



PROJECT MUSE®

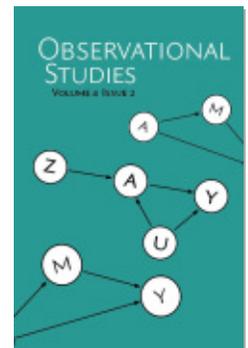
Statistical Criticism and Causality in *Prima Facie* Proof of
Disparate Impact Discrimination

William B. Fairley, William A. Huber

Observational Studies, Volume 4, Issue 2, 2018, pp. 11-16 (Article)

Published by University of Pennsylvania Press

DOI: <https://doi.org/10.1353/obs.2018.0001>



➔ *For additional information about this article*

<https://muse.jhu.edu/article/793366/summary>

Statistical Criticism and Causality in *Prima Facie* Proof of Disparate Impact Discrimination

William B. Fairley
William A. Huber
Analysis & Inference, Inc.
Springfield, PA 19064, U.S.A.

wfairley@analysisandinference.com
whuber@analysisandinference.com

Statistical criticism in a legal case need not follow the same “rules of the game” as those set out by Bross (1960) for criticism in a scientific context. The law has its own stylized requirements for criticism engaged in by experts – statistical or otherwise. For example, the law formally assigns a “burden of proof” depending on the legal theories advanced by the parties. When that burden of proof is not laid on a statistical expert he or she may make a “hit and run criticism” (see Bross (1960, p. 395)). However, insofar as the expert is in the business of being persuasive, then hit and run in many contexts will still not suffice regardless of the formal assignment of burdens.

Law has another device for handling criticism. The decision maker, such as a judge or jury, is understood to be giving different “weights of evidence” to different witnesses depending on their persuasiveness. A hit and run criticism or a counter-hypothesis that is not tenable may be given little or no weight.

Writing for judges about how scientific evidence can be best used in legal proceedings (as applied particularly for econometric regression models), Finkelstein (1973, p. 146) has proposed that judges adopt certain protocols, one of which echoes Bross’s proposal that a critic should offer a tenable counter-hypothesis. Finkelstein paints a picture of arguments converging towards “tacit agreement among the experts”:

When parties express criticisms by introducing alternative models, the process need not necessarily be a seesaw battle of conflicting econometric demonstrations. [T]here may instead be a progression towards greater refinement and correctness in statistical methodology which will not only be apparent to the decision maker, but which may also achieve results meriting at least tacit agreement among the experts.

Finkelstein echoes Bross in requiring a “superior alternative analysis” from an “objecting party”:

This experience suggests as the second protocol that (i) a party objecting to an econometric model introduced by another party should demonstrate the numerical significance [size of effect] of his objections whenever possible, and (ii) a party objecting to an econometric model of data designated by the decision maker for econometric analysis should produce a superior alternative analysis of that data.

A quantitative evaluation of effect size (“numerical significance”) will be more persuasive (although it is not always available). An example is a suit brought by high ranking officers in the New York City Police Department who had been passed over for promotion to top posts by a new Commissioner. *Courtney v. City of N.Y.*, 20 F. Supp. 2d 655 (S.D.N.Y. 1998).¹ They alleged age discrimination. While these officers were in their fifties or older, many of the top posts had been given to younger officers; it was alleged that the new Commissioner had said he wanted to sweep out “old dead wood.” On behalf of the plaintiffs, a city university professor testified that age and promotion were related because a chi-square test of the data showed a statistically significant difference in the rates of promotion between officers over 53 and officers under 53. The expert also provided data showing the older officers also had more experience (years of service), with the implication being that they were more qualified for that reason. In response to questioning by counsel for the Police Department, the expert declined to comment on whether the younger officers might have had more leadership ability, noting that while “experience” was “objective,” no objective evidence on leadership ability was at hand.² The jury (perhaps impressed by this testimony of “numerical significance,” or perhaps influenced by other facets of the evidence not presented here) found for the plaintiffs.

The expert seemed to infer age discrimination from an association between age and promotion. An obvious alternative explanation³ was that the younger officers promoted did indeed have stronger leadership abilities – at least as judged by the new Commissioner regarding his agenda of taking radical departures in policing from his predecessor. Here was an instance of the numerical data at hand appearing, at least in the mind of the expert, to drive out the arguably more relevant but non-numerical data.⁴

In employment discrimination law, and in some other contexts, there is a doctrine of disparate impact in which a plaintiff may advance a prima facie case of discrimination

¹One of the authors of this article testified on behalf of the Police Department.

²In other contexts, such as disparate impact in employment discrimination, courts have accepted subjective criteria. For example, in *Watson v. Ft. Worth Bank & Tr.*, 487 U.S. 977 (1988), the court wrote:

In the context of subjective or discretionary employment decisions, the employer will often find it easier than in the case of standardized tests to produce evidence of a “manifest relationship to the employment in question.” It is self-evident that many jobs, for example those involving managerial responsibilities, require personal qualities that have never been considered amenable to standardized testing. In evaluating claims that discretionary employment practices are insufficiently related to legitimate business purposes, it must be borne in mind that “[c]ourts are generally less competent than employers to restructure business practices, and unless mandated to do so by Congress they should not attempt it.” (“[The] criteria [used by a university to award tenure], however difficult to apply and however much disagreement they generate in particular cases, are job related....It would be a most radical interpretation of Title VII for a court to enjoin use of an historically settled process and plainly relevant criteria largely because they lead to decisions which are difficult for a court to review”). (citations omitted)

³This explanation is supported by the *Peter Principle* (Peter and Hull, 1969), which posits that “every new member in a hierarchical organization climbs the hierarchy until he/she reaches his/her level of maximum incompetence.” Accordingly, older members will tend to be less competent and less qualified for promotion than younger members, especially in the higher ranks of the hierarchy. Simulations bear out this conclusion: see, *inter alia*, Pluchino et al. (2009).

⁴Tribe (1971), in an influential law review article, complained that mathematical or statistical data and models could over-awe the legal decision maker, especially juries, and unfairly take the place of more subjective but more relevant evidence.

by demonstrating an imbalance in the proportion of a protected class (distinguished, for example, by age, race, or sex) among those gaining some advantage such as promotion or not being laid off.⁵ Usually such an imbalance must be shown to be statistically significant and not negligible.⁶ Thus, a numerical difference between proportions of minority and majority class members being laid off may in some instances be taken as sufficient for the case to proceed to a further stage.

Most of the time, stronger evidence is needed. As the Supreme Court noted in [*Watson v. Fort Worth Bank & Tr.*, 487 U.S. 994 (1988)], the Supreme Court’s “formulations...have consistently stressed that statistical disparities must be sufficiently substantial “that they raise...an inference of *causation*.” In other words, the statistical disparities must “show that the practice in question has *caused* the exclusion of applicants for jobs or promotions because of their membership in a protected group” *Id.* (emphases added).⁷ See also *Stagi v. National R.R. Passenger Corp.*, 391 F. App’x 133, 136 n.3 (3d Cir. 2010).

A statistical difference is not always enough to establish even a *prima facie* case. Other factors than age could explain an association between age and being laid off. *Sheehan v. Daily Racing Form* was a case in which a layout editor at a newspaper in Chicago was laid off after the introduction of computer methods replaced his work.⁸ 104 F.3d 940 (7th Cir. 1997). A greater proportion of older than younger employees were laid off. Table 1 gives the data ($\chi^2 = 9$, $p = 0.009$ (Fisher’s Exact Test); no one was between 42 and 48).

In his opinion in *Sheehan* Judge Richard A. Posner wrote:

More important is the expert’s failure to correct for any potential explanatory variables other than age. Completely ignored was the more than remote possibility that age was correlated with a legitimate job related qualification, such as familiarity with computers. Everyone knows that younger people are on average more comfortable with computers than older people are

⁵The plaintiff must also identify an action or policy that (allegedly) caused the imbalance. See *Wards Cove Packing Co. v. Atonio*, 490 U.S. 657 (1989).

[A]...plaintiff does not make out a case of disparate impact simply by showing that, “at the bottom line,” there is racial imbalance in the work force. As a general matter, a plaintiff must demonstrate that it is the application of a specific or particular employment practice that has created the disparate impact under attack. Such a showing is an integral part of the plaintiff’s *prima facie* case in a disparate-impact suit under Title VII [of the Civil Rights Act of 1964].

⁶Courts have placed an emphasis in finding discrimination on establishing the statistical significance of a difference. That emphasis is traced to the Supreme Court’s decision in *Castaneda v. Partida*, 430 U.S. 482, 496 n.17(1977): “As a general rule for such large samples, if the difference between the expected value and the observed number is greater than two or three standard deviations, then the hypothesis that the jury drawing was random would be suspect to a social scientist.” The court did not comment on the importance of the model or assumptions underlying a test of statistical significance. In practice, these are often ignored. For example, see cases discussed in Sugrue and Fairley (1983, pp. 953-955), especially *Bryan v. Koch*, 627 F.2d 612, 616-618 (2d Cir. 1980).

⁷Showing of causation is the key requirement. There is, however, the statement that causation can be shown by “sufficiently substantial” disparities. We have already stated above that disparities must be substantial in certain ways to demonstrate a *prima facie* case, There is the further suggestion here that the more substantial the disparities the better the demonstration of causation. This may be, but is not always, the case.

⁸The plaintiff in this case advanced not only a legal claim of disparate impact but disparate treatment (intentional discrimination).

Table 1: Laid Off Versus Age

Age	Laid Off?		Total
	Yes	No	
≤ 42	0	6	6
≥ 48	9	3	12
Total	9	9	18

Source: *Sheehan vs. Daily Racing Form*

on average more comfortable with manual shift cars than younger people are. 104 F.3d 940, 942 (7th Cir. 1997).

Bross (1959, p. 395 left hand side) wrote:

When a counterhypothesis involves a well-known real factor, e.g., age or sex in an epidemiological study, it would be sufficient to mention the relationship, e.g., death rates from cancer tend to increase with age.

Posner might say that his counter-hypothesis (computer facility was a factor correlated with age) is obviously superior to the hypothesis of age discrimination, which was supported by nothing more than a recital of who was laid off and their ages.

Also, see *Diehl v. Xerox Corp.*, 933 F. Supp. 1157, 1167-68 (W.D.N.Y. 1996).

Although both experts concluded that the redeployment and dismissal rates among certain portions of the [employer's] workforce who were 40 years of age or older and [were] male employees were greater than those of other employee groups which disparities could not be caused by chance factors, Dr. Honig failed to conduct further statistical tests to determine what other factors *could* have accounted for those disparities. Instead, she concluded that since the disparities were not caused by chance, the high probabilities (1 in 100) demonstrated that they were caused by age or gender. Without conducting any other statistical tests to rule out factors other than age or gender and by relying solely on age and gender ratios, Dr. Honig's testimony is fatally flawed pursuant to *Wards Cove*.

Nevertheless, it is not uncommon to see experts rely solely on a statistical difference to support a *prima facie* case of discrimination. This is in marked contrast to the view in science or social science that correlation, whether statistically significant or not, does not (by itself) establish causality.⁹ From a scientific perspective, statistical significance is not

⁹To cite just one example, Madden (1985, p. 80) writes of how finding sex discrimination in employment requires by definition adjusting for other factors that could account for pay differentials:

Empirical work by economists has concentrated on measuring and accounting for sex differentials in productivity. The basic procedure used in empirical investigations of sex discrimination is to determine the magnitude of the pay differential *after adjusting for quantifiable sex differences in productivity-enhancing characteristics*. (emphasis supplied)

as important a desideratum in finding a difference – that can at least in part be assigned to being a member of a protected class – as “causal significance,” if by this phrase is meant: “a causal relation between two variables is supported.”¹⁰ Establishing causal significance requires a very different and more extended inquiry than showing statistical significance.

In discrimination cases, we see a marked tendency for testifying experts to draw causal conclusions from tests of association, which of course goes beyond what such tests can support. For example, in cases of age discrimination in promotions, it is common to see experts opine based only on statistical differences that the effects are “age-based” or that age must be a “factor” in creating the differences. Along the same lines, it is common to see experts interpret a null hypothesis of independence or no association as a causal hypothesis. For example, in a sex discrimination case the null hypothesis might be described as a “gender-neutral” policy, suggesting that the alternative hypothesis is the policy is “gender-biased.” Often such characterizations are unaccompanied by any evidence suggesting actual bias, and in many cases the absence of such evidence is actually stipulated. Similarly, the term “effect” is used, as in “the effect of race,” to suggest to a lay audience a causal effect of a variable (like age, race, or gender) without having established causal significance and without explicitly claiming the effect is causal.

A good causal argument examines more than one variable: in Bross’s terms, any variables suggested by plausible counterfactual hypotheses could be incorporated in the analysis. Courts have not, however, required that all possible explanatory variables be included to support a causal conclusion. *Diehl v. Xerox Corp.*, 933 F. Supp. 1167 (W.D.N.Y. 1996) is a case alleging age discrimination in employment following a reduction in force. The court wrote (933 F. Supp. 1157, 1169 (W.D.N.Y. 1996)):

[Dr. Bloom testified that]...the probative value of a regression analysis is not eliminated simply because the analysis uses fewer than all critical variables. In *Bazemore v. Friday*, the Supreme Court reached the same conclusion:

Importantly, it is clear that a regression analysis that includes less than all measurable variables may serve to prove a plaintiff’s case. A plaintiff in a Title VII suit need not prove discrimination with scientific certainty; rather, his or her burden is to prove discrimination by a preponderance of the evidence.

From our point of view as sometime expert witnesses (and critics thereof) it often is unclear whether a demonstration of causality is required to support a given case. We see instances where experts equate association with causation. In some cases, but not all, courts criticize (or even exclude) experts for such interpretation. However, our experience is consistent with Bross’s thesis insofar as judges and juries are more likely to be persuaded when a tenable counter-hypothesis can be supplied than when the only critical rejoinder is the mere suggestion that another factor might explain an outcome. However, the situation is not symmetrical between plaintiffs and defendants. A plaintiff always has the burden of showing causation; a defendant does not have this burden and only needs to rebut the plaintiff’s case – without necessarily even suggesting other factors than those advanced by plaintiff.

¹⁰Tinkham (2010), a legal commenter, suggests that a *prima facie* case need only be supported by some relevant evidence and need not require a demonstration of causality. In practice, as in *Sheehan* and *Diehl*, we see cases where a failure to demonstrate causality called down unfavorable opinions from judges.

References

- Bross, I. D. J. (1960). Statistical criticism. *Cancer*, 13, 394-400.
- Finkelstein, M.O. (1973). Regression Models in Administrative Proceedings. *Harvard Law Review*, 86, No. 8.
- Madden, J.F. (1985) The Persistence of Pay Differentials: The Economics of Sex Discrimination. In Laurie Larwood et al, Editors, *Women and Work: An Annual Review*, Volume 1, Sage.
- Peter, L. and Hull, R. (1969). *The Peter Principle. Why Things Always Go Wrong*. William Morrow.
- Pluchino, A., Rapisarda, A. and Garofalo, C. The Peter Principle Revisited: A Computational Study. arXiv:0907.0455 [physics.soc-ph].
- Sugrue, T.J. and Fairley, W.B. (1983). A Case of Unexamined Assumptions: The Use and Misuse of the Statistical Analysis of Casteneda/Hazelwood in Discrimination Litigation. *Boston College Law Review*, July 1983.
- Tinkham, T. (2010). The Uses and Misuses of Statistical Proof in Age Discrimination Claims. *Hofstra Labor and Employment Law Journal*, 27, Issue 2.
- Tribe, L.H. (1971). "Trial by Mathematics: Precision and Ritual in the Trial Process." *Harvard Law Review*, 84, No. 6.

Cases Cited

- Bazemore v. Friday*, 478 U.S. 385 (1986).
- Castaneda v. Partida*, 430 U.S. 482 (1977).
- Courtney v. City of N.Y.*, 20 F. Supp. 2d 655 (S.D.N.Y. 1998).
- Diehl v. Xerox Corp.*, 933 F.Supp. 1157 (W.D.N.Y.1996)
- Sheehan v. Daily Racing Form, Inc.*, 104 F.3d 940 (7th Cir. 1997).
- Stagi v. National R.R. Passenger Corp.*, 391 F. (3d Cir. 2010).
- Wards Cove Packing Co. v. Atonio*, 490 U.S. 657 (1989).
- Watson v. Ft. Worth Bank & Tr.*, 487 U.S. 977 (1988).