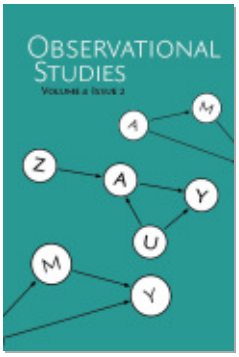
# Judging Statistical Criticism

Daniel E. Ho




Observational Studies, Volume 4, Issue 2, 2018, pp. 42-56 (Article)

Published by University of Pennsylvania Press
*DOI:* *https://doi.org/10.1353/obs.2018.0005*

Observational Studies 1 (2018) 42-56 Submitted 7/17; Published 1/18

# Judging Statistical Criticism

**Daniel E. Ho** **dho@law.stanford.edu**
*Stanford Law School*
*Stanford, CA 94305, U.S.A.*

## Abstract

Bross (1960) proposes rules for statistical criticism, chiefly that critics bear the responsibility of proving the tenability of a counterhypothesis. This Comment makes three points. First, the higher the tenability standard, the more statisticians will be drawn into the local ground rules of a substantive field. Bross feared this prospect, yet his work exemplifies it. Second, more content needs to be given to the tenability standard across domains. Proving tenability may be untenable, for instance, when data is unavailable. Third, Bross's proposal ultimately led him to espouse a quasi-judicial "adversary science" proceeding to resolve controversial issues of public policy (Bross, 1980). But Bross's own involvement in a pilot at the Nuclear Regulatory Commission illustrates the difficulties with a "science court" model, with adversarialism potentially exacerbating rather than muting political conflict. I illustrate these points with the common setting of statistical evidence in an antidiscrimination suit, using data from the University of Texas at Austin School of Law. Ultimately, Bross's work raises profound questions about the institutions for judging statistical criticism.

## 1. Introduction

The hydrogen bomb. Nuclear power. Genetically modified organisms. In the 1970s, Washington DC was abuzz with a newfangled way to resolve the most contentious public policy issues of the day: "Science Court" (Kantrowitz, 1967; Mazur, 1973). The White House organized a Task Force. The Department of Commerce, the National Science Foundation, and the American Association for the Advancement of Science co-sponsored a conference, convening scientists and policymakers to design the court — neutral scientific judges, scientific advocates presenting their case with opportunity for cross-examination, and a verdict on scientific facts — giving the green light for an experiment. Allan Mazur, one of the two principal proponents of science court, reminisced, "lunching at a small table with a Nobel laureate, the President's science advisor, and Margaret Mead was testosterone inducing" (Mazur, 1993, p. 165). One journal dubbed it "the ultimate in peer review" (Seagrave, 1976) and others referred to the idea as a "supreme court of science" (Mazur, 1993, p. 164). When asked for suitable topics, EPA Administrator Russell Train suggested global warming (Leeper, 1976, pp. 718-19).

As quickly as the model gained traction, it engendered sharp criticism. Detractors charged that there was a mismatch between adversarialism and the scientific process. Commingling science and advocacy would encourage abuse of science (Matheny and Williams, 1981). Others argued that the idea of a court judgment clashed with science as a communal enterprise. Ecologist Barry Commoner, for instance, decried the "attempt to reintroduce

Figure 1: Illustration of the "Science Court" appearing in the *New York Times*. Thanks to Carlos Llerena Aguirre for permission to excerpt.

authoritarianism in science" (Seagrave, 1976, p. 378). The *New York Times* ran a cartoon of lab instruments donning judicial robes (Figure 1).[1] Philip Abelson, editor of *Science*, questioned its likely efficacy: "You could put a bunch of scientists in white robes and they could ... make a solemn judgment of truth. And a lot of people would still think the devil is lurking out there in the Bermuda Triangle" (Seagrave, 1976, p. 379).

Irwin Bross himself was a fan. Indeed, Bross participated in one of the few pilots sponsored by the Nuclear Regulatory Commission (NRC), arguing that the scientific evidence proved negative health effects of low-level radiation exposure. Bross's advocacy for the science court, at least in his conception, was a natural extension of the ideas in "Statistical Criticism."[2] While Bross's advice was phrased in terms of professional ethics and norms in 1960, it led him to advocate for institutions to judge statistical and scientific criticism. His involvement also illustrates how adversarial science and science court ultimately fizzled.

## 2. The Virtues of Statistical Criticism

I am grateful for the invitation to comment on the re-publication of Bross's "Statistical Criticism" in *Observational Studies*. The article raises profound issues of scientific truth-seeking and how we should judge statistical criticism. Bross lamented the "superficial and sophomoric" statistical criticism levied against tobacco-cancer studies. He provided

1. John Noble Wilford, Science Considers Its Own 'Court,' N.Y. Times, Feb. 29, 1976, at 140.
2. Citing to Bross (1960), Bross (1980) makes the case as follows:

> A large number of standard put-downs are in wide use in academia. These put-downs are criticisms by innuendo rather than statements that show specific errors in the data, methods, or findings. What can be done to raise the depressingly low standards of statistical criticism and scientific controversy that now exist? What can be done to stop this kind of nonsense from delaying or blocking essential public health action? Is there any way to resolve these statistical issues more quickly and scientifically? One device that has promise is called adversary science. This device adapts to scientific questions the adversary procedure used in courtroom trials.

.

a typology of statistical criticism: (a) the hit-and-run, which points to a flaw without developing a counterhypothesis; (b) the dogmatic, which appeals to statistical theory to categorically dismiss bodies of work; (c) the speculative, which proposes a counterhypothesis but makes no attempt to reconcile it with extant evidence; and (d) the tubular, which fails to see evidence contrary to the favored hypothesis (tubular / tunnel vision). Bross argued persuasively that the responsible critic should bear the responsibility for proving the tenability of a counterhypothesis.

To illustrate this practice, Bross revisited smoking and disease data analyzed by Berkson (1958).[3] Berkson was critical of the link between smoking and cancer. He observed that because smoking appeared to be associated with a wide range of diseases, selection bias may confound smoking studies. Wrote Berkson: "I find it quite incredible that smoking should cause all these diseases" (p. 32).

Bross believed Berkson to have fallen prey to tunnel vision. To illustrate how one might prove the tenability of selection bias, Bross formalized a kind of placebo test by distinguishing "specific diseases," where etiological evidence supports a link to chemical components of tobacco, and "nonspecific diseases," where no etiological link is evident. Conducting separate tests for each category, Bross showed how Berkson's observation appeared incorrect. While specific diseases were statistically significantly correlated with smoking, nonspecific diseases were not. On its own terms, Berkson's criticism fell short of proving tenability.

Bross's contribution stands up well. Substitute any contentious policy issue (taxes and economic growth, guns and crime, universal health insurance and cost) with "smoking and lung cancer", and many of the same observations hold. The quality of statistical criticism can remain poor, "obscur[ing] a scientific discussion rather than clarify[ing] it" (Bross, 1960, p. 394). Perhaps because they are easier than full reanalyses, hit-and-run, dogmatic, speculative, and tunnel vision critiques persist.[4] Tunnel vision and motivated reasoning continue to lead to divergent conclusions on factual inferences (see, e.g., Kahan et al., 2012). Just as Ronald Fisher discounted epidemiological data on smoking (Bross, 1960, p. 396), some present day observers "raise[] randomization to the level of dogma" (id.), without being willing to contemplate any observational, descriptive, or qualitative evidence (Cook, 2015). While randomization is understandably the gold standard for causal inference, such dogma is unhelpful in areas where randomization is infeasible or unethical and where natural experiments are sparse.

## 3. Unresolved Difficulties

I offer three comments on Bross's contribution. The first is about the tension internal to Bross's "Statistical Criticism," when the tenability standard increasingly requires the statistician to engage with local ground rules. The second is about the meaning of tenability across contexts (e.g., when data is unavailable or when it is uncertain how to weight observational and experimental evidence). The third is about the path of "adversary science" that "Statistical Criticism" paved for Bross.

3. The data originally come from Doll and Hill (1956).
4. Writing in the 1990s, Mazur (1993, p. 169) opined that "the perception of undisciplined, raucous and chaotic technical controversy has dissipated" since the 1970s.

### 3.1 The Gravity of Local Ground Rules

Bross advocated that critics go beyond simply raising objections. Instead, critics should develop a counterhypothesis and prove its tenability with existing data and knowledge. While proving the tenability of a counterhypothesis is indeed worthwhile, this responsibility also potentially conflicts with Bross's admonition for the statistician to stay close to her domain. "A statistician should be especially careful . . . in the domain of the subject matter field — he is functioning as an epidemiologist or sociologist or psychiatrist . . . rather than as a statistician." In Bross's view, this was acutely the case for substantive rather than methodological counterhypotheses, "since 'local' ground rules . . . come into play." Yet the tension pervades the large swath of efforts to prove the tenability of a counterhypothesis.

Consider the Berkson example. Applying a permutation test to specific and nonspecific diseases may *seem* methodological. At the very least Bross didn't seem to classify Berkson's as a "substantive hypothesis" raising concern over the limits of the statistician's domain, suggesting that "analytical tools" can guard against Berkson's tunnel vision. But the only way for Bross to distinguish disease types (e.g., coronary thrombosis, cardiovascular disease, other respiratory disease) was to resort to local ground rules. The statistician necessarily must engage with the etiological medical evidence to understand the plausibility of the mechanism. Bross classified 15 disease categories into (a) specific, (b) questionable, and (c) nonspecific diseases. Yet how is the statistician supposed to reach such a decision without entering the substantive domain? For instance, how are we to know that "other" cardiovascular disease should be classified as nonspecific, as Bross classifies, when smoking has since been shown to affect cardiovascular disease on a range of measures (e.g., Critchley and Capewell, 2003)? What coronary events and diagnoses were included by Berkson in this category and how do we know they are not plausibly related to smoking? Bross's classification is not necessarily wrong, but his examination of Berkson's counterhypothesis illustrates that proof of tenability requires engagement with local ground rules. And the higher the standard for tenability, the greater this tension.

In later writings, as he was drawn further into the substance of the radiation debate, Bross fleshed out these concerns, posing provocatively whether statisticians should serve as scientists or be relegated to "shoe clerks." Bross worried that the path of least resistance would be to serve as a shoe clerk, by which he meant simply pleasing the customer to earn a commission (e.g., running the power calculation, fitting the model). But as scientists, applied statisticians must engage with substantive problems and criticisms (Ho and Rubin, 2011 ("To ground the assumptions [of causal inference], substantive knowledge and research are required."); Rubin, 2008 ("[N]o amount of fancy analysis can salvage an inadequate data base unless there is substantial scientific knowledge to support heroic assumptions.")). And Bross worried that doing so would raise the potential for conflicts in collaborative research settings. Administrators, for instance, may desire particular outcomes: "telling the truth can be very hazardous when it contradicts an administrator's view of things" (Bross, 1974, p. 127).

Put differently, tenability pushes one away from serving as a shoe clerk.

### 3.2 The Tenability of Tenability

It is unclear how tenability would operate across different contexts. What is the criterion by which the critic (e.g., Fisher) *should* weight observational data? In some areas, the observational data may be so limited that Fisher's dismissal of a body of evidence may in fact be warranted. The literature on the causal effect of legalized capital punishment on crime, for instance, is fraught with so many methodological challenges (e.g., highly nonrandom adoption of capital punishment and capital prosecution) that it is not obvious whether *any* observational design can ever replicate the hypothetical experiment (see Donohue and Wolfers, 2005). Can we give more content to tenability by formally incorporating prior knowledge about counterhypotheses? Bross implicitly did so when finding that the genetic hypothesis was untenable because of the rise of the male death rate.

And what does tenability require of a critic when the data may be unavailable to conduct alternative tests? Berkson published the relevant data on half a page of the *Journal of the American Statistical Association*, conveniently available for Bross's reanalysis (see Berkson, 1958, p. 34). When the underlying data is more complex and not publicly available, such reanalysis may not be as feasible. Given the rise of proprietary datasets in the age of "big data," critics may be less able to engage in the reasoned reanalysis and criticism that Bross espouses of such data.

Just as Ronald Fisher should not have dismissed the body of epidemiological data for want of randomization, the body of criticism should not be dismissed for want of reanalysis.

### 3.3 The Limits of Adversary Science

My third comment pertains to Bross's appeal to law when seeking a set of evidentiary rules. "Statistical Criticism" appeals to law in calling for evidentiary rules. The proponent has the "burden of proof" for a hypothesis. The critic has the burden to prove that a counterhypothesis is tenable. When a tenable counterhypothesis is shown, the proponent must show it to be wrong. Indeed, this process closely mirrors the legal framework in employment discrimination (disparate treatment) cases: the plaintiff must establish a prima facia claim of discrimination by preponderance of the evidence; the defendant then has the burden of rebutting the prima facia case; and the plaintiff prevails by showing that this rebuttal is wrong.[5]

In later work, Bross went further and argued that the solution for "rais[ing] the depressingly low standards of statistical criticism" lies in "adversary science" borrowed directly from the courtroom (aka "science court") (Bross, 1980, p. 37). Kantrowitz (1967) first proposed a science court, and credited Bross as providing one of several motivating proposals in the Interim Report of the Task Force for the Science Court experiment (Kantrowitz, 1977, pp. 332, 340). Kantrowitz articulated three guiding principles. First, there should be sharp separation of value judgments from judgments of scientific fact. Second, neutral, independent, scientific judges (with no prior work on the issue) would preside, with advocates presenting evidence on either side, with opportunity for cross-examination. Third, the court should issue a published decision on the state of scientific fact.

5. McDonnell Douglas v. Green, 411 U.S. 792 (1973). The legal standard is more precise by specifying the standard of proof and by distinguishing burdens of proof and production.

While much has been written about the conceptual merits and challenges of science court (e.g., Aakhus, 1999; Bazelon, 1976; Burk, 1993; Martin, 1977; Matheny and Williams, 1981; Mazur, 1993), Bross's experience offers us a concrete sense of how well the institution might foster statistical dialogue. In 1978, Bross participated in an NRC public hearing that aimed to pilot the "science court" with the subject of health effects of low-level radiation. Bross tasked himself with "establish[ing] a ... prima facie case that there are serious human health hazards from dosages of ionizing radiation in the range between 100 millirads and 10 rads" (American Chemical Society, 1978). Bross also articulated the task for Harvard epidemiologist Kenneth Rothman: "My opponent must take the position contrary to mine that there is no hazard" (id.). Echoing "Statistical Criticism," Bross stated, "It is not enough for him to argue that there might be questions or doubts ... or that there alternative interpretations" (id.). In his view, the hearing — with a neutral chair, formal presentations, and a form of cross-examination — was a success, permitting a "clear public answer" to emerge (Bross, 1980, p. 37).

Yet while Bross wrote positively of his experience, contemporaneous reports suggest that the NRC hearing was a poor exemplar of science court. No NRC judges were actually present. (Two weeks before the hearing, the Supreme Court had struck down a lower court ruling urging the NRC to create more genuine dialogue on nuclear safety,[6] possibly explaining the lack of interest in this dialogue.) Rothman, whom the NRC had engaged to assess Bross's evidence, refused to engage in adversarialism. Perhaps he objected to Bross's charge to prove a negative (that "there is no hazard"). On his account, Rothman's role was to provide an independent, unbiased review. The lack of agreement on whether the hearing was adversarial epitomizes the normative clash of policy advocacy and scientific inquiry. To make matters worse, the NRC had actually engaged Rothman to specifically evaluate Bross's report, seeing Rothman's lack of expertise in radiation as a virtue. In that sense, Rothman, lacking subject matter expertise, was less peer reviewer / adversary than science court judge. While he agreed to some extent with Bross on potential dangers of radiation, he claimed that Bross had used the data twice: both to develop and test hypotheses. Rather than a proceeding about the broad evidence base about radiation risks, the hearing focused on a specific assessment of Bross's study and reanalysis. This was far from impersonal science. Concluded Rothman: "I cannot agree that his findings warrant any revision in our thinking about the health consequences of radiation exposure."

This NRC experience is consistent with how science courts, thrust into highly politicized issues, foundered. Matheny and Williams (1981) studied the proposal for a science court to resolve a power line dispute in Minnesota. Rather than mitigating conflict, the science court turned it into a "political hot potato" (p. 355). As many had feared, separating value judgments from scientific judgments proved challenging. Even when separated, Judge Bazelon feared that the science court would obscure the ultimate importance of value judgments (Bazelon, 1976). When used for delay, the proceeding itself became political. Asked about the science court proposal for licensing two nuclear power plants, a Con Edison representative complained about yet another barrier in the regulatory process. "We had five years of hearings ... [It's] a PR kind of thing, and maybe a science court would help" but "it just adds another layer to what we have to deal with already" (Seagrave, 1976, p. 379-80).

6. Vermont Yankee v. Natural Resources Defense Council, 435 U.S. 519 (1978).

Concern about unwieldy procedural requirements animated the Supreme Court to reverse the lower court's remand for more process in nuclear licensing. And even the lower court that mandated more process expressed deep trepidation about the suitability for quasi-judicial adversarialism at the agency level. "Factual issues in hybrid proceedings tend to be complex scientific or technical ones involving mathematical or experimental data peculiarly inappropriate for trial-type procedures."[7]

These factors ultimately contributed to science court's demise. "Like a sky rocket, [the science court] got a lot of attention as it ascended but just as quickly fell downward to crash and burn" (Mazur, 1993, p. 161).

## 4. Empirical Illustration: Adversary Science in Employment Discrimination

To further illustrate these points, we consider a common setting of statistical evidence offered by opposing experts in an employment discrimination suit. This setting has formal rules for the admissibility of evidence[8] and, as mentioned above, places burdens of proof that largely mirror Bross's proposed process. Statistical evidence often plays a large role in such cases and analyses and datasets are required to be submitted to opposing parties, with opportunity for deposition, testimony, and cross-examination.

The specific example comes from the University of Texas at Austin School of Law. In December 2011, Dean Lawrence Sager resigned, amidst allegations of improper use of a foundation to compensate faculty members, including claims of gender discrimination. While there was no formal litigation surrounding these claims, the allegations were covered widely in the media and the data are representative of the kind of case that could end up in trial.

Table 1 provides descriptive statistics for the dataset on 63 faculty members, compiled from public records. For simplicity of exposition and because such techniques rarely enter the courtroom, we do not consider more advanced, and arguably appropriate, methods for causal inference here (e.g., matching methods, panel techniques, causal intermediation). The conventional posture is that each side's expert offers statistical evidence, most commonly linear regression models. Typical debates are about the sample definition, measurement, and the proper specification, each potentially suggesting or contradicting discrimination.

The first row of Table 1 shows a statistically significant salary difference between male and female faculty members. Male faculty members earn, on average, \$35k more than female faculty members ($p$-value = 0.02). Potential covariates are listed below the outcomes in Table 1, suggesting considerable differences along gender lines. (As we will see below, whether they are true covariates (i.e., unaffected by the treatment) depends on the substantive theory of the case.) Men have been teaching on average eight years longer, reflecting the diversification of the legal profession over the past few decades (Chused, 1988). Women, on the other hand, are more likely to have held positions as clerks to federal judges. Because a common concern in estimating gender discrimination is about "productivity," we aug-

7. Natural Resources Defense Council v. Nuclear Regulatory Commission, 547 F.2d 633, 656 (D.C. Cir. 1976).

8. See Daubert v. Merrell Dow Pharmaceuticals, 509 U.S. 579 (1993).

| **Outcomes** | Male mean | Female mean | Difference | $p$-value |
|---|---|---|---|---|
| Salary ($1000s) | 266.52 | 231.19 | 35.32 | 0.02 |
| Forgivable loan ($1000s) | 71.74 | 8.82 | 62.92 | 0.00 |
| **Covariates** | | | | |
| Articles | 29.17 | 21.06 | 8.12 | 0.15 |
| Years teaching | 26.02 | 17.76 | 8.26 | 0.03 |
| Endowed chair | 0.61 | 0.35 | 0.26 | 0.08 |
| Federal clerkship | 0.35 | 0.71 | -0.36 | 0.01 |
| Doctoral degree | 0.13 | 0.18 | -0.05 | 0.67 |
| $n$ | 46 | 17 | | |

Table 1: Descriptive statistics for University of Texas at Austin School of Law faculty dataset. The first column represents the mean for male faculty members; the second column represents the mean for female faculty members; the third column represents the gender difference; and the fourth column presents the $p$-value from a $t$-test. Salary and forgivable loans are in $1000s, and salary is the twelve-month salary, excluding forgivable loans. $n$ represents the number of observations.

mented this dataset with counts of the number of articles published by the faculty member from publicly available CVs. Male faculty members have written eight more articles, on average, but the difference is not statistically significant.

Imagine that the plaintiff's expert introduces a regression of salary against gender and articles. Figure 2 plots the data on articles published (logged to adjust for skewness) on the $x$-axis and salary on the $y$-axis. The lines present simple fits from the regression model, with gender (coded as 1 if male and 0 if female) and articles (logged) as predictors, with 95% confidence intervals. Red (blue) colors correspond to female (male) observations. The gender coefficient remains statistically significant, and the expert may conclude that even controlling for productivity, women are underpaid, corroborating an inference of discrimination.

Of course, many specifications, even with such a sparse covariate set, are possible. The defendant's expert may counter that this regression fails to adjust for other confounding factors (e.g., years in teaching), offering the second regression in Table 2. The gender difference is no longer statistically significant. In a narrow sense, the defendant's expert has carried out the responsibility of proving the tenability of her counterhypothesis: the gender difference in Model (1) may simply be an artifact of academic rank.

Yet in a deeper sense, the statistician is necessarily drawn into the local ground rules. Just as Bross made coding decisions of diseases based on substantive grounds, the statistician must make substantive decisions in specifying the model. Perhaps the very mechanism by which the University of Texas is discriminating against women is by awarding endowed chairs only to men, so that controlling for such covariates introduces post-treatment bias (Rosenbaum, 1984). The only way to explore this hypothesis further – particularly if the question is about gender discrimination by the dean – is to understand the substantive mechanism. Does the dean in fact exercise discretion in endowment decisions or is a committee

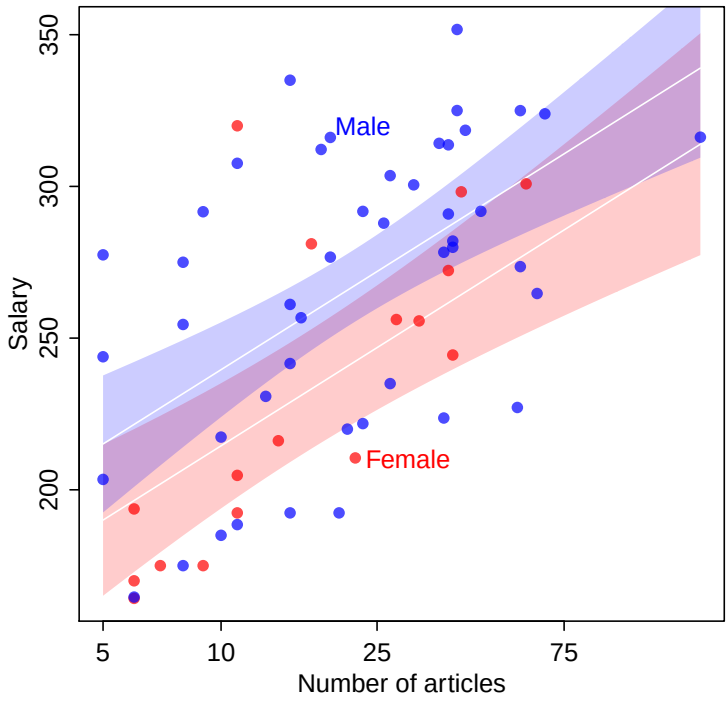



Figure 2: Correlation between number of articles (on a log scale due to skewness) and salary (in $1,000s). Red colors correspond to female faculty members, and blue colors correspond to male faculty members. 95% confidence bands from regression model are overlaid.

responsible, which might seek external letters for the promotion decision? The answer is informed at core by an understanding of the statistics of causal inference, but lies, in a sense, beyond statistics alone (Ho and Kramer, 2013; Ho and Rubin, 2011).

Consider now Models (3) and (4) of Table 2. Recall that one of the concerns that emerged was about Sager's use of a foundation, separate from the University of Texas system and not reported to the Regents, to recruit, retain, and compensate faculty members. The principal vehicle for faculty compensation constituted loans to be forgiven over a number of years. The models present the same specifications for the outcome of forgivable loan amount. Here we see that the difference increases from $64k to $91k when covariates are added and remains statistically significant in both models, suggesting that the plaintiff may have a stronger case on this outcome dimension.

To understand whether this evidence is consistent with discrimination again requires substantive knowledge. One theory of the case is that the dean used the foundation to respond specifically to outside offers to faculty (i.e., retention). On the idea that responding to such outside offers is not indicative of bias, Bross might advocate conducting separate analyses for cases of retention and recruitment. If forgivable loans are indeed merely responses to outside offers, the evidence in Models (3) and (4) may have nothing to do with

| | **Salary** | | **Forgivable loan** | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Male | 25.09** | 12.70 | 64.07** | 91.07*** |
| | (11.67) | (9.63) | (30.91) | (27.11) |
| log(Articles) | 35.31*** | 16.95*** | −3.97 | 5.10 |
| | (6.58) | (6.21) | (17.43) | (17.49) |
| Years teaching | | −0.74* | | −5.75*** |
| | | (0.42) | | (1.17) |
| Endowed chair | | 101.57*** | | 117.20** |
| | | (17.11) | | (48.16) |
| Full Professor | | 52.59*** | | 132.98*** |
| | | (15.09) | | (42.49) |
| Doctoral degree | | −7.28 | | 26.63 |
| | | (11.93) | | (33.59) |
| Federal clerkship | | −8.73 | | −9.52 |
| | | (9.56) | | (26.90) |
| Constant | 133.19*** | 147.17*** | 19.85 | 2.79 |
| | (20.74) | (19.06) | (54.94) | (53.65) |
| Observations | 63 | 63 | 63 | 63 |
| $R^2$ | 0.39 | 0.67 | 0.07 | 0.43 |

Table 2: Linear least squares regression estimates of gender discrimination. Models (1) and (2) are for the outcome of 12-month salary in $1000s. Models (3) and (4) or for the outcome of forgivable loans in $1000s. Coefficient estimates are presented with standard errors in parentheses. The salary models show sensitivity to the covariate set, with a statistically significant gender difference in Model (1) becoming statistically insignificant in Model (2). In the forgivable loan models, the gender difference magnifies as more covariates are added. Whether the predictors are true "covariates" in the sense of being unaffected by the treatment of gender depends on a substantive understanding of the theory of discrimination. *p<0.1; **p<0.05; ***p<0.01

discrimination by the dean per se. Instead, they may indicate either (a) that male faculty are more likely to seek opportunities for lateral movement, or (b) that the real parties engaging in discrimination are *other* schools by systematically making more lateral offers to male faculty.

The application also illustrates the difficult position generalist judges and juries are placed in by dueling expert reports. Precisely because experts are paid by adversarial parties, the testimony will be skewed to that side and experts may have an incentive to obscure rather than elucidate substantively important assumptions (Greiner and Rubin, 2011). Judge Richard Posner, a leading proponent of law and economics who has published econometric work, concluded, "Econometrics is such a difficult subject that it is unrealistic to expect the average judge or juror to be able to understand all the criticisms" (Posner, 1999). Yet how is a generalist judge to decide technical claims (e.g., about model fit, clustering of standard errors, principal stratification)? Even given the formal burden of proof, is it clear when the critics case has become "tenable" when the statistical evidence requires local ground knowledge? Under what conditions does such adversarial science outperform conventional statistical inquiry? Both the legal system and statistical science struggle with how to judge such statistical criticism, and the employment discrimination setting does not provide strong support for exporting adversarial science.

## 5. Conclusion

It's been a pleasure to have the opportunity to reflect on Bross's "Statistical Criticism." The article remains as relevant now as it was back then, raising profound questions about who judges whether rules of statistical criticism were adhered to.

While the science court per se may be flawed, contrasting the judicial and peer scientific models can be valuable to developing other possibilities for institutional reform. In the court room, importing scientific neutrality by greater use of court-appointed experts may make it easier for judges and juries to incorporate complex statistical evidence (Cecil and Willging, 1994).[9] In administrative agencies, importing peer review practices may improve the reliability and scientific judgment of regulatory enforcement (Ho, 2017).

On the flipside, in some circumstances, exporting elements of adversarialism might benefit statistical learning. Professional journals may incentivize higher quality science by (a) providing reviewers with datasets and replication code (just as courts mandate of experts) and (b) potential publication of discussion from reviewers to air out differences in analysis and interpretation (just as courts require opposing expert reports to be disclosed). To the extent that science court proponents were worried about representing the full range of opinions, scientific advisory committees might encourage members to write separately where their judgment of the tenability of a counterhypothesis diverges from the committee report. Scientific consensus building may benefit from forms of "adversarial collaboration," espoused by Daniel Kahneman: joint research by parties to resolve a debate, potentially with a neutral arbiter as part of the research team (Kahneman, 2003, pp. 729-30).

If properly designed and evaluated, these reforms would fall short of a full-blown science court, but could improve institutions for judging statistical criticism.

9. To be sure, there are criticisms of such techniques, as wresting power from litigants and importing a foreign inquisitorial technique (see Deason, 1998). Yet the alternative may be the mess Bross bemoaned.

Late in his career, Bross struggled with that broader question. No doubt, he was feeling embattled by the criticism of his claims for low-level radiation risk. In the *American Journal of Public Health* (Bross, 1979), he decried:

> Are the conflicts-of-interest hypothetical or real? The extraordinary editorial handling of our paper provides factual evidence on this point. The hostility of the editor to our findings is evidenced by the gratuitous comment in his introductory note: 'Dr. Bross stands virtually alone in his defense of his data.'

As critiques of his conclusions mounted (see, e.g., Boice and Land, 1979; Oppenheim, 1977; Rao, 1978), drawing Bross further into the substantive territory he had earlier warned of, he grew increasingly skeptical of the peer review system as enforcing the rules of criticism. Peers can of course have dramatically divergent standards of tenability (Simon et al., 1981), but Bross went further and charged that agencies manipulated peer review in low-level radiation "to suppress, vilify, or cut off the funding of the little scientists" (Walker, 2000, p. 95). Responding to the inability of peer review to avert the Summerlin scandal — where William Summerlin had faked tissue-culture skin transplants at the Sloan-Kettering Institute — Bross famously wrote in the *New York Review of Books*: "Big science is bad science."[10] Finding himself in the minority view on radiation risks may also have bolstered Bross's enthusiasm for adversarial science. But the judicial system itself, from which he had so heavily borrowed in proposing adversarial science, rebuffed him too. The Second Circuit affirmed a dismissal of Bross's suit against the Veterans Administration pertaining to his radiation research: "the interest of a scientist like Dr. Bross in seeking professional and governmental recognition of his views, although unquestionably genuine, [does not] fall within the [law's] zone of interest."[11]

## Acknowledgments

Thanks to Dylan Small for inviting this commentary, to Zoe Ashood, Sandy Nader, Sam Sherman, and Dylan Small for comments, and to Carlos Llerena Aguirre for permission to reprint the Science Court image.

10. Irwin D.J. Bross, A Better Mouse Trap, N.Y. Rev. Books, June 10, 1976.
11. Bross v. Turnage, 889 F.2d 1256, 1257 (2d Cir. 1989) (the specific law was the Veterans' Dioxin and Radiation Exposure Compensation Standards Act).

## References

Aakhus, M. (1999). Science court: A case study in designing discourse to manage policy controversy. *Knowledge, Technology & Policy*, 12(2):20–37.

American Chemical Society (1978). Nrc sponsors low-level radiation hazard debate. *Chemical & Engineering News*, 56(16):14. Available from: `http://dx.doi.org/10.1021/cen-v056n016.p014`.

Bazelon, D. L. (1976). Coping with technology through the legal process. *Cornell Law Review*, 62:817–832.

Berkson, J. (1958). Smoking and lung cancer: Some observations on two recent reports. *Journal of the American Statistical Association*, 53(281):28–38.

Boice, J. D. and Land, C. E. (1979). Adult leukemia following diagnostic x-rays? (review of report by bross, ball, and falen on a tri-state leukemia survey). *American Journal of Public Health*, 69(2):137–145.

Bross, I. D. (1960). Statistical criticism. *Cancer*, 13(2):394–400.

Bross, I. D. (1974). The role of the statistician: Scientist or shoe clerk. *The American Statistician*, 28(4):126–127.

Bross, I. D. (1979). Protection of the public health against radiation hazards. *American Journal of Public Health*, 69(6):609–610.

Bross, I. D. (1980). When speaking to washington, tell the truth, the whole truth, and nothing but the truth, and do so intelligibly. *The American Statistician*, 34(1):34–38.

Burk, D. L. (1993). When scientists act like lawyers: The problem of adversary science. *Jurimetrics*, 33(3):363–376.

Cecil, J. S. and Willging, T. E. (1994). Court-appointed experts. *Reference Manual on Scientific Evidence*, 527–573.

Chused, R. H. (1988). The hiring and retention of minorities and women on american law school faculties. *University of Pennsylvania Law Review*, 137(2):537–569. Available from: `http://www.jstor.org/stable/3312253`.

Cook, T. D. (2015). The inheritance bequeathed to william g. cochran that he willed forward and left for others to will forward again: The limits of observational studies that seek to mimic randomized experiments. *Observational Studies*, 1:141–164.

Critchley, J. A. and Capewell, S. (2003). Mortality risk reduction associated with smoking cessation in patients with coronary heart disease: A systematic review. *JAMA*, 290(1):86–97.

Deason, E. E. (1998). Court-appointed expert witnesses: Scientific positivism meets bias and deference. *Orregon Law Review*, 77:59–156.

Doll, R. and Hill, A. B. (1956). Lung cancer and other causes of death in relation to smoking. *British Medical Journal*, 2(5001):1071.

Donohue, J. J. and Wolfers, J. (2005). Uses and abuses of empirical evidence in the death penalty debate. *Stanford Law Review*, 58(3):791–846.

Greiner, D. J. and Rubin, D. B. (2011). Causal effects of perceived immutable characteristics. *Review of Economics and Statistics*, 93(3):775–785.

Ho, D. E. (2017). Does peer review work? an experiment of experimentalism. *Stanford Law Review*, 69:1–119.

Ho, D. E. and Kramer, L. (2013). The empirical revolution in law. *Stanford Law Review*, 65:1195–1202.

Ho, D. E. and Rubin, D. B. (2011). Credible causal inference for empirical legal studies. *Annual Review of Law and Social Science*, 7:17–40.

Kahan, D. M., Hoffman, D. A., Braman, D., and Evans, D. (2012). They saw a protest: Cognitive illiberalism and the speech-conduct distinction. *Stanford Law Review*, 64:851–906.

Kahneman, D. (2003). Experiences of collaborative research. *American Psychologist*, 58(9):723–730.

Kantrowitz, A. (1967). Proposal for an institution for scientific judgment. *Science*, 156(3776):763–764.

Kantrowitz, A. (1977). The science court experiment. *Jurimetrics Journal*, 17(4):332–341.

Leeper, E. (1976). Science court 'tried,' cleared for test case. *BioScience*, pages 717–719.

Martin, J. A. (1977). The proposed 'science court'. *Michigan Law Review*, 75:1058–1091.

Matheny, A. R. and Williams, B. A. (1981). Scientific disputes and adversary procedures in policy-making: An evaluation of the science court. *Law & Policy*, 3(3):341–364.

Mazur, A. (1973). Disputes between experts. *Minerva*, 11(2):243–262.

Mazur, A. (1993). The science court: Reminiscence and retrospective. *Risk*, 4:161.

Oppenheim, B. E. (1977). Genetic damage from diagnostic radiation. *JAMA*, 238(10):1024–1025.

Posner, R. A. (1999). The law and economics of the economic expert witness. *The Journal of Economic Perspectives*, 13(2):91–99.

Rao, P. (1978). Genetic damage from diagnostic radiation. *Investigative Radiology*, 13(1):100–101.

Rosenbaum, P. R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society. Series A (General)*, 147(5):656–666.

Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, 2(3):808–840.

Seagrave, S. (1976). The ultimate in peer review: Science court: Test case this year? *BioScience*, pages 377–380.

Simon, G. A., Cole, J., and Cole, S. (1981). Chance and consensus in peer review. *Science*, 214(4523):881–886.

Walker, J. S. (2000). *Permissible Dose: A History of Radiation Protection in the Twentieth Century*. University of California Press.