 4:08-cv-0160, 2018 WL 3861840, at *7 (N.D. Ohio Aug. 14, 2018) (holding, based on precedent, that an expert's choice of date for an "event study was entirely improper because you are supposed to hypothesize and then see your results").

[87] *See generally* Gelbach, *supra* note 79, at 131–34 (discussing "expert mining" as analogous to data mining).

[88] *See infra* Part IV.

[89] *See* FED. R. CIV. P. 26(b)(4)(D) ("Ordinarily, a party may not, by interrogatories or deposition, discover facts known or opinions held by an expert who has been retained or specially employed by another party in anticipation of litigation or to

statistical evidence by (1) hiring a nontestifying expert to perform exploratory analysis on a dataset pertinent to the case and to identify a methodology (including a model) that is favorable to the litigant's position; and then (2) hiring a separate testifying expert to test hypotheses using the methodology identified by the nontestifying expert and to testify regarding the results. The litigant must be careful to ensure that the nontestifying expert's analysis remains within the protection of the work-product privilege—for example, avoiding reliance by the testifying expert on the nontestifying expert's analysis. But, in general, the protections established for nontestifying experts will allow a litigant to explore the data, form a theory and methodology around the results obtained from such exploration, and then hire a testifying expert to testify regarding these results while excluding any information regarding the nontestifying expert's analysis.[90] This way, the testifying expert can assert on cross-examination, for example, that she did not engage in data fishing, and that, to the contrary, she only tested a particular hypothesis using a particular methodology.[91]

Even when this method is not used, it would not be uncommon for an expert to testify regarding statistical results produced by data fishing. Data fishing is prevalent and poorly understood, and courts often require an extreme case of data fishing to exclude the evidence under Rule 702 and *Daubert*.[92] Additionally,

---

prepare for trial and who is not expected to be called as a witness at trial."). *See generally* Hickman v. Taylor, 329 U.S. 495 (1947) (attorney work product); FED. R. EVID. 705.

[90] A litigant can likely convey information to a testifying expert, or at least bias a testifying expert in the direction of one methodology or another based on a nontestifying expert's analysis, while avoiding direct reliance on it by the testifying expert. *See generally* Robertson, *supra* note 4, at 185–86 (discussing "psychological heuristics" used by litigants (or their attorneys) to bias experts in their favor).

[91] A similar way for a litigant to engage in a type of indirect data fishing while avoiding the negative repercussions of overt data fishing by an individual expert is to engage in a problematic practice called "expert mining," or "witness shopping." Gelbach, *supra* note 79, at 131 ("expert mining"); Posner, *supra* note 79, at 1541 ("witness shopping"). This practice involves "hiring multiple experts, asking each to provide an expert report on the same issue, and then put[ting] on the stand only the one who provides the most favorable report," or, in particular, "directing each to conduct a single test until one turns up a helpful result." Gelbach, *supra* note 79, at 131, 136; *see also* Posner, *supra* note 79, at 1541–42 ("Suppose the lawyer for the plaintiff hired the first economist, agronomist, physicist, physician, etc. whom he interviewed, and the lawyer for the defendant hired the twentieth one whom *he* interviewed. A reasonable inference is that the defendant's case is weaker than the plaintiff's. The parallel is to conducting twenty statistical tests of a hypothesis and reporting (as significant at the five percent level) the only one that supported the hypothesis being tested.").

[92] *See supra* note 86 and accompanying text. *Compare In re* Roundup Prods. Liab. Litig., 390 F. Supp. 3d 1102, 1137 (N.D. Cal. 2018), *and* Karlo v. Pittsburgh Glass Works, LLC, 849 F.3d 61, 82–84 (3d Cir. 2017), *with* Ohio Pub. Emps. Ret. Sys. v. Fed. Home Loan Mortg. Co., No. 4:08-cv-0160, 2018 WL 3861840, at *7 (N.D. Ohio Aug. 14, 2018), *and* Bell v. Ascendant Sols., Inc., No. Civ.A. 301-cv-0166N, 2004 WL 1490009, at *3 (N.D. Tex. July 1, 2004) (excluding study that identified "information days" that "appear[ed] to be consciously chosen in order artificially to support [the expert's] hypothesis").

although a jury may consider an expert's engagement in data fishing in evaluating the expert's testimony, an opposing party may encounter various difficulties in using data fishing to discredit the evidence. These include proving that an opposing party actually engaged in data fishing, explaining to the jury why data fishing is problematic, and, relatedly, avoiding giving the jury the impression that the litigant opposing the evidence is focused on data fishing only because the litigant has a weak substantive argument.

To be sure, data fishing in litigation is somewhat distinct from data fishing in legal scholarship. Litigants generally do not purport to assert neutral research claims. Instead, they advocate for a legal position and offer support for it. Litigants are overt in offering evidence that supports their positions, and the exploration of data by experts (as well as the prior vetting of experts) is often expected. On the other hand, expert testimony is expected to reflect the witness's expert opinion. And, in any case, data fishing in litigation, like data fishing in scholarship, is very damaging. Courtroom battles between experts frequently boil down to the hired-gun problem: each expert offers testimony to support his sponsor's position. The experts search for and selectively present methodologies that lead to results that are favorable to their respective sponsors. This disingenuous battle causes confusion and distrust among jurors, and it leads to a wide range of harms, including inaccuracy, unpredictability, and loss of faith in the courts.

Finally, as in scholarship, regardless of the precise prevalence of data fishing in the courtroom, the *opportunity* for experts to engage in this practice is sufficient to cause substantial harm. The jury is aware of the expert's bias in favor of his litigant sponsor, as well as his opportunity to search for and selectively report favorable results. The jury, therefore, has substantial reason to distrust the expert's claims even if the expert's methodology seems sound.[93] Furthermore, in this environment, experts are incentivized to engage in data fishing in order to maintain a substantive advantage over an opposing expert, or at least to prevent an opposing expert from herself gaining an advantage. The litigation may consequently devolve into one in which experts engage in data fishing and juries justifiably distrust expert testimony and choose winners based on inappropriate criteria. This devolution applies equally in cases in which a litigant hires a nontestifying expert to perform exploratory analysis prior to hiring a testifying expert.

---

[93] For obvious reasons, this analysis does not apply equally to court-appointed experts. *See* FED. R. EVID. 706.

Again, this model may be overly simplistic in certain respects, but it illustrates the point that the opportunity to engage in data fishing is itself sufficient to cause inaccuracy, unpredictability, and loss of faith in the courts.

II. ELIMINATING DATA FISHING WITH DASS

One of the most remarkable features of data fishing is how comprehensively it can be resolved, relative to the harm that it causes, with attentiveness to the problem by researchers and readers. In this Part, I draw on methods in statistics and other fields to propose a concrete framework for eliminating data fishing in law. Its steps can be summarized using the acronym DASS, for Design, Analyze, Scrutinize, and Substantiate. The first three steps (Design, Analyze, and Scrutinize) directly safeguard against data fishing. The fourth step (Substantiate) indirectly safeguards against data fishing by requiring evidence that the researcher has adhered to this framework. In this Part, I explain these steps and discuss their implementation. Then, in the following Part, I discuss a method that would allow a researcher to perform useful exploratory analysis while adhering to DASS.

A. Design Before Analysis

The first principle of DASS is based on the idea of methodological prespecification: the researcher should design the study and record it in a protocol prior to analyzing the data.[94] This means that, prior to beginning analysis—and, ideally, prior to having any access to outcome data—the researcher must carefully determine the design of the study, where "design" means "all contemplating, collecting, organizing, and analyzing of data that takes place prior to seeing any outcome data."[95]

As part of the design, the researcher should define all estimands and estimators—the objects that the researcher intends to estimate and what the researcher intends to use to

---

[94] *See* Donald B. Rubin, *For Objective Causal Inference Design Trumps Analysis*, 2 ANNALS OF APPLIED STAT. 808, 810, 816–17 (2008) ("[O]utcome-free design is absolutely critical for objectivity."); D. James Greiner, *What Do Statisticians Really Need to Know, and When Do They Need to Know It?*, *in* BLINDING AS A SOLUTION TO BIAS: STRENGTHENING BIOMEDICAL SCIENCE, FORENSIC SCIENCE, AND LAW 167, 175–78 (Christopher T. Robertson & Aaron S. Kesselheim eds., 2016).

[95] Rubin, *supra* note 94, at 810, 812; Donald B. Rubin, *The Design* versus *the Analysis of Observational Studies for Causal Effects: Parallels with the Design of Randomized Trials*, 26 STAT. MED. 20, 21 (2007); *see also* Ho & Rubin, *supra* note 17, at 27–28.

estimate them.[96] The design also specifies all other fundamental components of a study—for example, the research units, treatment levels, covariates, outcome variables, sampling methodology, etc.[97] It identifies how the researcher intends to analyze the outcome data.[98] This includes details regarding the researcher's intended hypothesis testing, levels of significance, data transformations, multiple-comparisons methods, and other components of the researcher's intended analysis.

In short, the researcher should specify the study's methodology[99] and should record its details in a research protocol. To the extent that it is not possible or practicable to record all important methodological details in the study's protocol, the researcher should at least specify the primary details of the study's methodology.[100] The more detail the

---

[96] For example, in a study to determine the average age of personnel in a certain branch of the military, the researcher may define her primary estimand as $\mu = \frac{x_1+x_2+\ldots+x_N}{N} = \frac{\sum_i x_i}{N}$, the mean age of every person in that branch of the military, where x_i is the age of person i; and she may define her primary estimator as $\bar{x} = \frac{x_1+x_2+\ldots+x_n}{n} = \frac{\sum_i x_i}{n}$, the mean age of a random sample of n people in that branch of the military. Defining estimands and estimators is frequently not simple, let alone obvious. For example, a researcher studying whether a drug has a causal effect on pain must precisely define a causal effect, as well as how such an effect will be estimated. Defining an estimand may involve the researcher asking herself, for example, "if I had all of the data I could ever need or want—even data that, in the real world, would be impossible to obtain—how would I compute the causal effect?" Defining an estimator, then, may involve the researcher asking herself, "given that I cannot obtain all of the data that I would need to compute the estimand, what computation can I perform to best estimate it using the data that I will have?" For a discussion surrounding the definition of estimands and estimators in a recent experimental study, see *The Effects of CCG*, *supra* note 37, at 429.

[97] *See* Rubin, *supra* note 95, at 21, 33.

[98] *See* Rubin, *supra* note 94, at 811–12 ("[F]or example, design includes conceptualization of the study and analyses of covariate data used to create matched treated-control samples or to create subclasses, each with similar covariate distributions for the treated and control subsamples, as well as the specification of the primary analysis plan for the outcome data.").

[99] I use the term *methodology* to include all facets of the study's design and analysis (or intended analysis, depending on the context).

[100] Additionally, in certain circumstances, it may be appropriate to have multiple design phases. For example, in the experimental setting, a researcher will often be interested in testing whether the randomization of units to treatment levels led to covariate balance—that the background characteristics of units are balanced across treatment levels. In this case, prior to randomization, the researcher may complete an initial design that contemplates a number of analytical possibilities based on whether and to what extent the randomization achieved covariate balance. Then, after randomization and data collection, the researcher can analyze the covariate data without accessing the study's *outcome* data. Finally, once the covariate data has been analyzed but prior to accessing the outcome data, the researcher would complete the study's design based on the results of the researcher's analysis of the covariate data. *See, e.g.*, *The Effects of CCG*, *supra* note 37, at 426–28 (describing analysis of covariate data in "secondary design phase"). *See generally* Rubin, *supra* note 94 (emphasizing the importance of careful study design); Greiner, *supra* note 94 (discussing "blinding"—keeping certain information from a researcher—as a tool for improving statistical analysis and causal inference in particular).

researcher provides, the more credible she will be in asserting her statistical claims.[101]

The researcher should complete the design phase of the study prior to seeing the outcome data, and, ideally, prior even to having *access* to the outcome data. I discuss this component further in Section II.C.

Once the design is complete and the researcher has specified her methodology, the researcher may proceed to the analysis phase of the study. Ideally, this simply involves following the methodology established and recorded in the study's design phase. If deviations from this prespecified methodology are necessary, they should be declared and explained in the study's report.

B. Scrutinizing the Study's Methodology and Results

Once the researcher has completed the design and analysis phases of her study, she should turn her attention to scrutinizing her methodology and results. Of course, a good researcher carefully scrutinizes her methodology during the design phase of a study. At this later point of the study, however, the researcher has access to the outcome data and to the results of her study. The primary point of this element of DASS is for the researcher to use "sensitivity analysis" to examine the robustness of her results to different methodologies and to explain the reasoning for and the weaknesses associated with the study's methodology and results.[102]

Note that terms such as "validate" and "support" would be less suitable than "scrutinize" to describe this element of DASS. The key to fulfilling this element is for the researcher to perform an earnest examination of her methodology and results. The point is not for the researcher to *advocate for* or to *prove* the robustness of her results or strength of her methodology. Rather,

---

[101] *See generally* Mitchell, *supra* note 1, at 176, 186–87, 197–204 (discussing importance of transparency in empirical research and recommending adoption of "stringent disclosure requirements for reports of original empirical research [in law], including disclosure of detailed information about methodology, data analysis, and the availability of raw data for replication and review"). Mitchell emphasizes the importance of providing "the details necessary for others to simulate one's methods to check for similar results." *Id.* at 185. DASS is of course consistent with the overarching message of Mitchell's proposal—the need for transparency. Note, however, that DASS is premised on the notion that strict disclosure requirements (and, particularly, those imposed by journals) are neither sufficient nor necessary to safeguard against data fishing. Additionally, DASS is intended to serve as a simple and concrete standard, but one that is sufficiently flexible to be applied (by researchers and readers directly) to a wide range of research settings and conditions.

[102] The researcher can incorporate guidelines for such analysis in her design. Note that sensitivity analysis has numerous functions in statistics. *See, e.g.*, ANDREW GELMAN ET AL., BAYESIAN DATA ANALYSIS 141–62, 184–85, 435–36 (3d ed. 2014) (discussing sensitivity analysis in Bayesian statistics).

it is more in line with Richard Feynman's comments regarding "cargo cult science"—a term Feynman used to describe a sort of fake science—in his 1974 Caltech commencement address:

> [T]here is *one* feature I notice that is generally missing in cargo cult science. That is the idea that we all hope you have learned in studying science in school—we never explicitly say what this *is*, but just hope that you catch on by all the examples of scientific investigation. . . . It's a kind of scientific integrity, a principle of scientific thought that corresponds to a kind of utter honesty—a kind of leaning over backwards. For example, if you're doing an experiment, you should report everything that you think might make it invalid—not only what you think is right about it: other causes that could possibly explain your results; and things you thought of that you've eliminated by some other experiment, and how they worked—to make sure the other fellow can tell they have been eliminated.
>
> Details that could throw doubt on your interpretation must be given, if you know them. You must do the best you can—if you know anything at all wrong, or possibly wrong—to explain it. If you make a theory, for example, and advertise it, or put it out, then you must also put down all the facts that disagree with it, as well as those that agree with it. There is also a more subtle problem. When you have put a lot of ideas together to make an elaborate theory, you want to make sure, when explaining what it fits, that those things it fits are not just the things that gave you the idea for the theory; but that the finished theory makes something else come out right, in addition.
>
> In summary, the idea is to try to give *all* of the information to help others to judge the value of your contribution; not just the information that leads to judgment in one particular direction or another.[103]

I finally settled on the term "scrutinize" to express this element of DASS. It entails a sincere examination by the researcher of her results. As Feynman emphasized,

> The first principle is that you must not fool yourself—and you are the easiest person to fool. So you have to be very careful about that. After you've not fooled yourself, it's easy not to fool other scientists. You just have to be honest in a conventional way after that.[104]

This element is more of an art than a science. The goal is to examine the study's methodology and whether and how the study's results would differ under reasonable alternative methodologies. It involves identifying the study's primary methodological features and examining the sensitivity of the

---

[103] RICHARD P. FEYNMAN, "SURELY YOU'RE JOKING, MR. FEYNMAN!": ADVENTURES OF A CURIOUS CHARACTER 385–86 (W.W. Norton & Co. 2018) (1985).

[104] *Id.* at 387; *see also* Mitchell, *supra* note 1, at 176 (discussing importance of transparency in empirical research and recommending adoption of "stringent disclosure requirements for reports of original empirical research [in law], including disclosure of detailed information about methodology, data analysis, and the availability of raw data for replication and review").

study's results to changes in those features. It also entails explaining why the study's primary methodological features were chosen and examining how the results would differ using reasonable alternatives. For example, the researcher might test whether her results differ when she measures central tendency using the median rather than the mean, or whether her results differ when she addresses outlying data by winsorizing the data at the 99th percentile rather than the 95th percentile, or a combination of the two.

How extensively a researcher should examine alternative methodologies depends on a number of factors. First, how plausible are the alternatives? If the researcher's choice of a particular methodological feature is an obvious choice that stands above all other choices, there may be less of a need to examine alternatives for this feature. If, on the other hand, there are multiple options that stand on equal footing, then a thorough sensitivity analysis may be necessary. If there are many plausible alternatives, a researcher may attempt to identify the most important alternatives, test them, and use the results to determine whether further tests are necessary.

Second, how detailed is the research protocol; and did the researcher complete the study's design prior to accessing the outcome data, and can she prove it? If the researcher is unable to evidence that she completed the study's design prior to analyzing the outcome data, the reader should read the study skeptically. But the reader's skepticism may be overcome if the researcher can prove, by scrutinizing her results, that the study's results should be trusted. If, on the other hand, the researcher can show that she completed the study's design prior to beginning her analysis, the researcher's scrutinization, although still important to demonstrate the robustness of the study's results, may be less critical, since readers have more reason to trust the study based on the researcher's adherence to the design-before-analysis principle.

It is important to realize that, if the researcher tests alternative methodologies and obtains results that are inconsistent with those in her primary analysis, this does not necessarily invalidate her study's results. The researcher must consider why her results differ based on different acceptable methodologies and whether the results under her chosen methodology are meaningful notwithstanding their sensitivity to such changes. Generally, she should report the results of her sensitivity analysis and explain to the reader her theory regarding why the inconsistencies arise and whether they invalidate her study's results. Inconsistencies should be expected, and frequently there are good explanations for them

and for why a study's results are meaningful notwithstanding their sensitivity to certain changes in methodology. A cautious researcher may also incorporate in her study design a specific plan to account for the possibility of such inconsistencies.

In summary, the scrutinize element is about examining and being transparent regarding a study's methodology and the robustness of the researcher's results to changes in methodology. The spirit of this element of DASS is in stark contrast to that of data fishing, which involves searching for methodologies that produce favorable results and effectively hiding (or at least lacking transparency regarding) alternative methodologies that produce unfavorable results. The process of scrutinizing one's own results with sensitivity analysis is a key component to safeguarding against data fishing and to gaining a reader's trust generally. At the same time, the reader should expect this in a study and should be cautious when an empirical study does not include such self-scrutinization.

C. Substantiating Adherence to DASS

The aim of this article is to facilitate the elimination of data fishing and the restoration of trust in empirical legal research. This goal requires not only principles for safeguarding against data fishing, but also a process that allows readers to know whether and to what extent a researcher has applied these principles. In turn, the researcher's awareness that a reader will know the extent to which the researcher has followed DASS and will credit or discredit the study's results accordingly will incentivize careful adherence to its elements. Thus, providing readers with evidence of the researcher's adherence to DASS is fundamental for achieving both the aims of the researcher, who is interested in promoting her research, and the aims of readers, who are interested in consuming *credible* research. This element of DASS is central to its effectiveness and represents an important advancement over current practices in the natural and social sciences.[105] Moreover, as mentioned in the

[105] Although the focus of DASS's substantiate element is the provision of evidence to readers directly rather than the fulfillment of requirements enforced by journals, law journals have a strong incentive to consider evidence of an author's adherence to DASS's principles in deciding whether to extend an offer of publication. Journals are incentivized to publish credible research, and an author's adherence to DASS's principles is significant in this respect. In this sense, the substantiate element allows researchers to signal credibility to journals as well as to readers. This effect may help to counterbalance a researcher's concern regarding her adherence to DASS in light of the well-evidenced phenomenon that journals have a tendency to publish results that are statistically significant over those that are not. *See* Zeiler, *supra* note 1, at 82; Ioannidis et al., *supra* note 1, at F241; NASEM REPORT, *supra* note 13, at 91–103.

Introduction, it is arguably of particular importance in legal scholarship, which frequently does not involve a peer-review selection process for articles.[106]

The first step is for the researcher to establish evidence that she completed the study's design prior to analyzing the study's outcome data and that her analysis follows her prespecified methodology. There are various methods for accomplishing this. First, the researcher should specify her design in writing and record it, ideally, by publishing it, "preregistering" it, or at least privately uploading it electronically with a timestamp. Electronic forums have been developed specifically to facilitate protocol preregistration and similar practices.[107] Alternatively, SSRN (Social Science Research Network), a popular forum for posting working papers in law and other fields, can be used to post research protocols privately or publicly.[108] SSRN is well suited for this practice and would likely be willing to incorporate additional features to further accommodate it.

Alternatively, a researcher can establish a small committee of colleagues and use this committee to facilitate evidencing her adherence to the design-before-analysis principle by submitting a timestamped protocol to the committee prior to beginning data collection or analysis. In some cases, this may be as simple as emailing the committee a final version of the study's protocol. The researcher can also use the committee to facilitate restrictions on the researcher's access to data and other safeguards. Like preregistration or uploading a research design to SSRN, this practice is simple, and it does not place any substantial burden on the researcher or on the members of the committee.

Whichever practice the researcher chooses to apply, the researcher should substantiate her adherence to DASS by specifying in her report which of these safeguards she used (e.g.,

---

[106] In legal scholarship, the substantiate element may therefore be of particular importance to both readers and journals.

[107] *See, e.g.*, *Preregistration*, CTR. FOR OPEN SCI., https://www.cos.io/initiatives/prereg [https://perma.cc/NEZ7-3G8J].

[108] Some researchers may have concerns regarding the idea of publicizing a research protocol, especially at an early stage of a study. Researchers, however, have numerous options for recording the protocol privately. Although sometimes ideal, it is frequently not necessary to publicize a study's design prior to publishing the study's report. Furthermore, in some circumstances, it may be necessary or desirable to avoid sharing a record of a study's protocol with readers, even after the study's report has been published. In these circumstances, a researcher would nevertheless have options for evidencing that she completed the study's design prior to analyzing the outcome data and that her analysis complied with the study's prespecified methodology. For example, the researcher could utilize a committee of colleagues (*see* text accompanying notes 108–109), a neutral third party, or a confidentiality agreement to facilitate such proof.

preregistration or establishing a committee) and by signaling to the reader the researcher's ability to provide further evidence of such safeguards.[109]

Experimental studies are ideal for exercising and evidencing design-before-analysis practice. This is because the researcher generally has control over the timing of her access to the study's outcome data. In this context, the researcher can complete and record her protocol prior to collecting data and therefore can easily prove to readers that she has prespecified her methodology. If the researcher wishes to begin data collection while completing certain aspects of the protocol, the researcher can still evidence her strict adherence to design-before-analysis practice by restricting (verifiably) her access to the data (e.g., with the help of a committee, IT department, or otherwise) until the protocol is recorded as complete.

Although experimental studies are ideal for exercising and proving design-before-analysis practice, most empirical studies in law are observational rather than experimental. Generally, contrary to experimental studies, the data in an observational study already exists, and the researcher often lacks control over the timing of her access to the data. Consequently, it can be more difficult for a researcher to prove that she prespecified her methodology. Indeed, the researcher will frequently have full access to the data—for example, public data available on the internet—prior to beginning the study design or even conceiving the idea for the study in the first instance. These circumstances call for special care: while proving adherence to DASS can be more difficult for observational studies, adherence to DASS is arguably most important for these studies in particular.[110]

For some observational studies, the researcher will have control over when she gains access to the outcome data. For example, if a district attorney promises to provide certain prosecution data to a researcher, the researcher can instruct the

---

[109] The researcher should be able and willing to provide certain relevant records—e.g., a timeline (*see infra* notes 110–112 and accompanying text), a record of when the researcher gained access to the study's outcome data, preregistration records or a letter from an established committee attesting to the researcher's completion of the study's design prior to accessing any outcome data, a record of the study's protocol, or other records, depending on the circumstances. *See also supra* note 108. Note that I do not mean to suggest that DASS, and the substantiation element in particular, should *replace* safeguards instituted by journals. To the contrary, journal requirements aimed at safeguarding against data fishing only complement and further DASS's principles.

[110] Experimental studies are frequently considered the "gold standard" for empirical research. Rubin, *supra* note 94, at 808 ("For obtaining causal inferences that are objective, and therefore have the best chance of revealing scientific truths, carefully designed and executed randomized experiments are generally considered to be the gold standard.").

district attorney to refrain from giving her access to the data until she informs him otherwise. Alternatively, the researcher can create a committee or other "neutral" entity that would receive the data and withhold access to it until instructed otherwise by the researcher. In either case, the researcher can then complete and record her study design, and establish evidence of her completion time, prior to gaining access to the outcome data. She would maintain her research protocol and a verifiable record of when she completed (and recorded) her design, when she requested the data, when she obtained the data, and when she began her analysis. Most importantly, the researcher should establish and maintain evidence of (1) when she completed her study design, and (2) when she gained access to and when she actually accessed the outcome data (as applicable).[111]

For observational studies in which the researcher has little or no control over the timing of her access to the outcome data, the researcher must turn to weaker modes of proof and therefore must rely more heavily on DASS's scrutinization component. First, regardless of whether she can substantiate her timing, she should maintain a record of when she completed her design, when she accessed the outcome data (and, if applicable, when she gained access to the outcome data), and any other timepoints and occurrences that may be important for establishing that she completed her design prior to analyzing the data. For example, if a researcher has had access to a public dataset on the internet, she may maintain a record of when she completed her design, when she discovered the dataset, occasions on which she accessed the data prior to completing her design and what she did with the data on those occasions, and when she began her analysis. Second, in reporting the study's results, the researcher should be transparent regarding her access to and use of the data prior to completing the study's design. She may also include relevant timepoints.

In addition to evidencing the researcher's completion of the design prior to her analysis of the outcome data (and reporting any access to and use of the data prior to completing the study's design), the researcher should substantiate her adherence to the scrutinize element simply by reporting the

---

[111] If, for example, the researcher had no option other than to gain access to the data from the prosecutor prior to completing the study's design, but assuming that she could nevertheless establish and maintain evidence that she refrained from actually accessing the data prior to completing her design—e.g., evidence from an access record maintained by the technology hosting the data—she should record and maintain this evidence. This evidence is somewhat weaker than evidence that the prosecutor had not yet granted her access to the data, but it is likely convincing nevertheless.

primary results of her sensitivity analysis and providing a summary of her analysis and any other relevant information (such as weaknesses in the study's methodology) that would help the reader to assess the study's results.

Regardless of whether and to what degree the researcher has control over the timing of her access to the study's outcome data, and regardless of the strength of the *evidence* of her adherence to DASS, a key component of DASS's substantiate element is this: the researcher should include in the study's report a statement attesting to whether and to what degree she adhered to DASS's principles. In essence, adhering to DASS means taking deliberate steps to avoid data fishing and to present results transparently and honestly—and, in particular, following its design-before-analysis, scrutinization, and substantiation principles.[112] The first step in substantiating such adherence is simply stating (in the text or a footnote, perhaps at the start of the report's methodology section) whether the researcher has in fact adhered to DASS by following these principles.[113] Beyond this statement, it is the researcher's responsibility to convince the reader of such adherence with additional details.

In particular, once the researcher has indicated her adherence to DASS's principles, she should provide details that would strengthen or weaken this claim. These may include any steps that the researcher has taken to fulfill the design-before-analysis, scrutinization, and substantiation elements of DASS and a reference to the reported results of the researcher's sensitivity analysis. They should also include any weaknesses in the researcher's adherence to DASS, such as instances in which the researcher accessed the data prior to completing the study's design, as well as any qualifications or steps that the researcher took to rectify these weaknesses.

Note that a reader's expectations may sensibly depend on the circumstances of the study. For example, a reader may be more forgiving of a researcher's access to data prior to her completion of the study design if preventing such access would have been impossible or impractical. On the other hand, the reader may be less forgiving if the researcher did not take proper precautions to ensure completion of her design prior to obtaining access to outcome data in an experimental setting. Of course, this is only part of the story: some fixed level of substantiation may be expected

---

[112] See *infra* Conclusion for a detailed discussion of what it means to adhere to DASS.

[113] *See generally* Joe Simmons et al., *A 21 Word Solution*, DIALOGUE (Soc'y for Personality & Soc. Psych.), Fall 2012, at 4–7, https://spsp.org/sites/default/files/dialogue_26%282%29.pdf [https://perma.cc/G2PN-N767] ("If *you* did not *p*-hack a finding, *say it*, and your results will be evaluated with the greater confidence they deserve.").

and required to afford the study credibility, regardless of the circumstances. But some dependence on circumstance is reasonable and useful. It is reasonable, because a researcher's failure to fulfill the substantiate element when substantiation is feasible may signal that the researcher has not adhered to DASS's principles. And it is useful, because many important studies may require investigation under subideal circumstances in which a high degree of substantiation would be impossible; and an uncompromising refusal to afford such studies credit would disincentivize potentially important research.

In any event, the reader should consider whether the researcher has at least *attested* to her adherence to DASS's principles. Based on the circumstances, this, in combination with the researcher's scrutinization of her results, may satisfy the reader that the researcher's results are credible.

The substantiate element of DASS is critical. Ideally, a researcher will strive to adhere to the design-before-analysis and scrutinize standards in a way that could be described, in Feynman's words, as "a kind of utter honesty—a kind of leaning over backwards."[114] Ultimately, however, it is up to the researcher to substantiate this adherence, and for the reader to judge it based on the evidence—and to credit the researcher's results accordingly.

III. THE NEED FOR DATA EXPLORATION

A researcher should complete her study design prior to analyzing the outcome data; however, good study design often requires data exploration. Pilot studies and training datasets satisfy this need without violating the principles of DASS. In particular, the design-before-analysis principle requires that a researcher complete her study design prior to analyzing the outcome data *to which she intends to apply her study design*. Engaging in exploratory analysis and using the results of that analysis to inform the researcher's study design is not violative of DASS and does not constitute data fishing so long as the researcher is transparent regarding her exploration and uses a dataset for her exploratory analysis that is *different* from the one to which she intends to apply her study design.

In the experimental context, "pilot studies" are commonly used to perform exploratory analysis and hone study design.[115]

---

[114] FEYNMAN, *supra* note 103, at 385.

[115] *See generally Pilot Studies: Common Uses and Misuses*, NAT'L CTR. FOR COMPLEMENTARY & INTEGRATIVE HEALTH, https://www.nccih.nih.gov/grants/pilot-studies-common-uses-and-misuses [https://perma.cc/32L5-9JW6] (discussing uses and misuses of pilot studies).

Pilot studies provide an opportunity for a researcher to perform exploratory research and to learn from such exploration prior to beginning the "main study." In the observational context, "training datasets" are used to perform exploratory analysis, while the researcher's study design is ultimately applied to the study's "testing dataset." I begin this Part by explaining how a researcher can engage in exploratory analysis using pilot studies without violating the principles of DASS. I then explain why the same reasoning applies to exploration using training datasets in the observational context and discuss methods for partitioning a dataset into training data and testing data for exploration and testing, respectively.

A. Pilot Studies

Pilot studies serve an important purpose: they are used to detect design flaws and gain other useful information for a study's design prior to undertaking a costly full-blown experiment. Pilot studies involve a researcher's analysis of the pilot study's outcome data for the purpose of orienting the design of the main study.[116] We must therefore ask whether this practice causes the same problems that arise from data fishing.

A key distinction between exploration using pilot studies and data fishing is that, in the former, the researcher analyzes outcome data from the pilot study rather than the main study—that is, from a dataset that is distinct from the data that the researcher ultimately uses to test her hypotheses. Consequently, contrary to data fishing, exploration in a pilot study does not cause false positives. Furthermore, although pilot studies can be used to give false impressions, this risk can be minimized through attentiveness to the scrutinize element of DASS.

1. False Positives

Data fishing causes false positives.[117] However, analyzing data in a pilot study and using the pilot study results to decide on methodological factors in the main study does not cause false positives in the main study. This is because the main study involves a new set of data.

A 0.05 level of significance reflects a certain tolerance for false positives (a rate of 5%). False positives occur because there is randomness in a sample of data, and 5% of the time, the

---

[116] *See id.*

[117] *See supra* Section I.C.1.

observed test statistic (e.g., difference in means) will be sufficiently extreme to be "significant" as a result of the randomness from sample to sample rather than a true effect. If, for example, a researcher engages in data fishing by performing one hundred hypothesis tests in search of significant results, and, in each, the null hypothesis is in fact true (meriting acceptance rather than rejection via a finding of significance), then we would nevertheless expect five significant results (5% = 5/100) due to randomness alone—that is, five false positives. Therefore, if a researcher only reports significant results without informing the reader of her exploratory analysis or otherwise accounting for expected false positives, then she substantially understates the rate of false positives, and her findings of significance may well be spurious. On the other hand, if the researcher uses a pilot study to identify methodological combinations that yield significant results and then applies those particular combinations to the data in the main study—a new sample entirely—then, if the results are indeed spurious, these methodological combinations are highly unlikely to produce significant results in the new sample. After all, if a finding of significance in the pilot sample is spurious, it occurred by randomness alone; at a 0.05 level of significance, there is only a 5% chance of recurrence in the main study.

Similar reasoning applies to the related problem of overfitting. If, in the pilot study, a methodology is selected to fit the pilot data too tightly, and thereby to obtain a significant result in the pilot study, then the researcher is, in a sense, misinterpreting sampling variation as a true characteristic of the population; therefore, as in the example above, applying this methodological combination to a new set of data in the main study is likely to yield a nonsignificant result. Contrary to data fishing, using a pilot study for exploration facilitates a balance between fitting a model, or methodology, to the pilot data on the one hand and ensuring that it is sufficiently general to apply to the main study's data on the other hand. In other words, exploration of data in a pilot study discourages overfitting because an overfitted model will generate poor results in the main study.

2. False Impressions

Even putting aside false positives, data fishing misleads the reader.[118] By searching for and selectively reporting

[118] *See supra* Section I.C.2.

significant results, the researcher causes readers to believe that the study's results are more robust and replicable than they actually are. Readers are generally interested in experimental results only insofar as the results tell us about the real world—that is, only insofar as the results apply with some robustness rather than apply only under the precise set of conditions in the experiment, even if the results could be replicated using an identical methodological combination in a new sample. In other words, readers are interested in results that are method replicable as well as sample replicable.[119] Data fishing, however, often involves reporting results in a way that gives the reader the impression that the results are generalizable to a robust set of real-world conditions when in fact they are not.

Exploratory analysis in a pilot study can also be used to create false impressions in the main study if the researcher cherry-picks a methodology in the pilot study and then applies it in the main study. The risk of overfitting prevents this practice to some extent, since overfitting will result in false positives (in the pilot study) and therefore nonsignificance in the main study. Sometimes, however, the researcher, through her exploration of different methodologies, will detect a true characteristic of the population; but her selective reporting of the result will cause the reader to believe that it is more robust and replicable than it is.

Consider again the example above from *The Effects of CCG*.[120] As ***Table 2*** in Section I.B shows, only one methodological combination out of eight yields a negative causal effect of the intervention on the outcome variable. It is possible that this result could be the product of overfitting and therefore not replicable in a new sample. However, it is also possible that this negative effect reflects a true signal—for example, that defining the outcome variable in terms of a mean rather than median damages award (in the punitive damages setting) yields a negative effect. This effect may well be sample replicable. But, even assuming that it is, the result may be misleading to readers if a researcher discovers this negative effect in a pilot study, cherry-picks the one methodological combination that yields this effect for application in the main study, and then, after obtaining a negative effect in the main study, reports this result in isolation of the other seven possibilities, none of which yields a negative effect.

Here, the result reported may be perfectly correct—that is, it may be correct to conclude that the intervention causes a

---

[119] *See supra* notes 55–57 and accompanying text.

[120] *See supra* notes 37–40 and accompanying text.

negative effect on the mean of punitive damages awards when winsorizing at the 95th percentile. This result is not a false positive. The problem is that repeating this methodology in the main study and reporting the result in isolation of its context (i.e., without explaining why the researcher chose the particular methodology that she chose) may give the false impression that the result is more robust than it really is—that it is method replicable, and that, for example, the observed negative effect would occur if the median or log mean were used to define the outcome variable. As discussed above, this false impression may arise due to the reader's assumption that the researcher has designed the study based on neutral criteria rather than on what methodology would produce a favorable result. If the researcher has cherry-picked a favorable result, it is far less likely to be robust to modest changes in methodology; and, if the researcher reports the result without accurately explaining how she arrived at her methodology, she will likely give the reader a false impression regarding the robustness of the study's results.

This problem can, however, be addressed with attentiveness to the scrutinize element of DASS. It is the researcher's responsibility to explain her methodology and to perform and report basic robustness checks. She should also be transparent regarding her use of pilot studies. The reader should expect this. She should note whether the researcher has attested to following DASS's principles, and she should be diligent in considering how the researcher has used pilot studies and why the researcher has chosen a particular methodology. She should also examine the study's sensitivity analysis with respect to the study's major methodological features. For example, in the foregoing illustration, a reader should expect the researcher to provide sensitivity analysis displaying some of the results shown in ***Table 2***. The reader's attentiveness to the researcher's attestation to her adherence to DASS, use of pilot studies, and sensitivity analysis, combined with the risk of overfitting, significantly limits the risk of a researcher using a pilot study to mislead readers with respect to the robustness of a study's results.[121]

---

[121] Importantly, in addition to the elements discussed, peer review and "workshopping" can serve as fundamental safeguards against false impressions and data fishing generally. When a researcher presents a finding to a particular research community, members of that community—whether workshop participants or journal referees—may insist on methodological specifications that differ from those employed by the researcher, thereby implicitly or explicitly testing the robustness of the researcher's results. These community members, in general, do not have access to the data and have neutral perspectives relative to that of the researcher, who may be subject to certain conscious or subconscious motivational biases.

B. Training Data

In the experimental setting, pilot studies allow the researcher to benefit from exploration without compromising the principles of DASS. In the observational setting, this function is fulfilled by a process of partitioning a dataset into training data and testing data, where the training data is used for exploratory analysis and the testing data is used for confirmatory analysis.[122] Once the data is validly partitioned, the discussion above regarding data exploration using pilot studies generally applies to data exploration using training data.[123]

A researcher can use training data to engage in exploratory analysis legitimately, *without data fishing*, as follows: First, without analyzing the outcome data (or, ideally, without looking at it or even accessing it), the researcher should obtain a small random sample from the dataset. Assume that a dataset contains five-thousand data points. The researcher can take a random sample of, e.g., 5% or 10% of the data (250 or 500 data points, respectively) and set it aside as a training dataset while preserving the remaining data for the main study.[124] The researcher can then use the training data to perform exploratory analysis. This data is analogous to the data in a pilot study. The researcher can view the data, explore it, and test it for patterns. As with a pilot study, the researcher can use her results from the training data to inform her design for the main study—for example, to inform the researcher's modeling decisions, choice of estimands and estimators, subgrouping decisions, and handling of outliers.

This useful method, sometimes referred to as "train-test splitting," allows the researcher to conduct exploration on a sample of data from the population of interest in the main study while preserving the bulk of the data for testing in the main study.[125] Further, in certain circumstances, more complex "cross validation" methods may allow a researcher to perform exploratory analysis on a sample of the data without having to then discard the training data during the testing phase in the main study.[126] These methods may be particularly useful when

---

[122] The terms "training data" and "testing data" (or "test data") are borrowed from the machine-learning context, where, for example, a model is trained to recognize patterns in data using training data before it is evaluated using test data. *See* Yarkoni & Westfall, *supra* note 43, at 1102, 1111.

[123] *See supra* Section III.A.

[124] *See* Yarkoni & Westfall, *supra* note 43, at 1110–11.

[125] *See id.*

[126] *See id.*

the entire dataset involves a relatively small sample and the researcher cannot afford to discard any data.[127]

The researcher should be cautious to obtain training data in a way that allows her to evidence her adherence to DASS. In the experimental context, this is simple because the researcher generally has control over when the data is generated and thus can perform pilot studies and complete her design of the main study prior to accessing any main-study data. Observational studies can present additional challenges in this regard. However, establishing good evidence of a researcher's adherence to DASS is frequently possible, even when the researcher begins her research by performing exploratory analysis on training data. For example, a researcher obtaining data from a third party may be able to request that the data be transferred to the researcher in two phases—first, a small random sample of the data would be transferred for use as training data, followed by the remaining data to be transferred at a later date for use as testing data.

In some circumstances, it can be very difficult to obtain relevant data, and convincing the data provider to coordinate a phased transfer of the data may be out of the question. In these situations, however, the researcher can consider using a committee to facilitate her use of training data for exploration. In particular, the committee would receive the data and transfer training data to the researcher while restricting access to the remaining data until a later date. The committee would maintain a record and timeline of all data released to the researcher.[128]

---

[127] As Yarkoni & Westfall explain, in the context of selecting and testing the performance of a model,

> instead of assigning each observation exclusively to either the training or the test datasets, one can do both, by repeating the cross-validation twice. In one "fold" of the analysis, one half of the data is used for training and the other half for testing; in a second fold, the datasets are reversed, and the training set and test sets exchange roles. The overall model performance is then computed by averaging the test performance scores of the two folds, resulting in a single estimate that uses all of the data for both training and testing yet never uses any single data point for both. More generally, this approached is termed *K-fold cross-validation*, where K, the number of "folds," can be any number between 2 and the number of observations in the full dataset (but is most commonly set to a value in the range of 3 to 10).

Id. at 1111.

[128] In certain circumstances, a data provider may be reluctant to transfer potentially sensitive data to a committee rather than to the researcher directly. In these circumstances, the researcher might consider assigning a member or members of her *research team* to act as a third-party entity to restrict the primary researcher's access to the data and to transfer the data to the researcher in phases. This is of course less ideal than having a neutral third-party entity to facilitate data transfer, but with proper precautions (e.g., creating an informational screen between individuals who have access to the test data and individuals who will perform exploratory analysis in the training phase, carefully maintaining records, etc.), this procedure may allow a researcher to

In still other circumstances, the researcher will not be able to provide strong evidence that her exploration has not violated DASS's principles. These situations generally correspond to those in which evidencing adherence to DASS is difficult in the first instance, except through the researcher's attestation and scrutinization. In these situations, the researcher should follow the same procedures recommended in Section II.C, such as careful record keeping, attesting to the researcher's adherence to DASS in the study's report, and detailing the results of the researcher's sensitivity analysis. Now, however, the researcher should also account for the exploration phase in her record keeping and attestation. The researcher should be transparent. For example, she should report her steps for partitioning the data into training data and testing data, her timeline for accessing each component of the data, and the steps she took to prevent data fishing, as well as associated evidence and records.

The observational context frequently requires additional care to ensure that the researcher can obtain training data without compromising the dataset generally and that she can substantiate her adherence to DASS, notwithstanding the study's exploration phase. However, the discussion regarding exploration using pilot samples generally applies equally to exploration using training data in the observational context. If the dataset is validly partitioned, exploration in the training data will not cause false positives in the main study using the testing data. Similarly, false impressions can be managed, as in the experimental context, using reasonable diligence and sensitivity analysis.

In summary, although the observational context frequently requires additional care to ensure that the researcher will obtain training data without compromising DASS's principles, the discussion above regarding exploration using pilot samples generally applies equally to exploration using training data in the observational context. DASS does not require sacrificing the benefits of exploratory analysis, even in the observational context. It only requires thoughtful planning and principled use of the data. In a sense, DASS facilitates a balance between the benefits of exploration using pilot studies and training data on the one hand and the need to safeguard against data fishing on the other.

---

establish convincing evidence of her adherence to DASS notwithstanding certain challenges associated with completing a phased data transfer of sensitive data.

IV. ADDRESSING DATA FISHING IN THE COURTROOM

As discussed in Section I.E, data fishing is common and more overt in litigation, notwithstanding protections against unreliable evidence. In cases involving empirical evidence, data fishing causes inaccuracy and facilitates disingenuous battles over methodology. More generally, data fishing plays a substantial role in causing the hired-gun problem and issues associated with battling experts. At the very least, experts have a substantial *opportunity* to engage in data fishing, and this alone can have far-reaching consequences.[129] Ultimately, data fishing causes unpredictability, inaccuracy, and loss of faith in experts and the judicial system, among other serious problems.[130] Eliminating the opportunity for experts to engage in data fishing, while not a silver bullet, would go a long way toward increasing the reliability of expert evidence.

An in-depth discussion of eliminating data fishing in litigation or of solving the hired-gun problem is beyond the scope of this article. Below, however, I briefly consider two applications of the discussion above for addressing the problem of data fishing in the courtroom.

First, trial courts should consider at least the basic elements of DASS in determining the admissibility of statistical evidence. Empirical results that are unreliable should be excluded from evidence under *Daubert* and Rule 702. Statistical results that are produced by data fishing are unreliable. They are misleading and suffer from unreasonably high error rates—error rates that are far higher than those asserted by the expert. As such, they should be excluded as unreliable evidence.[131]

My intention is not to argue that courts should necessarily exclude all analyses for which the researcher is unable to establish strict adherence to the principles of DASS. Rather, courts should view data fishing as a substantial threat to reliability and should consider these principles when applying

---

[129] *See supra* Section I.E.

[130] One consequence of the hired-gun and battle-of-the-experts problems is that juries can obtain a grossly lopsided image of the relevant scientific community's position on an issue. For example, even if 95% of experts in a field would agree with the defendant and only 5% of experts would agree with the plaintiff, the jury may well not become aware of this disparity from the testimony at trial. The jury would hear testimony from the plaintiff's and the defendant's experts, and frequently would not have any information as to the consensus in the field. *See infra* notes 136–139 and accompanying text. Exacerbating this problem, the jury often lacks a clear understanding of the experts' testimony and may decide which expert to believe based on criteria other than the substance of their testimony—for example, an expert's ability to convey information with clarity or confidence.

[131] *See supra* notes 79–86 and accompanying text.

standards of admissibility. Courts should at minimum consider whether the expert has refrained from data fishing and made a good faith effort not to mislead the jury. That is, courts should confirm that the expert contemplated the study's design prior to beginning analysis and that the expert performed basic sensitivity analysis. Ideally, experts will have established evidence to substantiate their adherence to DASS, but, at minimum, courts should consider requiring experts to attest to following DASS's principles at a basic level.

Additionally, lawmakers should consider revising disclosure protections for nontestifying expert analysis to make it more difficult for a litigant to introduce evidence that is based on undisclosed exploratory analysis. Technically, exploration on which a testifying expert's analysis is based *is* discoverable. But a litigant can bypass such discovery requirements, for example, by employing a nontestifying expert to perform exploratory analysis prior to hiring a testifying expert to perform confirmatory analysis.[132] Lawmakers should consider the reliability implications of data fishing, as well as the elements of DASS, in evaluating disclosure requirements and protections for both testifying and nontestifying experts.[133]

Second, litigants and experts should consider adhering to DASS's principles in order to improve the credibility of an expert's testimony. In particular, a litigant whose expert adheres to DASS's

---

[132] *See supra* notes 89–91 and accompanying text.

[133] *See generally* Gelbach, *supra* note 79, at 134–35 (discussing disclosure requirements for testifying and nontestifying experts); *id.* at 135–37 (discussing how "the Federal Rules of Civil Procedure and Evidence work together to enable a well-prepared party to punish an adversary for its expert's use of data mining" but how a litigant can instead achieve the same effect by "[h]iring [multiple] experts and directing each to conduct a single test until one turns up a helpful result"); *id.* at 144–46 (suggesting that a "possible reform [to prevent expert mining] would require disclosure not just of the number of experts hired, but also of the contents of reports provided by a party's nontestifying experts (including the contents of any oral report)"); Posner, *supra* note 79, at 1541 (proposing that "lawyers who call an expert witness could be required to disclose the name of all the experts whom they approached as possible witnesses before settling on the one testifying" to "alert the jury to the problem of 'witness shopping'"). Gelbach is correct in emphasizing that, while "[v]arious aspects of evidence and civil-procedure law disincentivize data mining," "[n]othing in the Federal Rules of Evidence or the Federal Rules of Civil Procedure . . . prevents expert mining." Gelbach, *supra* note 79, at 131–32. But, while certain aspects of the rules of evidence and procedure may discourage data fishing, they far from eliminate it; and, under certain circumstances, data fishing can be accomplished more cheaply than expert mining. In any event, the problem of data fishing via a nontestifying expert, like the problem of expert mining, raises issues regarding disclosure loopholes that permit litigants to engage in undisclosed exploration and selective reporting. It is possible that revising disclosure protections would address individual-level data fishing and expert mining simultaneously. However, applying DASS to expert testimony in the absence of revisions to the current disclosure protections leaves open the possibility of expert mining or individual-level data fishing via a nontestifying expert. *See supra* notes 89–91 and accompanying text; Gelbach, *supra* note 79, at 144–46.

principles can later present evidence of such adherence, or at least have their expert attest to it, in order to increase the credibility of the litigant's empirical evidence. Litigants should also consider the elements of DASS when cross examining an opposing party's experts and attempting to discredit their testimony.

Again, disclosure protections for nontestifying expert analysis may interfere with the objectives of DASS in the litigation setting. Disclosure rules may prevent cross examination regarding a nontestifying expert's analysis and may have implications for privilege and waiver of privilege if a litigant introduces evidence regarding a nontestifying expert's adherence to DASS. But, at minimum, experts can indicate their adherence to DASS in their own analysis and litigants can cross examine a testifying expert regarding whether she has engaged in data fishing or has relied on a nontestifying expert's exploratory analysis.[134] Additionally, as suggested above, lawmakers should reevaluate disclosure rules in light of these considerations.

In deciding whether to adhere to DASS, a litigant may have the concern that not engaging in data fishing would be too risky. After all, data fishing allows an expert to search for and select the methodology that most favors the sponsoring litigant. But, this reasoning does not hold water. Although DASS insists on developing one's methodology blindly—that is, without accessing the study's outcome data—the litigant maintains control of her case, and particularly, the material that she chooses to offer as evidence. Although following DASS requires commitment to one's methodology prior to analyzing the data, the litigant does not commit to offering the results of their expert's analysis prior to reviewing them. Instead, the litigant would review their results and then decide whether to offer them as evidence. The litigant could likely even avoid having to disclose their unsuccessful attempts to follow DASS by employing a nontestifying expert or multiple experts generally.[135]

---

[134] In certain contexts, it may be appropriate for a court to apply a procedure, based on the principles of DASS, in which parties litigate to arrive at a suitable methodology that would then be applied by all parties to a specific dataset.

[135] If the sponsoring litigant attempts to adhere to DASS via nontestifying expert analysis, it may be protected from disclosure. *See supra* notes 89–91 and accompanying text. If, however, the litigant uses a testifying expert and the failed attempt is discoverable, the party opposing the evidence can highlight for the jury that the sponsor of the evidence sought to prove their argument legitimately and only decided to resort to data fishing once realizing that the legitimate approach yields unfavorable results. Even so, however, a litigant may (unfortunately) be able to avoid disclosure of their failed attempt to adhere to DASS by hiring a new expert. *See supra* notes 132–133 and accompanying text; Gelbach, *supra* note 79, at 135–37. In any event, an expert may be able to employ legitimate data-exploration methods, discussed *supra* Part III, while remaining consistent with DASS's principles.

Furthermore, expert evidence that relies on data fishing or that fails to involve safeguards against it signals to the jury that the evidence is not trustworthy. In particular, if a litigant resorts to data fishing or cannot establish that steps have been taken to safeguard against it, the opposing litigant can and should seek to discredit the expert's claims as unreliable, regardless of whether the litigant has abandoned an earlier attempt to adhere to DASS or has simply never made any such attempt.

Thus, both parties have a strong incentive to adhere to DASS: if one party does and the other does not, the jury may credit only the evidence sponsored by the adhering party. This effect is similar to that described in Robertson's analysis in *Blind Expertise*, in which Robertson proposes that parties hire experts anonymously through intermediaries so that the experts are unaware of which party hired them and are therefore incentivized to provide unbiased opinions.[136] As is the case here, once a party learns of the opinion of their expert, they can then decide whether to introduce it as evidence.[137] As Robertson explains, a litigant's choice to introduce or exclude the evidence after learning of the expert's opinion only improves reliability:

> When both litigants in a case try the procedure, *two* experts will independently render opinions on the same case, and the procedure sends a signal to factfinders only when the two blind experts agree and one litigant discloses his favorable expert to the jury. An erroneous signal from a blind expert is thus exponentially less likely than from a single court-appointed expert. If, on the other hand, the two blind experts disagree, the jury will see neither or both of them and will thus be left in the same situation as the status quo.[138]

Similar to Robertson's proposal, whether or not an expert adheres to DASS should be an important factor in a jury's determination as to whether to credit the expert's claims; therefore, reliability is improved even when a party is permitted to attempt adherence to DASS and exclude their analysis, without having to disclose their failed attempt, after learning the results.

This method is of course distinct from Robertson's procedure of blinding experts. It is not a procedure for generally "eliminat[ing] . . . litigant-induced selection, compensation, and affiliation biases."[139] Rather, it simply aims to improve the state of expert testimony by preventing litigation parties from introducing evidence that involves invalid statistical methodology—although

---

[136] Robertson, *supra* note 4, at 179–80, 201–19.
[137] *Id.* at 215.
[138] *Id.* at 180.
[139] *Id.* at 179.

such methodology may arise, in part, from an expert's biases. It does this by providing a concrete standard, DASS, by which courts, litigation parties, and juries can judge whether an expert has taken steps to safeguard against data fishing.

Importantly, in addition to facilitating more reliable expert testimony, DASS has important implications for litigants' substantive arguments. As courts and juries become more conscious of the dangers of data fishing, litigants will become more prone to adhering to DASS, since they know that not doing so will result in the court excluding or the jury discrediting their evidence. In turn, as litigants feel more pressure to adhere to DASS, this pressure may well impact their substantive arguments. In particular, litigants may adopt more extreme positions when they know that they can support those positions through data fishing. Extremeness may, for various reasons, help a litigant to achieve their litigation goals. But, if a litigant feels pressure to adhere to DASS, then a tradeoff may arise between the extremeness of a litigant's position and their ability to support that position with data. For example, a prespecified methodology may be less likely to yield a statistically significant result in support of an extreme position relative to a more moderate position. This effect may lead to more moderate and more genuine litigation positions. It may also lead to more settlements, as parties realize that their abilities to support extreme positions are limited.

In short, empirical results produced by data fishing in litigation, like those in scholarship, are unreliable, leading to a wide range of harmful effects. Courts should consider DASS's principles in assessing reliability, and, in certain circumstances, should exclude expert evidence that fails to adhere to them. Additionally, litigants and experts should adhere to these principles in developing evidence and should discredit evidence developed by an opposing party that has not adhered to them.

CONCLUSION: WHAT IT MEANS TO ADHERE TO DASS

Data fishing invalidates statistical results by causing false positives and false impressions, and it creates an environment in which, at best, readers are highly skeptical of statistical claims and, at worst, readers base important decisions, such as policy decisions and jury verdicts, on incorrect information.

DASS is intended to serve as a framework and standard for safeguarding against data fishing. It is for both researchers and readers: researchers should follow it in their research and readers should expect adherence to it and should use it to evaluate a researcher's claims. It builds on established

statistical methods to form a framework that is concrete but sufficiently flexible to accommodate a very wide variety of research settings and conditions. Its focus is not only on taking steps to safeguard against data fishing; rather, it is also concerned with establishing evidence that such steps have been taken—that is, with substantiating a researcher's anti-data-fishing practices. This is crucial to incentivizing researchers to follow such practices, to allowing readers to evaluate statistical studies appropriately, and to reversing the norm of data fishing and replacing it with one of transparency and reliability, and one by which researchers are expected to persuade readers of a study's safeguards against data fishing—just as researchers do for other components of a study.

As such, an important component of DASS, and the substantiate element in particular, is for the researcher to attest to her adherence to DASS's principles. Let us consider what it means for a researcher to state that she has adhered to DASS.

First, note that many types of statistical analysis do not call for a researcher to follow the principles of DASS. Also, data fishing should not be confused with exploratory analysis. Exploration is important. Indeed, it is a fundamental component of scientific development.[140] A researcher engaged in data fishing, on the other hand, *hides* the exploratory context from the reader and reports statistical results misleadingly as though they arose from confirmatory analysis. It is perfectly valid to conduct exploratory analysis. Many studies are entirely exploratory in nature. Others include both confirmatory and exploratory components—sometimes described as a study's "primary" and "secondary" analyses.

The key is for a researcher to be transparent regarding the nature of the research. Additionally, it is important for the researcher and the reader to understand that exploratory analysis will produce a particularly high level of false positives and should not be treated as confirmatory analysis. Researchers should refrain from discussing exploratory results as though they are conclusive findings of statistical patterns; and readers should not understand them as such. Exploratory analysis is better used to corroborate confirmatory results, to add color to results of primary analyses, and to discover potential patterns to be examined further in separate studies.

---

[140] As noted in the NASEM Report, "some of the most important discoveries in the annals of science have come from unexpected results that did not fit any prior theory." NASEM REPORT, *supra* note 13, at 96.

Second, perfection is not necessary for a researcher to state that she has adhered to DASS. DASS signals that the researcher has followed a set of principles. If a researcher has not followed these principles but has not engaged in data fishing, the researcher should at least attest to not having engaged in data fishing. Every empirical paper that is reported as confirmatory in nature—e.g., making use of hypothesis tests—should include at least this much.

The problem with attesting only to not having engaged in data fishing is that it is ambiguous. For example, if such a statement is included in a study's report, should we assume that the researcher completed, or at least contemplated, her design prior to analyzing the data? Perhaps she began her analysis without prespecifying her design, scanned over the data informally, and then performed three or four hypothesis tests and reported the test that she decided involved the best methodology—although only *after* observing the tests' results. A well-meaning researcher could easily take these steps and innocently attest to not having engaged in data fishing. But these steps are exactly that. A well-meaning researcher may have an idea of data fishing that entails an image of a devilish character sitting in front of a computer examining thousands of methodologies to see which produces the most favorable results; as such, and perhaps with minimal training regarding the actual meaning and risks of data fishing, she may report, without any burden on the conscience, that she has not engaged in any such practice.

Attesting to DASS, on the other hand, indicates that the researcher has followed a particular set of principles. It is flexible and therefore applicable to a wide range of research settings, conditions, and researcher styles. Some researchers will follow it more meticulously than others. But, at its core, following DASS means that the researcher (1) contemplated and specified the study's design before engaging in analysis, and analyzed the data consistently with that design; (2) thought carefully about whether the researcher has been upfront regarding the study's methodology and results and took steps to avoid misleading the reader with respect to the robustness of the study's results; and (3) is willing to take and has taken steps to substantiate the study's safeguards against data fishing—at least by attesting to them.

These elements constitute the minimal standard for adherence to DASS's principles. Researchers should go beyond this standard. They should prespecify and record their study designs, they should perform thorough sensitivity analyses, and

they should take care to establish and maintain evidence of these steps—for example, a timeline and record of when the researcher completed a study's protocol and when the researcher gained access to the study's data. Researchers should discuss these steps in their reports. They should get credit for it: it strengthens their results and increases the credibility of their claims. And readers should certainly examine these details when evaluating a study and deciding whether and to what degree to credit a researcher's claims. But, adhering to DASS means, at minimum, following the principles above.

Importantly, in passive terms, the principles of DASS essentially state: do not data fish and do not mislead the reader. But DASS requires going beyond passively not data fishing and not misleading the reader. It involves proactive steps for safeguarding against data fishing.

No study is perfect. All studies will have flaws with respect to methodology and to safeguarding against data fishing. But, while there are many factors that can invalidate a statistical study, perhaps none is so widespread, so damaging, and so capable of being addressed as the problem of data fishing. To be sure, solving the big-picture problem requires completely reversing a relatively strong norm. But at the level of the particular case or the particular paper, there are concrete proactive steps that a researcher can and should take to safeguard against data fishing. And readers—whether scholars, courts, juries, or other consumers of empirical research—should expect and tolerate no less than a researcher's performance of these steps. This is the idea behind DASS.