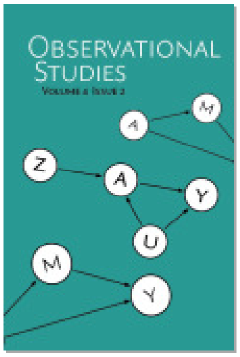
The Potential Usefulness of Bross's Principles of Statistical Criticism for the Evaluation of Statistical Evidence in Law and Public Policy

Joseph L. Gastwirth

Observational Studies, Volume 4, Issue 2, 2018, pp. 17-31 (Article)

Published by University of Pennsylvania Press
DOI: *https://doi.org/10.1353/obs.2018.0002*



Observational Studies 4 (2018) 17-31 Submitted 12/17; Published 1/18

The Potential Usefulness of Bross's Principles of Statistical Criticism for the Evaluation of Statistical Evidence in Law and Public Policy

Joseph L. Gastwirth **jlgast@gwu.edu**
Department of Statistics
George Washington University
Washington, DC 20052, USA

1. Introduction

The classic paper by Bross (1960) should be read in conjunction with Cornfield's inequality, in the appendix of Cornfield, Haenszel et al. (1959) and described in Gastwirth (1988), Greenhouse (1982), Greenhouse (2009), Rosenbaum and Krieger (1990) and Rosenbaum (2002). The inequality states conditions that a suggested omitted variable needs to satisfy in order to "explain" a difference in the proportions of successes (or failures) between two groups. Many extensions of the original result allowing for sampling error or matched pairs and other designs have been developed over the years (Rosenbaum, 2002; Guo at al. 2013) and suggested for use in legal cases (Gastwirth, 1992). Although one may question the thoroughness of the analysis of the data in Table I of the article, the criteria Bross gives for statistical criticism remain relevant today.[1]

This commentary will focus on the applicability of the framework suggested by Bross for the evaluation of criticisms of statistical evidence in law and public policy as other commentators will discuss the advances in statistical methodology that provide a more comprehensive analysis of epidemiologic and related data sets. Section 2 reviews the role of statistical evidence in discrimination cases. These cases are brought under civil, rather than criminal law, so the trier of fact (jury or judge) decides the case based on the preponderance of the evidence or "more likely than not" standard, rather than the "beyond a reasonable doubt" standard used in criminal cases. Section 3 discusses how courts have considered criticisms of statistical evidence. Because our focus is on the usefulness of Bross's principles, some important legal aspects, such as whether the cases discussed in Section 3 concerned an appeal of a summary judgment or were a class action will not be emphasized.[2] Section

[1] Table I examines the death rates of many diseases, however, most toxic agents only affect one or a few diseases, e.g. workers exposed to benzene have an increased risk of leukemia. The sign test gives equal weight to each of the types of mortality, although epidemiologic studies had shown that smoking had a strong association with lung cancer. Statistical methods designed to detect trends in dose-response data, e.g. the Cochran-Armitage (1954, 1955) test or its extension (Mantel, 1963) to stratified data or combining the results of several studies would be more powerful.

[2] When a party moves for summary judgment, it is claiming that the opposing party has no case and the entire case should end. Summary judgment is warranted when "the pleadings, depositions, answers to interrogatories and admissions on file, together with the affidavits, if any, show that there is no genuine issue

4 describes the main studies that led to warning the public about the association between the use of aspirin to treat children with colds or chicken pox and their risk of subsequently developing a rare but serious disease, Reye Syndrome. Because the industry was able to raise questions about the early studies, without being held to the criteria stated by Bross (1960), slightly over three years elapsed from the time the FDA (November, 1982) felt the public should be notified and the start of the warning campaign in the United States in early 1985.

2. The Role of Statistical Evidence in Disparate Impact and Disparate Treatment Discrimination Cases

There are two categories of EEO cases, disparate impact and disparate treatment. Disparate impact cases concern the legitimacy of a job requirement, e.g. passing a written or physical test or possessing a certain level of education. When the proportion of applicants from a legally protected group, typically a race-ethnic minority or females, satisfying the requirement is significantly less than the corresponding proportion of majority applicants, an employer has the opportunity to discredit the plaintiffs' analysis by showing that the data contain serious errors or omit a major relevant variable. If the employer cannot show that the plaintiffs' statistics are defective, then they need to demonstrate that the requirement is necessary for the job. If the employer can demonstrate that the requirement is job-related, the plaintiff is given the opportunity to suggest an alternative criterion that achieves the objective of the requirement at issue but has less of a disparate impact. This approach was established by the Supreme Court in *Griggs v. Duke Power*, 401 U.S. 424, 91 S.Ct. 849 (1971).

In *International Brotherhood of Teamsters v. United States*, 431 U.S. 324 (1977), the U.S. Supreme Court defined disparate treatment as discriminatory acts in which "[t]he employer simply treats some people less favorably than others because of their race, color, religion, sex, or national origin." The plaintiff needs to show that the employer's decision was motivated by the employee's membership in a protected group and statistical evidence is relevant to the process of shifting burdens of production during the proceedings:

1. The plaintiff has a burden of establishing a *prima facie* case that, left unrebutted, raises an inference of discrimination.

2. If the plaintiff establishes a *prima facie* case, the defendant has the burden of producing a legitimate and non-discriminatory reason for its action.

3. Then the plaintiff has the burden of showing that the non-discriminatory justification given by the defendant is a mere pretext.

as to any material fact and that the moving party is entitled to judgment as a matter of law." Fed. R. Civ. R. Civ. P. Rule 56(c). Courts evaluate the evidence giving the opposing party, who would lose the case at that point, the benefit of the doubt. For example, even if the trial judge thinks the evidence favors the moving party, if a reasonable jury could view it in favor of the opposing party, summary judgment should not be granted. The seminal case in the U.S. is *Anderson v. Liberty Lobby, Inc.*, 477 U.S. 242, 242-243, 106 S. Ct. 2505, 2511 (1986). Thus, a party can prevail at summary judgment but still lose the case. Because class actions concern a large number of plaintiffs and usually involve a large amount of money, plaintiffs need to satisfy special rules, see Fed. R. Civ. Proc. 23. Statistical evidence is often used to show there is a common question, e.g. underpayment of minority or female employees.

In both types of cases, statistical evidence may be used by a plaintiff at the first or third stage and may be used by the defendant at the second stage to support its claim that its job requirement is related to successful performance of the job or its employment decisions arose from legitimate considerations. After the plaintiff submits statistical evidence at the first stage, the defendant may rebut the analysis by showing that it is flawed. Similarly, at the second stage of either type of case the plaintiff may point out defects in the defendant's analysis.

Using the framework suggested by Bross, at the first stage the plaintiff has the role of the *proponent* of a scientific hypothesis. Typically, their evidence will be a comparison of the success rates or average wages of minority and majority success rates, often adjusting for the effect of other appropriate factors, such as seniority or relevant educational background, with a regression or stratified analysis. Then the defendant has the role of the critic and the protocol suggested by Bross is especially relevant as it implies that courts should require the defendant to

1. Specify the deficiencies in the plaintiff's statistical presentation.

2. Show that they are sufficiently severe that they raise serious doubts about the inference made by the plaintiff's expert from the statistical analysis.

If, after taking into account the defendant's criticism of the plaintiff's analysis and the other evidence, the judge decides that the plaintiff's evidence is still sufficient to establish a *prima facie* case of disparate treatment, the defendant becomes the *proponent* of the hypothesis that any disparity is due to legitimate factors, rather than discrimination. Often they will submit a more detailed analysis, e.g. by incorporating more legitimate factors or stratifying the data into more appropriate subgroups.[3] At the similar stage of a disparate impact case, the defendant needs to submit a validation study showing the job requirement at issue is correlated with performance of the job. After the defendant submits its evidence, both statistical and non-statistical, the plaintiff has the role of the *critic* and courts should require them to offer criticisms meeting the above two criteria.

In disparate impact cases, if the defendant demonstrates that the job requirement at issue is job related, the plaintiff is given an opportunity to propose an alternative requirement that has less of a disparate impact but achieves the purpose of the original one, in selecting successful employees. Now the plaintiff is the *proponent* of an alternative requirement and the defendant may *criticize* the proposal by showing that it is not have the same predictive ability of on the job success as the original requirement.

[3] For example, in a promotion case, if an employer has several locations, say k, and typically promotes from within each of them, the defendant may rebut a plaintiffs' analysis of the data from all locations aggregated into a single 2x2 table, by stratifying the data into k sub-tables and showing that the Cochran-Mantel-Haenszel test is not significant.

3. The Applicability of these Principles of Statistical Criticism to Equal Employment Cases

The Supreme Court's discussion of the regression analyses[4] submitted by both parties in *Bazemore v. Friday*[5] illustrates how the principles of statistical criticism could assist the courts when they assess statistical evidence. The lower courts had rejected plaintiffs' regression analysis submitted to support their request that a class action be certified for the claim of the plaintiffs, black employees of the North Carolina Agricultural Extension Service, of discrimination in salaries and in the performance evaluation system. Prior to the Civil Rights Act of 1965, the service paid black employees less than whites and a secondary issue was whether pre-Act disparities in pay that continued after the Civil Rights Act was passed violated the new law.[6]

The expert for the United States analyzed salary data for 1974, 1975 and 1981 by regressing pay on race, education, tenure and job title because an official at the Service stated that salaries were determined from four factors, education, tenure, job title and job performance. He found that black employees earned $331 less than whites did in 1974, the disparity was $395 in 1975 and these were statistically significant at the .05 level. While there was disparity disadvantaging blacks it 1981, it was not statistically significant.

The defendant submitted a regression analysis for 1975 using similar predictors and found a significant disparity of $384, quite similar to those found by the plaintiffs. Furthermore, when they included quartile rank for that year, the disparity increased to $475.[7]

Apparently, the defendant offered a number of criticisms of the plaintiffs' regression, which the lower courts accepted. In particular: the County Chairman who were mainly white might skew the data to show whites earning more than blacks, several variables relating to job performance and variation in pay between counties. The Court noted that job title was included in plaintiffs' regressions, so the inclusion of County Chairman would be accounted for and that defendant's own regression showed that including job performance as reflected in the quartile rankings, did *not* explain the disparity.

The defendant did *not* submit evidence to support its claim that the disparities could be explained by county variations. The government, however, did present evidence that black employees were not located disproportionately in counties that contributed only a small amount to employee salaries.[8]

[4] There is a large literature concerning the use of statistics and regression in legal cases. See Gastwirth (1988), Fienberg (1989), Zeisel and Kaye (1997) and Finkelstein, and Levin (2000). Gray (2009) and Hersch and Bullock (2014) discuss how courts have evaluated data from many cases and suggest that they may under-weight plaintiffs regressions by giving too much credence to defendants' criticism. Greiner (2009) reviews the applicability of causal inference methods in civil rights cases. Sinclair and Pan (2009) and Graubard (2009) discuss the use of Peters-Belson regression, which fits a model to the majority group and compares the actual outcomes (e.g. pay or promotion) of the protected group to its predicted value from the majority equation

[5] 478 U.S. 385 (1986)

[6] The decision, Id. at n.8 noted that if the pre-1965 pay disparities continued, the employer violated the law. For employees hired afterwards, plaintiffs would need to provide evidence that new disparities were created. Prior to 1965, the Service maintained two racially segregated branches and paid black employees less than white employees.

[7] The defendant's expert said he was unable to explain why adding those rankings increased the disparity but mentioned that the ranking data was missing for 20% of the employees.

[8] The Extension Service was funded by money from the national, state and local governments.

The Court stated that an analysis, which accounts for the *major* factors, is acceptable and that failure to include additional predictors would affect its probativeness, rather than its admissibility as evidence. In a footnote, the opinion stated that a regression analysis could be so incomplete that the results would not be admissible, but this concern did not apply to plaintiffs' evidence in *Bazemore*. The Court rejected the standard adopted by the appellate court that the plaintiffs' regression should include *all* measurable factors. The Court reminded the lower courts that plaintiffs are not required to prove discrimination with scientific certainty; their burden is to prove discrimination by the preponderance of the evidence.

Had the courts known of the principles of statistical criticism in Bross (1960), they would have required the defendant to do more than suggest some factors, especially pay differentials among the state's counties; they would expect the defendant to offer evidence in support of that claim. Thus, the plaintiffs would not have needed to show that black employees were not concentrated in counties that contributed a small amount of funds.

Comments: (1) The evidence in the case is somewhat unusual in that the Service submitted a regression that included a possible explanatory variable reflecting job performance that actually increased the disparity. Normally, a defendant would not suggest a factor that increased a disparity as an explanation for it. (2) The Supreme Court did not discuss the data on quartile rankings as the appellate opinion examined the data for 1981 in each of the five (of six) districts with agents of both races, separately.[9] According to the appellate court, 751 F. 2d 662 (1984) at 673-74, the disparity in the proportions of black and white employees receiving evaluations in the lowest quartile, who would not receive a merit raise, were *not* statistically significant in *any* of the five districts.[10] Even though the proportion of blacks being ranked in the lowest quartile exceeded the proportion of whites in the five districts, the court apparently required that statistical significance be observed in at least one-half of the districts. The proper analysis uses the Cochran-Mantel-Haenszel method for combining 2x2 tables and finds a statistically significant difference at the .01 level (Gastwirth, 1988, p. 267-78), and the odds a black employee ranked in the lowest quartile was 2.17 or twice those of a white.

3.1 Examples of cases where judges adopted an approach similar to Bross's principles

The plaintiffs in *Randall vs. Rolls Royce*[11] were women who complained that they received less pay than similar male employees did. After the district court rejected their request that the case be certified as a class action, their appeal was considered by the 7th Circuit. The company established broad pay ranges for compensation categories for employees in jobs that were of equal value to it. In order to meet competition for job types that were in

[9] There were no black employees in one district.

[10] The court used a t-test to compare the proportions. Had it used Fisher's exact test, it would have observed a statistically significant result in the Northwest district as the test yields a two-tailed p-value of .04. More importantly, one should examine stratified data by an appropriate combination test, e.g. the Cochran-Mantel-Haenszel test after checking that the odds ratios are similar. Gastwirth, Miao and Pan (2017) reanalyze stratified data from an actual case illustrating the methodology and providing references to the literature.

[11] 637 F 3d 818 (7th Cir. 2011).

demand, the company created an additional narrow range for each category to allow it to match the "prevailing market wage", i.e. wages offered by other employers in the area.

In 2003, the year just before the complaint period, the plaintiffs noted that the average base pay of male employees in the five compensation categories relevant to the case was 5% higher than that of females. This differential persisted throughout the complaint period and if the difference were attributable to sex, the firm's failure to eliminate it would perpetuate discrimination and violate the law. The plaintiffs apparently submitted a regression analysis. By including the type of job in a more comprehensive regression, the defendant showed that gender was no longer a significant predictor of salary.

The opinion mentions that the plaintiff's expert made other errors, in addition to failing to adjust for differences in the jobs occupied by male and female employees. He included employees hired after the beginning of the complaint period, which did not make sense since the claim was that females were discriminated against because the company failed to erase a disparity that existed at the outset of the period. Moreover, he did not make a study of the reasons for differences in the starting salaries of these male and female hires.

The opinion followed the precepts of Dr. Bross, as it did not simply accept a suggestion that the type of job could possibly explain the 5% disparity in the salaries of male and female employees; rather it required the defendant to submit an appropriate regression. Had the plaintiffs found a significant gender difference in an analysis that included job type, then the issue of the coverage of the database used by the plaintiffs should be explored, i.e., the analysis should be redone by examining the appropriate employees. Irwin Bross' article, "Statistical Criticism," gives advice that is surprisingly current, given that it appeared in the journal *Cancer* nearly sixty years ago. Indeed, the only obviously dated aspects of this paper are the use of the generic male pronoun and the sense that it was still an open question whether cigarette smoking caused lung cancer.

The *Allen v. Seidman*[12] case concerned the disparate impact on African-American employees of a written exam used to determine whether to promote bank examiners employed at the FDIC. Fourteen of the 36 or 38.9% of the African-Americans passed, while 329 of 391 or 84.1% of Whites did. The pass rate of African-Americans was less than one-half that of Whites and the difference is highly significant.[13] The defendant suggested that the minority candidates might have had lesser qualifications, e.g. education. The opinion noted that all of the examinees had worked for five to fifteen years and at least one year at their current grade level at the agency and had obtained a recommendation from their regional director. Thus, the minority and White candidates appeared to be reasonably homogeneous. The defendant did not submit any evidence showing an imbalance between the two groups with respect to a job-related factor, e.g. seniority or education. The opinion continues with "since the defendant, while taking pot shots – none fatal at the plaintiffs' statistical comparison, did not bother to conduct its own regression analysis, which for all we know would have confirmed and strengthened the plaintiffs' simpler study."

The *Allen* opinion is a good example of a court following sound principles of statistical criticism. Even its description of the defendant's suggestions of possible explanatory factors,

[12] 881 F.2d 375 (7th Cir. 1989) upholding the district's finding of disparate impact in Allen v. Isaac. 39 FEP Cases, 1142 (N.D. Ill. 1986).

[13] Fisher's exact test yields a p-value of about 10^{-8} or less than one in a million.

without submitting evidence that they could explain the disparity, as taking "pot-shots" is similar to the words "hit and run" used by Bross.

In support of a claim of sex discrimination *EEOC v. General Telephone Co. of Northwest*, 829 F.2d 885 (9th Cir. 1989), the Commission's expert introduced a regression including the legitimate factors, showing that gender was statistically significantly negatively related to pay. The defendant did not submit a regression including additional variables or submit a different statistical analysis. Rather, it argued that differences in job interest among men and women would explain the disparity. The District court accepted this explanation and found for the defendant. Consistent with Bross's criteria for a critic, the Ninth Circuit reversed the trial court and remanded the case for reconsideration. The opinion notes "GenTel had to produce credible evidence that curing the alleged flaws would also cure the statistical disparity – proof which GenTel did not offer. Thus, we hold that the district court erred in uncritically accepting GenTel's assertion that the plaintiff's analyses were flawed where GenTel had failed to show that if the EEOC had "adequately" accounted for the alleged flaws, the disparities in its analyses would have been eliminated."

The statistical evidence in *Sheehan v. Purolator*, 839 F.2d 99 (2nd Cir. 1988) concerned whether female exempt employees, as a class, were discriminated against in pay, job assignment and promotion. The plaintiffs submitted a regression analysis, which showed disparities in pay between male and female employees.[14] The opinion notes that the defendant supported its claim that plaintiffs' regression was flawed because it did not include education and prior experience by *introducing* evidence that these factors indeed could explain the disparities.

3.2 A case where courts accepted "explanations" without requiring evidence they could explain the disparity

The statistical evidence in the *Equal Employment Opportunity Commission v. Sears, Roebuck & Co.*, 839 F.2d 302 (7th Cir. 1988) sex discrimination case focused on whether females had the same opportunity as males to be hired or promoted to commissioned sales jobs, which had more risk but generally had higher pay than non-commissioned sales positions. Plaintiffs' expert selected six factors that he thought might affect an applicant's chance of selection: (1) job applied for; (2) age; (3) education; (4) job type experience; (5) product line experience; and (6) commission product sales experience. A logistic regression including gender and these variables yielded a statistically significant negative coefficient for gender when the model was fit to national data, but not for all regions or territories

In addition to questioning the coding of some of the predictors, Sears argued that the predictor variables chosen by the expert did not include major variables such as interest in a commission sales position and females had less relevant experience and less education. Furthermore, some characteristics such as assertiveness, friendliness and motivation are better evaluated in an interview and cannot be obtained from a written application form. The appellate court thought that the experience variables used by the plaintiff were adequate but that Sears provided sufficient evidence from national surveys and some limited studies

[14] The appellate opinion does not report the independent variables used to predict an employee's salary.

by the company, demonstrating that women had noticeably less interest in the type of position than men did.[15]

Apparently, the courts did *not* require Sears to demonstrate that the general difference in interest in commissioned sales jobs, which was also true for individuals who actually applied, was of sufficient magnitude that it and incorporating years of relevant experience, could explain the disparities. The dissenting opinion of Judge Cudahy states, "Perhaps the most questionable aspect of the majority opinion is its acceptance of women's alleged low interest and qualifications for commission selling as a complete explanation for the huge statistical disparities favoring men."

This case has important implications for statisticians, who might become involved as an expert witness. The courts gave less credibility to the statistician's choice of predictors because he was not a labor economist and they were not necessarily the major predictors considered by Sears. Thus, it is important for a statistician to be involved in discovery in order to learn the criteria actually considered by an employer and obtain the information on those factors.

3.3 Two cases illustrating the need for principles of criticism to consider the availability of data on the major variables and for statistical experts to be given all the relevant data

In a very useful review of the need for statistical evidence to meet the standards of reliability set out by the Court in Daubert and the Federal Rules of Evidence, Rosenblum (2015) summarizes several cases where the plaintiff's evidence was insufficient.[16] Rosenblum (2015) discusses how courts have implemented these standards in equal employment cases. Typically, this occurs when the statistical presentation does not account or adjust for known job-related factors. Sometimes plaintiffs' expert has pooled data from a wide variety of positions or locations into one large sample, e.g., ignoring the fact that most hires for the job at issue come from the local area or a difference in the type of product in each location.[17] The *Bickerstaff v. Vassar College*, 196 F. 3d 435 (1999) case, however, is interesting for our purposes as it is less clear that the plaintiff's regression was so flawed that it should have been disregarded.

The African-American plaintiff was an Associate Professor, who in 1994 claimed that she was discriminated against because she had not been promoted to Full Professor. Her expert submitted a regression addressing whether salaries at Vassar varied due to race or

[15] The coding of some of the experience variables as yes (1) or no (0), instead of using the amount of relevant experience was noted as a limitation by the trial and appellate courts.

[16] In *Daubert v. Merrill Dow* 509 U.S. 579, 584-587 (1993), the Court established guidelines for the admission of scientific testimony to ensure it was relevant to the issues involved in the case and reliable. Two of the factors were whether the method had been peer reviewed and published and whether it had a known and acceptable error rate.

[17] The *Penk v. Oregon State Board of Education*, 816 F.2d 458, case illustrates this issue. Plaintiffs claimed that women were paid less than similarly qualified men throughout the college and university system. Thus, the two major research universities were included with colleges with a different purpose. Again, the courts rejected a regression analysis submitted by plaintiffs that did not include several major variables. The opinion states "Missing parts of the plaintiffs' interpretation of the board's decision-making equation included such highly determinative quality and productivity factors as teaching quality, community and institutional service, and quality of research and scholarship."

sex. Salaries were regressed on experience, rank, productivity and discipline. The opinion reported that salaries were determined based on scholarship (0-3 points), teaching (0-3 points) and service (0-2 points). The court criticized the analysis because it omitted two of the key variables, teaching and service.[18] Furthermore, the court observed that in light of the point system that the College employed, these variables were quantifiable and could be controlled for in a statistical analysis.

Because the quality of teaching and research should be reflected, in part, in their experience and rank, it is reasonable to ask whether the points awarded each year in the three categories were available. Employers are required to keep personnel records for a period, so they might well have been. If the plaintiff had or could have had access to this data[19], then the court would be justified in concluding the regression omitted too many key factors. On the other hand, if the College did not keep that information, it is unclear that the point system was actually the main determinant of faculty salaries or pay raises.

In contrast to *Bickerstaff*, plaintiff's expert in *Diehl v. Xerox Corp.* 933. F. Supp. 1157 (W.D.N.Y. 1996), that concerned whether older male workers were unfairly laid off, did not include performance ratings of employees. She claimed that the managers would be predisposed to give older employees lower evaluations given the general level of discrimination in society. The court noted that the expert had not conducted a statistical analysis to check whether there was evidence of a pattern of lower performance ratings at the firm and deemed defendant's regression analysis that included these factors more relevant.

It is interesting to contrast the *Diehl* and *General Telephone* cases with the opinion in *Sears Roebuck.* The *Sears* court accepted the defendant's evidence that women, in general, had a lower level of interest in certain types of jobs. The other two decisions said that the party asserting that a general societal pattern would explain a disparity or justify its not being included, as a factor in its analysis should demonstrate that the general pattern applies to the applicants or employees in the particular case.

The issue of whether performance ratings need to be included in statistical analyses submitted in promotion or layoff cases occurs frequently.[20] When possible, statistical analyses submitted by a plaintiff should demonstrate that members of the protected class received lower performance ratings or that, the ratings in the review used to determine the promotions or layoffs were significantly lower than previous ones.[21] Courts also consider noticeably

[18] The opinion 196 F. 3d 435 at 449-450 raised questions about how well the productivity variable controlled for scholarship.

[19] In *Carpenter v. Boeing Co.*, 456 F.3d 1183 (10th Cir. 2006) the plaintiffs did not request information on some of the factors listed in the Collective Bargaining Agreement that specified how opportunities for overtime were to be allocated and the court discounted their regression. The *Bickerstaff* opinion does not discuss the issue of data availability and it is possible that plaintiffs only requested data in electronic form, such as payroll data that included job-title and date of hire.

[20] For example, *Nitshke v. McDonnell Douglas Corp.*, 68 F. 3d 249 (8th Cir. 1995) and *Hutson v. McDonnell Douglas Corp.* 63 F. 3d 771 (8th Cir. 1995). The decision in *Smith v. Virginia Commonwealth University*, 84 F. 3d (4th Cir. 1996) states that whether performance ratings should be included is a question of fact. Thus, the facts to the specific case will determine the appropriateness of using performance ratings.

[21] Gastwirth (1997) illustrates an analysis showing that older workers received worse ratings than younger employees did from a case that settled just prior to trial. Yu (2009) presents an alternative analysis of the data, which yielded a p-value of .07 for a two-tailed test, just above the usual .05 level. In age discrimination cases, however, one-sided tests are appropriate as only employees over the age of 40 are covered by the law. Hence, both procedures would conclude that age affected the performance reviews. The fact that the

more negative reviews that occur after an employee has filed a charge of discrimination as evidence of retaliation.[22]

3.4 Implications for statistical experts

While the Court in *Bazemore* set out useful guidelines on the factors a statistical analysis needs to satisfy in order to be admitted in evidence, it did not discuss the criteria courts should use in evaluating criticisms. An important issue that was not addressed in Bazemore is how to determine the legitimate major factors. In cases involving universities or colleges, teaching, research and service are well known determinants of pay and promotion.[23] Legitimate factors in other cases might be seniority, education and prior experience in hiring cases, along with performance ratings or reviews in promotion and equal pay cases. Information about these factors, however, should be available in order for them to qualify as a major factor. Otherwise, there is no way for courts to oversee that they were applied fairly to all applicants or employees or even used in the decisions that are being scrutinized.

These considerations and the cases discussed indicate that statistical experts request information about all the factors used in making the employment decisions. If one does not utilize information about a particular factor, one should be prepared to justify doing so.[24] For equal employment and other civil cases, such as product liability, it is important that an expert be involved at the time of discovery, so they can ensure that the major factors and information about how they were used in the decision process can be studied.[25]

evaluations of several older employees were lower than the previous ones the received sufficed to defeat the defendant's summary judgment motion in *Woods v. The Boeing Company*, 2009 WL 4609678 (10th Cir. 2009).

[22] See *Wyatt v. City of Boston*, 35 F.3d 13, 15-16, (1st Cir. 1994) for a list of actions indicating retaliation and *Kim v. Nash Finch*, 123 F. 3d 1046 (8th Cir. 1997) for an example where many more negative comments were put in an Asian employee's file after a complaint.

[23] In *Fisher v. Vassar College*, 70 F.3d 1420 (1995), the court noted that service is amorphous. In part, this is because the decision makers, Chairs and Deans, typically appoint faculty members to Committees etc. Thus, information about the willingness of faculty members to participate as a member of a committee, rather than the actual appointments may be more relevant. In addition, the relative importance of community service may vary with the type of institution.

[24] The *Wado v. Xerox Corp.*, 991 F. Supp. 174, 184 (1998), affirmed *Smith v. Xerox*, 196 F. 3d 358 (2d Cir. 1999) discussed by Rosenblum (2015) is an example. In an age discrimination case, the expert did not include performance ratings because they assumed they were biased. This case is similar to *Diehl* discussed in the text.

[25] After a civil case is initiated, there is a period set aside for discovery. The parties ask questions to the other designed to ascertain relevant information. During this period, other employees and the testifying expert may be deposed so both sides are aware of the evidence that will be presented. In product liability cases concerning a new drug, the plaintiff will ask for studies the manufacturer made as well as other complaints from individuals who may have been harmed from using the drug. The defendant, will ask about the health of the plaintiff to see whether the illness may have arisen from another health problem, rather than the drug in question.

4. The Reye syndrome story: How the use of the principles of statistical criticism might have saved lives in both the U.S. and U.K.[26]

For a number of years, pediatric specialists suspected that the risk of children contracting a rare but very serious disease, Reye syndrome, increased after they received aspirin to alleviate symptoms of a cold or similar childhood disease.[27] After the case control study by Halpin et al. (1982), which confirmed a statistically significant association found in two earlier studies, the FDA (1982) initiated the process of warning the public. The proposed warning and background studies was submitted to the Office of Information and Regulatory Analysis at the Office of Management Budget for review.[28]

During the reviewing process, the Aspirin Institute, which represented the interests of the industry, raised several questions about the 1982 study, the FDA had relied on. The Halpin et al. (1982) study matched each case to one or two (when available) controls on their age, race, geographical, time and type of illness.[29] A logistic model controlled for the presence of fever, headache and sore throat yielded an estimated relative risk of 11.5 (p-value ¡ .001). In addition to submitting a detailed reanalysis of the data, which questioned the coding of some of the answers the respondents gave, the Institute argued that the association could be due to "recall" bias and the possibility that parents of cases, who knew about the association, might say they administered aspirin because the child developed Reye syndrome.

The principles of statistical criticism would have required the Institute to submit evidence that a much higher proportion of parents or guardians of cases knew about the association than the parents of the controls. Furthermore, it should have demonstrated a meaningful difference in the ability of parents of cases and controls to recall what they gave their child, even though Reye syndrome tends to occur within a couple of weeks after the prodromal illness.

Two criticisms seemingly had some validity. First, a higher proportion of cases reported a fever than controls and the maximum fever reached by cases generally exceeded that of the controls. The logistic equation had just used the presence (1) or absence (0) of fever. Second, the parents of cases, who were interviewed while their child was very ill, were under much more stress than the parents of controls were and this could have affected their response. The industry suggested the controls formed from children hospitalized for other diseases or visited the emergency room would have been more appropriate as the parents who responded to the questionnaire had been under stress.

Table 3 in Halpin et al. (1982) addressed the issue of the effect of fever by stratifying the data into four levels of the highest fever (none, low, middle, high) and observed that while the prevalence of fever was higher in the cases, that for each level of fever, a higher fraction of cases had taken aspirin than controls. The data in their Table 3 ignored the matching,

[26] The author served as a statistical consultant to the Office of Statistical Policy at OMB and assisted in the review of the epidemiologic studies the FDA relied on.

[27] See Trauner (1984) for a medical perspective and FDA (1982) for a chronology of the early studies.

[28] Section 6(a)(3)(c) of Executive Order12866, and the Regulatory Right-to-Know Act, require that proposed regulations be reviewed by that office. Circular A-4, issued in September 2003, describes the review process, which considers the costs and benefits of a proposed regulation.

[29] Most controls were obtained by locating classmates of a case who were absent from class, presumably with the prodromal illness, at the same time as the case.

i.e., it is reported as several 2x2 tables. Analyzing it with the Cochran-Mantel-Haenszel test yields an estimated odds ratio of 14.7 (p-value $< 10^{-5}$ and 3.5 was the lower end of a 95% confidence interval. Furthermore, at the highest level of maximum fever, all 41 cases used aspirin and 33 of 44 (75%) of the controls did. Thus, the prevalence of high maximum fever in the cases could not have met Cornfield's conditions for an omitted factor.

By the time the decision was made, little information was submitted concerning the potential effect of stress on the accuracy of recall. The government decided that another study should be conducted.[30] A Public Health Task Force was formed and planned the study was during 1983 and a pilot study was undertaken during mid-February through May 1984.[31] The data analysis was reviewed and made available in December 1984. The logistic regression model that controlled for fever and other symptoms yielded an estimated relative risk of 19.0 and the Task Force recommended that the public be warned.[32] When the cases were compared to each of the control groups, the two controls suggested by the industry had the highest estimated odds ratios (28.5 for emergency room controls and 70.2 for inpatient controls). This highlights the importance of the requirement stated by Bross that a critic do more than suggest a possible explanation; they need to submit evidence supporting a conjecture, here a care-giver being under stress, could create the relative risks around 10, found in the earlier studies.

The industry raised doubts about the study and the association to the Regulatory Office at OMB. In particular, it argued that the decision should be based on one study. Based on all the studies, in January 1985, the Office said it would approve the request to warn the public and the industry voluntarily conducted a warning campaign. The CDC reported that the number of cases of Reye syndrome reported to it dropped from 204 in 1984 to 98 in 1985 (see Table 4 in Gastwirth, 2013) for a longer time series and references).

The United Kingdom did not institute a public education campaign until a further study in 1986 confirmed the association and the number of cases declined soon afterwards (Hardie et al. 1996; Porter et al. 1990). Porter et al. (1990) studied the effect of the warning campaign. They report that children with febrile illnesses were 17 times more likely to have taken aspirin before hospitalization in 1985-6 than in 1998-9. They also indicate that about 15 to 25% of the caregivers in Belfast and London were aware of the association between aspirin use in children and Reye syndrome and less than 50% had heard of Reye syndrome.

5. Conclusion and Implications

The events leading up to warning the public about how to avoid a major cause of Reye syndrome illustrate the importance of the statistics profession developing and endorsing sound principles of statistical criticism. The principles suggested by Bross remain a sound guide, however, the context of the particular application should also be considered. In

[30] The review of the regulation needed to be expedited because of an unusual circumstance, totally unrelated to the scientific issues. President Reagan had given speech in Iowa and answered questions from the audience. One person asked about Reye syndrome and the proposed regulation, and he responded that the review would be completed within a month.

[31] Putting aside the few questionnaires, the industry raised doubts about, there still was a statistically increased relative risk, although lower than 10. A non-statistical aspect of the review was whether the increased risk was sufficiently high to justify the proposed wording of the warning.

[32] See Public Health Service (1985) for further details.

situations like the Reye syndrome one, where alternative medications, without such serious side effects, are available, the critic should be required to demonstrate that their criticisms or suggested alternative explanations really explain away the association. There may be other situations in public health where an effective alternative treatment does not exist. Then the proponent of the studies showing an increased risk may also need to demonstrate that the risk of harm out-weighs the benefit of the medicine.

In the context of discrimination cases under the disparate treatment approach, the cost to a plaintiff of a court's total rejection of their statistical evidence is likely to be that they will lose the case at that point. In contrast, if a court admits the plaintiff's evidence and allows the case to go forward, the defendant only has the burden of explaining how the disparity between the success rates of minority and majority employees or applicants arose from legitimate considerations. Thus, at the first stage of the proceedings, the plaintiff, who is the proponent in the framework of Bross, should be required to use the information about the major factors considered by the employer, provided the employer has preserved this information. If the plaintiff succeeds in establishing a statistically significant disparity between similarly qualified (with respect to these major factors) minority and majority employees, the defendant should be required to demonstrate that, the flaws or omitted factors in the plaintiff's evidence are *sufficiently severe* that the ultimate inference could be changed.

In the context of a criminal case, where the prosecution must prove the defendant "beyond a reasonable doubt", the scientific support underlying some types of evidence, e.g. bullet residue and even fingerprint evidence has been questioned. Bolck and Stamouli (2017) discuss some of the statistical issues and refer to many useful articles. The principles of statistical criticism will be helpful to courts in assessing whether a minor violation of an assumption in a statistical calculation or small amount of missing data in study showing a forensic technique has a certain degree of accuracy are severe enough to alter the ultimate impact of the evidence on a jury.

References

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics*, 11, 375-386.

Bolck, A. and Stamouli, A. (2017). Likelihood Ratios for categorical evidence; Comparison of LR models applied to gunshot residue data. *Law, Probability and Risk*, 16, 71-90.

Bross, I. D. J. (1960). Statistical criticism. *Cancer*, 13, 394-400.

Cochran, W.G. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics*, 10, 417-451.

Cornfield, J., Haenszel, W. Hammond, E.C., Lillienfeld, A.M. , Shimkin, M.B. and Wynder, E.L. (1959), Smoking and Lung Cancer: Recent evidence and a discussion of some issues. *Journal of the National Cancer Institute*, 22, 173-203.

FDA (1982). Labeling for Salicylate-Containing Drug Products. Federal Register, 47, Dec. 28, 1982, 57886-57901.

Fienberg, S.E. (Ed.) (1989). *The Evolving Role of Statistical Assessments as Evidence in the Courts.* New York, NY: Springer-Verlag.

Finkelstein, M.O. and Levin, B. (2001). *Statistics for Lawyers*, 2nd Ed. New York, NY: Springer-Verlag

Gastwirth, J.L. (1988). *Statistical Reasoning in Law and Public Policy* Vol. 1 Statistical Concepts and Issues of Fairness. Orlando, FL: Academic Press.

Gastwirth, J.L. (1992). Methods for assessing the sensitivity of statistical comparisons Used in Title VII cases to omitted variables. *Jurimetrics Journal*, 33, 19-33.

Gastwirth, J.L. (2013). Should law and public policy adopt practical causality' as the appropriate criteria for deciding product liability cases and public policy? *Law, Probability and Risk*, 12, 169-18.

Gastwirth, J.L., Miao, W. and Pan, Q. (2017). Statistical issues in Kerner v. Denver: a class action disparate impact case. *Law, Probability and Risk*, 16, 35-54.

Graubard, B.I (2009) Comment on "Using the Peters-Belson method in equal employment opportunity personnel evaluations" by Sinclair and Pan. *Law Probability and Risk*, 8, 119-122.

Gray, M. (1993). Can statistics tell us what we do not want to hear? The case of complex salary structures. *Statistical Science*, 8, 144-179,

Greenhouse, J.B. (2009). Commentary: Cornfield, epidemiology and causality. *International Journal of Epidemiology*, 38, 1199-1201.

Greenhouse, S.W. (1982). Jerome Cornfield's contributions to epidemiology. *Biometrics*, 38, Supplement, 33-46.

Guo Z, Cheng J, Lorch S. A. and Small D. S. (2014). Using an instrumental variable to test for unmeasured confounding. *Statistics in Medicine*, 33, 35283546.

Halpin, T. J., Holtzhauer, F.J., Campbell, R.J. et al. (1982). Reye's syndrome and medication use. *Journal of the American Medical Association*, 248, 687-691.

Hardie, R.M., Newton, .H, Bruce, J.C., Glasgow, J.F.T., Mowat, A.P., Stephenson, J.B.P and Hall, S.M. (1996). The changing clinical pattern of Reye's syndrome 1982-1990, *Archives of Disease in Childhood*, 74, 400-405.

Hersch, J. and Bullock, B.D. 2014). The Use and Misuse of Econometric Evidence in Employment Discrimination Cases. *Washington and Lee Law Review*, 71, 2365-2429.

Mantel, N. (1963). Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel test. *Journal of the American Statistical Association*, 68, 690-700.

Porter, J.D.H., Robinson, P.H., Glasgow, J.F.T., Banks, J.H. and Hall, S.M. (1990). Trends in the incidence of Reye's syndrome and the use of aspirin. *Archives of Disease in Childhood*, 65, 826-829.

Public Health Service (1985). Public Health service study on Reye's syndrome and medications. *New England Journal of Medicine*, 313, 849-857.

Rosenbaum, P.R. (2002). *Observational Studies* (2nd ed.).New York: Springer.

Rosenbaum, P.R. and Krieger, A.M. (1990). Sensitivity analysis for two-sample permutation tests in observational studies. *Journal of the American Statistical Association*, 85, 493-498.

Rosenblum, M. (2015). Strategic evidence issues in equal employment litigation. *Touro Law Review*, 16, 1299-1317.

Sinclair, M.D. and Pan, Q. (2009). Using the Peters-Belson method in equal employment opportunity personnel evaluations. *Law, Probability and Risk*, 8, 95-117.

Trauner, D.A. (1984). Reye's syndrome, *Western Journal of Medicine*, 141, 206-209.

Yu, B. (2009). Modelling an omitted factor in employment discrimination cases. *Law, Probability and Risk*, 8, 153-158.

Zeisel, H. and Kaye, D. H. (1997). *Prove It with Figures: Empirical Methods in Litigation*, Springer-Verlag, New York, USA.