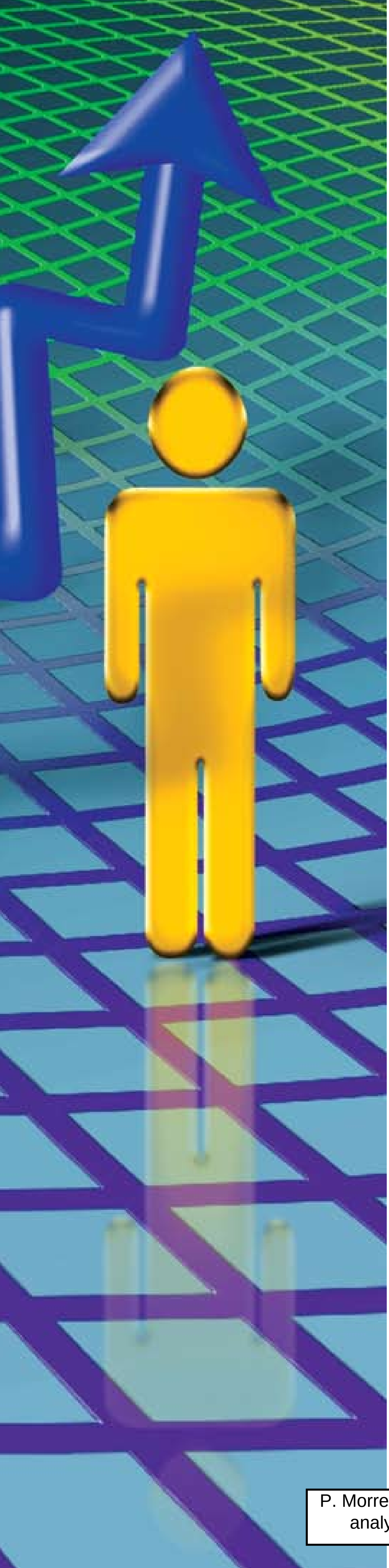


Who, What, and Where:

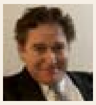
Guidelines for the Statistical Analysis of Disparate Impact in EEO Litigation

By Palmer Morrel-Samuels and Edward Goldman

This might come as a shock: Employees in large corporations sometimes mistakenly believe that they have been discriminated against. Admittedly, discrimination does occur, both in society and in the workplace. And as most attorneys know, many discrimination cases concern claims of either adverse treatment or adverse impact. In both types of litigation, employees believe that they have been discriminated against because of their minority status. In disparate treatment cases, plaintiffs must show that they were treated differently because of that status, and incriminating statements—express or implied—must be admitted as evidence to suggest a discriminatory intent. In contrast, disparate impact cases typically address the discriminatory impact of an ostensibly neutral policy, decision, or program, so plaintiffs rely upon objective data from the entire corporation to prove a discriminatory outcome. In essence, disparate treatment cases often (though not invariably) rely on the admissibility of prior statements or admissions to demonstrate discriminatory intent, whereas disparate impact cases typically rely on statistical analysis of quantitative data to demonstrate a discriminatory outcome that cannot be explained by chance or external societal factors (like gender-based differences in strength, education, etc.) alone. Disparate impact cases, unlike disparate treatment cases, do not require proof of the employer's motive, as *International Brotherhood of Teamsters v. U.S.* (S.Ct. 1977) shows.

This article is designed as a guide for corporate counsel when refuting an erroneous discrimination claim at a large corporation (*i.e.*, having between one thousand and several hundred thousand employees). How do you formulate specific guidelines for using statistics in such litigation so that you can, if and only if it is indeed justified, prove that your corporation acted properly? How do you know that the expert you are hiring has used the right methods for obtaining and analyzing the data? This article can guide you when working with a statistician or psychologist, from either inside or outside your corporation, as you build your case. These guidelines have made it possible to analyze very large datasets from the workplace[1], using quantitative data from both printed[2] and electronic[3] sources.

P. Morrel-Samuels and E. Goldman, "Who, What, and Where: Guidelines for the statistical analysis of disparate impact in EEO litigation," Docket, volume 25, (2) 2007, pp. 54-74.



PALMER MORREL-SAMUELS was formerly a research scientist at the University of Michigan Business School, and is currently the president of Employee Motivation and Performance Assessment in Chelsea, MI. He can be reached at *palmer@umich.edu.*

EDWARD GOLDMAN is associate vice president and deputy general counsel in the Office of the General Counsel, Health System Legal Office at the University of Michigan. He can be reached at *egoldman@med.umich.edu.*

A Few Words About Statistical Evidence

Numerous cases have established standards and precedents for the use of statistical evidence in disparate impact litigation. In fact, as *Watson v. Fort Worth Bank & Trust* (S.Ct. 1988) shows, adverse impact plaintiffs must identify the specific procedure causing the alleged disparity, and "…must offer statistical evidence of a kind and degree sufficient" to show that members of the protected subgroup were negatively impacted by the policy or program in question. That is, plaintiffs must use statistical evidence to prove their case, and must do so with precision, as *New York City Transit Authority v. Beazer* (S.Ct.1979) shows. Moreover, defendants can insist on holding proof of that impact's cause in their hands during litigation, as *Holder v. City of Raleigh* (4th Cir. 1989) shows; impact cannot be assumed just because it seems obvious, or logical, or likely.

Yet some experts overlook an important aspect of *Watson v. Fort Worth Bank & Trust* (S.Ct.1988)*,* which adds to the specificity requirement mentioned above: The plaintiff's burden to establish a *prima facie* case goes beyond the need to show the presence of specific "statistical disparities." Plaintiffs must also show that those observed disparities were not caused by innocuous or unavoidable factors associated with external forces, as *EEOC v. Joe's Stone Crab, Inc.* (11th Cir. 2000) subsequently confirmed. That is, to use statistics appropriately, it is necessary to rule out alternative explanations.

The only way to do an adequate job of "… isolating and identifying the specific employment practices that are allegedly responsible for any observed statistical disparities" (as *Watson* specifies), is to build a comprehensive statistical model of independent variables (also called "predictor variables") and dependent variables (also called "outcome variables"). This model must be compelling enough to withstand scrutiny by academic colleagues, adversarial experts, and decision makers in court. As much research in social psychology shows, any such models must include variables that control for factors such as socioeconomic status, years of education, years of experience, job tenure, skill, and the like. Without the inclusion of these potentially confounding variables (also called "covariates"), none of the analyses will stand up adequately in court.

Selecting an Expert

It is imperative to select a statistical expert who has extensive first-hand experience so that your case need not rely exclusively on findings from published research. Such evidence is vulnerable to a hearsay objection unless it has been read with the benefit of a specialist's expertise, as shown in *United States v. Dukagjini (*2nd Cir.2003). Accordingly, it is wise to select an expert who combines solid knowledge of published research with practical first-hand experience analyzing workplace data.

Social psychologists are in a particularly good position to help jurors, judges, plaintiffs, defendants, and attorneys by using rigorous statistics to identify and measure the causes of adverse impacts in the workplace. Even a brief description of disparate impact litigation introduces notions pertaining to societal norms, subtle unintended consequences, and the need to distinguish between those two statistically. Social psychologists are uniquely equipped to address these issues, in part because their tradition of using statistics to analyze impacts goes back to the late 1890s.

It was the social psychologist Norman Triplett in 1898 who first quantified the impact of gender, age, and an audience's presence on athletic performance[4]. In a set of carefully controlled analyses he isolated the impact of bystanders on an athlete's bicycling speed. The work is germane in this discussion for one simple reason: Triplett used straightforward statistical analyses to disentangle subtle inter-connected factors (such as age, gender, encouragement, anxiety, and mental fatigue) and to measure their impact on an objectively determined outcome.

Similarly, in disparate impact cases, particular importance attaches to statistical methodology and the complex interaction of social and psychological factors. Now the courts are becoming especially receptive to social psychologists' quantitative statistical approach, in part because of their ability to meet the "will assist" clause of the Federal Rules of Evidence (FRE) 702. Just as Triplett provided a helpful analysis of the factors leading to a win or a loss in races more than a century ago, social psychologists today can use advanced statistical tools to disentangle complex causes leading to a promotion or termination in the workplace.

Social psychologists are also in an exceptionally good position to measure impacts in EEO lawsuits because—ever since Sewall Wright's work developing statistical models for experimental research in 1921[5]—their flagship journals like the *Journal of Experimental Psychology,* which first appeared in 1916, and their references like Cohen and Cohen's 1983 text on applied multiple regression[6], have promulgated guidelines ensuring agreement about what it means to run analyses using "sufficient facts or data," using "reliable principles and methods," and "applying those principles and methods reliably to the facts" just as FRE 702 specifies. Such agreement ensures that "statistical validity" is preserved and maximized—just as the Supreme Court required in its 1993 decision *Daubert v. Merrell Dow Pharmaceuticals, Inc.* (S.Ct. 1993).

Moreover, many social psychologists have experience with large datasets containing literally millions of rows and hundreds of variables. They also typically have experience examining the impact of social-psychological variables in the large complex datasets that businesses conventionally collect on production, absenteeism, pay, training, promotions, selection, and the like—just as FRE 803(6) allows. Their approach is especially consistent with the court's requirement that experts use a clear statistical rule of exclusion to minimize the role of random chance while analyzing an observed disparity, a key element in the discrimination case *Hazelwood School District v. United States,* (S.Ct. 1977*)*. In addition, just as FRE 902 (11) or (12) requires, they can also run those analyses properly without tipping their hand—even when they provide opposing counsel with a copy of those certified datasets and written notification that they intend to analyze them. In short, social psychologists are eminently appropriate and qualified to serve as expert witnesses in EEO lawsuits by virtue of what FRE 702 requires in the way of "knowledge, skill, experience, training, or education."

A Baker's Dozen: Thirteen Guidelines for Using Statistics to Prove Impact

When using statistical evidence in EEO litigation, it is important to compile an argument that relies on tests and methods that are as compelling as possible. To service that goal, it is advisable—whenever possible—to adopt the recommendations directly below: (In the thirteen sections that follow, we include some rudimentary statistical information; if you encounter an opposing expert who disregards these guidelines, you will want to learn additional details about the mechanics and limitations of the statistics involved.)

1. In general, it is preferable to avoid a statistical test called the chi-square. The chi-square is a common test of criterion validity, but the test is problematic because in large datasets, the statistic converges to significance. That is, as the number of observations increases, the chi-square gets more and more likely to furnish a significant result, which in this case would suggest, of course, that discrimination is present and that the observed difference cannot be explained by chance alone. (The chi-square test was developed in 1900 to test whether two variables are significantly associated. In simplest form, it counts the frequencies of observations in a 4-cell table, where the columns define one variable, the rows define another, and the chi-square statistic computes the likelihood that data from the two variables are related. A good alternative to the chi-square is called a logistic regression, where the outcome variable is dichotomous, i.e., divided into two categories, and the analysis incorporates a full set of predictor variables that control for potential confounding effects; we'll have more to say about regression techniques and statistical confounds in the parenthetical material below.)

2. It is also sensible to avoid a statistical test called the ANOVA. The ANOVA, a workhorse of statisticians for generations, is less preferable than current methods because its results vary slightly depending on the order in which variables are entered into the test. (The ANOVA was developed in 1925 as a method for carrying out an "…exact analysis of the causes of human variability." It can only handle categorical predictors, i.e., predictors containing verbal labels such as "tall" or "short" or "medium;" accordingly, it cannot accommodate quantitative predictors such as actual height measured in inches. The ANOVA uses Type I sums of squares, also called sequential sums of squares—a value based on the sum of the squared vertical distances from each point in a scatter plot to the regression line that passes through the center of those points. Because the ANOVA relies on sequential sums of squares, its results depend upon the arbitrary order of the variables in the statistical model: Each predictor can only account for parts of the outcome variable that have not been explained by a previously-listed predictor.)

3. We also recommend that statisticians, psychologists and attorneys avoid multiple-measures tests, such as the MANOVA, MANCOVA, repeated-measures tests, and similar statistics, because their complexity makes comprehension by non-statisticians (e.g., counsel, judges, and juries) problematic. (The MANOVA, and the other tests mentioned in this recommendation, are uniquely powerful because they compute the impact of several predictor variables on several outcome variables all at once. The output from these tests typically contains a set of similar but non-identical statistics, each being appropriate for slightly different types of datasets that are, in themselves, differentiated only by virtue of complex preliminary analyses.)

4. Factor analysis is another common statistical tool that we strongly recommend be avoided in litigation. Although factor analysis is frequently encountered in personality research, some believe it is not well-suited for litigation because, unlike virtually every other test in statistics, factor analysis can generate dramatically different results depending on which formulas are used in the two steps of the process. (Specifically, in its first step, factor analysis computes every possible correlation between all the variables in an analysis, and typically uses that information to "crunch" down the variables into the smallest possible set of underlying factors, with each extracted factor including only those variables that are essentially redundant with each other; less typical extraction algorithms create the maximum possible number of factors, and numerous alternatives for extracting factors exist as well. In the second step, these inter-correlations are rotated in an imaginary three-dimensional space so that different factors will seem to gain prominence by becoming aligned with the X, Y, or Z-axis.) In the limited circumstances where a researcher wants to use factor analysis during litigation, it is recommended that the most conservative, replicable, and invariant algorithm be used (one called Principle Components Factor Analysis) and that the results be applied without rotation, so that absolute objectivity can be maintained.

5. Stepwise regression is also, in general, not a recommended procedure; it has the distinction of being singled out by Cohen and Cohen (1983) in their reference text on applied multiple regression as being an analytic method that has no place in the behavioral sciences, presumably because it is subject to intentional or unintentional misuse. (Stepwise regression is an exploratory technique that allows the researcher to compel the inclusion or exclusion of one or more predictor variables in a statistical model, a constraint that can slightly alter the significance of other predictive variables. Although the tool is not as capricious as it may appear at first glance, results can vary substantially if the strict rules for its application are violated. It is typically used with great caution only when it's necessary to formulate a novel statistical model in the absence of any published research.)

6. We also recommend that researchers who run analyses of disparate impact assiduously avoid exclusive reliance on mechanistic statistical models that neglect to include social psychological variables such as years of experience, job level, skill, etc. (In some circumstances, however, such simple models may be useful for convincing distractible jurors where the complexity of the variables can otherwise seem overwhelming.) Consider using hierarchical linear modeling or multiple regression—the latter being the method that Cohen and Cohen (1983) called the most well understood test in all of statistics. (In multiple regression, and its recent elaboration called hierarchical modeling, the researcher selects one outcome variable and as many predictor variables as are needed to accommodate the data, common sense, and the published literature. If the outcome variable is a continuous number like the set that includes 1, 2, and 3 up to some arbitrarily high value, then the regression is considered conventional; if the outcome is a dichotomous variable that includes two categories like alive vs. dead, or retained vs. fired, then the analysis is called a logistic regression.

   In either type of regression, the predictor variables can be continuous numbers, categories, or both. Rules for specifying statistical models in multiple regression are clearly established; moreover, a skilled expert will also know how to use supplementary statistics to assure that a regression model is sufficiently valid and well-formed.) When a full set of social psychological variables is included, it is typical to find that the plaintiff's proposed discrimination examples evaporate because the disputed program or business practice actually ceases to be statistically significant. This approach avoids the need to prove that a practice is "job related" and "consistent with business necessity," as the Civil Rights Act of 1991 specifies; without discussions of necessity in disparate impact litigation, plaintiffs are less able to introduce counterarguments about discriminatory intent or less discriminatory alternatives based on A*lbermarle Paper Co. v. Moody* (S. Ct. 1975), because intent is ultimately distinct from impact, as *International Brotherhood of Teamsters v. U.S.* (S.Ct. 1977) proves.

7. In general, we recommend that weighting formulas be avoided in statistical analyses of impact. When statisticians change the weight of data points—something that is often done in experimental work with good justification—they open themselves to the charge that, had the weightings been different, the results would have been different as well. (Just as the name suggests, this process assigns a weight to every observation in an analysis so that each datapoint has an adjusted impact on the final results. In a typical weighting scheme, observations are assigned weights that quantify their quality, importance, or frequency. This technique is largely a holdover from pre-computer days, when it was easier to multiply an observed value by its weight to generate a new data-

> Even a brief description of **disparate impact litigation** introduces notions pertaining to **societal norms,** subtle unintended consequences, and the need to distinguish between those two **statistically.** Social **psychologists are uniquely equipped** to address these issues.

point, than it was to run computations with a large set of predictor variables.) Although weights may be justified in some cases, the justification will often seem capricious or arbitrary to non-statisticians on a jury. Accordingly, it is sensible to avoid weightings whenever possible.

8. Do not rely on personality tests and/or discussion of personality traits. It is often true that the best approach is to focus on directly observable behavior throughout the process, so that the data reflect directly observable events. This focus on directly observable behavior, rather than personality traits, is well-documented by extensive research dating back to Walter Mischel's classic text of 1968.[7] The reason for this is that personality is not a particularly good predictor of behavior, and if you build your case on personality differences you run the risk of encountering very compelling counter-evidence that focuses on directly observable behavior rather than trait-based constructs that are not explicitly visible.

9. The use of protected t-tests like the Duncan range test or Hsu's test rather than a cascade of independent t-tests, helps prevent your opponent from charging you with fishing for significant results, while ignoring results that contradict your favored hypothesis. (Protected t-tests play an important role in behavioral research, where any one construct is typically measured with several different, nearly identical, variables. By putting all related tests into one group for simultaneous evaluation, protected t-tests control for the fact that large analyses containing hundreds or thousands of comparisons will, just by chance alone, seem to generate a small number of statistically significant results.)

10. In any litigation where statistical evidence is involved, it is important to prove the data's reliability. Reliability—which means "replicability" in the behavioral sciences, not "statistical validity" as it does in legal cases like Daubert (discussed above), can be measured by several common tests. For example, Cronbach's Alpha, the Spearman-Brown Rho, and Kappa are all frequently used in the behavioral sciences. However, an expert in disparate impact cases needs to be cognizant of the fact that some measures of reliability are more defensible than others: Cronbach's Alpha is almost universally accepted, both in and out of court. And, the Spearman-Brown Rho and Kappa are less favored for different reasons: Some researchers avoid the Rho unless it is being used to evaluate averages of correlations under limited circumstances; few experts will use Kappa in legal work because it, unlike other inferential statistics, has no known error rate—a clear requirement of Hazelwood School District (1977), which is discussed above. (Statistical tests of reliability evaluate the extent to which results are consistent in different occasions or circumstances. Chronbach's Alpha, for example, typically measures the consistency of a respondent's answers to similar survey questions. The notion of test reliability is critical in the behavioral sciences because humans are so consistently inconsistent. The analogy to determining one's weight is helpful here: If you have a bathroom scale that provides wildly different readings before and after you brush your teeth, then your scale is not reliable. Validity—also known generally as accuracy—is a subsequent issue: Your scale might be accurate and consistent, or consistent but entirely inaccurate. Because proving validity is not simple in the behavioral sciences, and because high reliability is so rare, proof of reliability takes on particular importance.)

11. In addition, we recommend that statistics in disparate impact cases rely as little as possible on results that do not reach the most stringent criterion for statistical significance. (There are two accepted levels of statistical significance in research. If the results could be explained in five out of one hundred cases just by chance variation alone, then the results are assigned a probability level of .05; if the likelihood of obtaining the observed results by chance variation alone is one in one hundred, then the "p value" drops to .01. Typically, results where p = .05 are described as statistically significant; if p = .01 or less then results are called highly statistically significant. In some fields, such as DNA analysis, smaller p values are reported, but the main criterion is unchanged: Anything smaller than p = .05 is considered statistically significant. Statistical significance is determined by three elements: The magnitude of the deviation from the expected value, the number of

observations, and the scattering of observations around the average. It is important to realize that statistical significance and practical importance are not redundant. In the legal domain, proving statistical significance is a necessary prelude to any discussion about importance.) Although virtually all statisticians and research psychologists use the less stringent requirement of p = .05, most are unaware that Hazelwood (1977) and cases cited in that opinion suggest that it's preferable to require p values of .01 or less. This more stringent criterion avoids the likelihood of counterproductive subsidiary arguments about degrees of statistical significance. Moreover, Footnote 17 of *Hazelwood* (1977) explicitly demonstrates the court's reliance on a criterion for statistical significance of "two or three standard deviations" from random chance, deviations that would typically entail p values of approximately .02 or .002 respectively in a conventional table of cumulative normal probabilities. Accordingly, it is sensible to adopt the more stringent requirement of p = .01 wherever practical.

12. It is recommend that, in analyses where results are not significant, the researcher use a power test to show that sample size was adequate. This is a far higher, and more useful, criterion for sufficiency than that allowed in *Harper v. Trans World Airlines, Inc.* (8^{th} Cir. 1975) where the size of the sample was at issue. (Power tests allow the expert to determine whether a given sample was large enough to definitively prove or disprove a specific explanation. The underlying issue concerns something called effect size: When the size of an effect is small, as it is in much toxicological research for example, it takes a very large sample for a statistically significant difference to emerge; a large effect on the other hand, as is common in mortality from high-caliber gunshot wounds, would be manifest in a much smaller sample. A power analyses can compute a predictor variable's observed effect size and tell the expert how big a sample needs to be for any valid inferences about causal impacts.)

13. We also recommend that experts in statistical litigation build a clear set of graphic displays (*e.g.*, using PowerPoint slides) showing all three pieces of evidence that prove a causal linkage according to Quine's (1941) classic text[8]: Specifically, if a predictor variable X has a genuine impact on an outcome variable Y, then X and Y must be associated at the beginning and the end of the observation period, AND X at the beginning of the observation period must predict Y at the end of the observation period, AND the change in X during the observation period must be statistically linked (e.g., by a statistically significant correlation or regression coefficient) with the change in Y (or at least with Y itself) as measured at the end of the observation period. (Note that these three required relationships cannot stand alone as evidence of causality between two variables, neither in the legal, philosophical, nor research domain: For example, if you only show that a change in predictor X predicts a subsequent change in outcome Y, on cross examination your expert would appear guilty of a *post hoc ergo propter hoc* fallacy. The two other relationships in isolation can look similarly circumstantial to casual observers.) Obviously, these requirements—for association, prediction, and dose-dependence respectively—are best demonstrated in data from regularly conducted business activity (FRE 803, Subsection 6) where the data can document a "flow" over time as discussed in *Malave v. Potter* (2^{nd} Circuit 2003).

As stressed throughout this section, these are merely guidelines for building an optimally compelling argument. Overriding circumstances in any given case may lead an expert to follow a slightly different course. Nevertheless, it is clear that these guidelines are practical and fruitful because they have been used in actual litigation where the charges involved large datasets, considerable complexity, and substantial financial risk.

For example, a large builder of welding robots successfully used these guidelines to defend itself from a breach of contract charge brought by one of its customers, a multinational vehicle frame manufacturer who purchased automated welders for their main assembly line. The frame maker believed that the robots were faulty, and sued the robot maker for $36M. In this case, multiple regression was used to analyze the causes of downtime in four million downtime events over a two-year period, admissible as "recordings of regularly conducted activities" of the business under FRE 803 Subsection 6. The results from the multiple regression, which accounted for nearly 98 percent of the variance in downtime, showed that the robots were not defective, and that downtime was virtually entirely determined by social psychological variables that included absenteeism, skill of the workforce, pay, number of overtime hours, and absence of burnout (operationalized as the proportion of the workforce being used during any given week).

However, in many disparate impact cases, attorneys are unsure about where the statistical analyses should concentrate their focus; after all, there are usually dozens (or even hundreds) of metrics that the corporation could use as evidence. Accordingly, in the next section we outline specific areas where data should be collected and analyzed.

Nine Arenas Where it is Important to Prove Lack of Bias

There are nine major areas for proving fairness, based on federal laws that prohibit job discrimination. These laws include the following: Title VII of the Civil Rights Act of 1964; the Equal Pay Act of 1963; the Age Discrimination and Employment Act of 1967; Title I and Title V of the Americans with Disabilities Act of 1990; Sections 501 and 505 of the Rehabilitation Act of 1973; and the Civil Rights Act of 1991, among others[9]. These laws prohibit discrimination in a number of arenas. We suggest focusing on these nine in a rigorous statistical manner. Ideally, you will be able to show that your company does not discriminate in any of these nine areas. Some courts have more patience than others for this type of analysis; in general, the greater the plaintiff's claim of intent to discriminate, the more leeway the court will allow you to present evidence in these areas.

It goes without saying that this statistical analysis cannot be performed unless the company has the relevant data. In most cases, the corporation's human resources information system contains a wealth of information about each employee's pay, merit bonuses, raises, performance appraisals, disciplinary actions, and job description, as well as unit-specific averages on staff retention, workloads, on-the-job injuries, and perceived working conditions as measured by the annual employee survey. In some instances you may be required to keep certain data but not all data. Maintaining and subsequently retrieving such data can be a lengthy and expensive task—one that may assist plaintiffs and defendants alike. It is understandable that most corporations archive such sensitive information with care, and only after thoughtful consideration.

Where is Everyone?

The first area for testing is *distribution*. Distribution asks, "Where are people distributed within the organization?" If members of a protected class of employees are all assigned to one location and never assigned to another location, then that disparity will show up in statistical tests that look at the distribution of members and nonmembers of that protected class. So, for example, in recent litigation against one of the nation's largest bakers, a minority employee felt that he was assigned to a particular facility because of his race. Support for his claim rested on the fact that his new assignment brought him into the only distribution center with another minority upper-level manager. Although the small population of minority employees in upper-level management positions seemed suggestive, the plaintiff's sole claim concerned his new job location. In this case statistics were used to test the likelihood that the plaintiff could have been assigned to the location in question on the basis of chance alone. Because the bakery maintained some 16 distribution centers with only a very limited number of managers at each facility, chance alone could easily account for the assignment of one or even two minority managers to any given location. This was clearly a case where, despite appearances, the statistics proved that minorities were equally distributed among all the corporation's units. In large corporations, when the proper variables are entered (for example, total number of locations, number of available openings, total number of upper-level managers, etc.) it's very often the case that distribution is not statistically different by protected class membership.

How Are Your Employees Compensated?

The second area to test is *pay*. It's often the case that members of protected classes believe that their pay is lower than the pay of nonmembers and, indeed, respected research shows that women tend to be paid less than men for equivalent work in identical positions. To prove that your company is not discriminating in pay, all you need do is to separate your employees into members and nonmembers of the protected minority of interest. Take their average pay and demonstrate with statistics that there is no statistically significant difference between the pay of the two or several classes involved. Again, the key is to select the appropriate variables for your statistical model.

For example, if a female employee believes that she has been discriminated against in her pay, classifying all employees according to their gender and entering their pay will enable you to test the hypothesis that salary in your company varies by gender. For any such test it is imperative to control for all the important confounding variables—like years of education, position, number of subordinates supervised, years with the company, job responsibilities, skill level, geographic region, etc.—so that you will be able to say that, controlling for all relevant variables, gender has no impact on overall pay. To run the statistical analysis properly, all of these predictor variables (and any others that are appropriate for the conditions at hand) need to be entered into a multiple regression seeking to account for pay. The results can be shown as leverage plots, which control all other predictor variables in the equation, and which will isolate the relationship between gender and pay. The heart of this analysis, of course, is the comparison between men and women in the company *when* all potentially confounding variables (like education, experience, skill, and the like) are controlled for. By following this approach, you can partial out confounding effects that might otherwise mislead casual observers. If any questions arise about whether or not to include a potentially confounding variable, it is best to run the analysis both ways so that you can demonstrate that your result is not contingent on the analysis method. If using the confounding variable *does* alter the results, most courts will allow its inclusion provided the variable does not merely function as a proxy for current discrimination.

Who Moves Up the Corporate Ladder?

The third area of concern is *promotions*. As mentioned above, most human resource information systems keep track of the number of promotions each employee has had since their date of hire. That can be tallied as a continuous number from zero to some high number, and handled in one of two ways: Promotions can be entered as predicted by membership or non-membership in the protected class of the plaintiff; alternatively, promotions can be tallied as a dummy variable, where "1" indicates that the employee was promoted sometime during his or her tenure, and a "0" indicates the absence of any promotions. Again, once you control for all appropriate confounding variables it is very rare indeed to see that number of promotions varies by religion, national origin, gender or age because there are far more important variables that determine promotions including the employee's annual performance appraisals, the amount of profit they generate, their geographic region, and a myriad of other variables.

Who Gets Fired?

The fourth area to analyze is *firing*, which should be treated just the same way that the previous areas were. Again, the human resource information system will be critical for looking at numbers of firings within a specified period—which can be defined as any calendar period, like one, two, or three years, or more. And the test is quite simple; however, as recommended in the guidelines above, it's sensible to avoid the chi-square even though some statisticians may suggest using it for dichotomous variables like this in other settings. As mentioned above, the preferred alternative is logistic regression with an appropriate set of predictor variables, an approach which is far more precise and (unlike the chi-square) will not converge toward significance as the dataset gets larger.

Are Your Training Opportunities Equal?

Accessed training is the fifth area of concern. Employees sometimes feel that they've been discriminated against because they did not have the opportunities for receiving the training that other employees enjoyed. The procedure is straightforward and involves measures such as total number of courses offered and courses taken, total number of class hours, and the like. Again, predictor variables based on membership or non-membership in the protected class will allow you to look at the impact of minority status on access to and use of training. Once the appropriate predictor variables are entered into the equation you will be in a good position to demonstrate that access to training and utiliza-

Do not rely on **personality tests** and/or discussion of personality traits. It is often true that the **best approach** is to focus on directly observable behavior throughout the process, so that the data **reflect directly observable events.**

tion of training are not driven by minority status. It is important, however, to control for other variables that have a known relationship to utilization of training, such as years of education and geographic region. When you control for years of education and you control for region, it is very rare to find any case where access to training is restricted by virtue of membership in a protected class.

Who Occupies the Leadership Suites?

The sixth area concerns *leadership roles*. A sensible question for a plaintiff to ask is whether protected employees are equally likely to occupy leadership roles within the company. Indeed, one of the primary statistics that people usually hold up in cases where there has been discrimination is that only *x*-percent of the top leaders in the company belong to this protected class and the vast majority, *y*-percent, do not. It is easy enough to envision a case like this where there is no discrimination, in part because employees may not have the training or experience that has been a prerequisite for assignment to a leadership position. So it is imperative to control for those variables in the analysis. In this case that means controlling for education, for years in the company, and for years of experience in the field. And, again, it is imperative that the chi-square be avoided, because in a large company, where there can be thousands of people in managerial positions, the results will become adverse even though there may have been no genuine discrimination.

Who Gets Hired?

The seventh area is a common area of concern: *hiring*. As with firing, the HR department's dataset is critical for examining the relationship between minority status and hiring. Some observations will certainly be excluded from this dataset, specifically people who were offered a job but went elsewhere. If it is possible, you may want to get access to that database as well. Large companies often keep track of the number of positions that were offered but not accepted, as well as some demographic information about potential employees who declined employment with the corporation. A number of complex issues are uniquely attached to the issue of hiring, such as disparities in the ratio between hiring rates and population rates in the surrounding area, and the available skill set in the local pool of people looking for work. Nevertheless, most questions about discrimination in hiring can be resolved by examining the ratio between the flow of potential employees applying for openings, and the flow of minority members into the company's workforce. However, here as in all other domains addressed in this paper, controlling for confounding variables is essential.

Are Your Performance Appraisals Fair?

The eighth area of interest is *performance scores*. In the classic 1971 Supreme Court case, *Griggs v. Duke Power Co.,* the aggrieved employee complained that requirements irrelevant to the job were holding him back. Performance appraisals are a critical issue. Large corporations, in general, need to have a methodical, regular, routinized method for evaluating each employee's work on an annual basis. It is also imperative that those performance appraisals be constructed so that no discrimination is built in.

We have seen performance appraisals, for example, that ask entirely inappropriate questions, or questions phrased metaphorically, which tend to evoke different associations in the mind of the appraiser when applied to members of different classes. For example, the performance appraisals at some respected corporations have asked about the leader's "grasp of problems," "ability to take bold strides," and ability "to speak with the voice of authority"—all of which are likely to give unfairly high scores to men because, on average, they have higher grip strength, longer stride length, and lower voices, as much research shows.

Aside from those obvious weaknesses, performance appraisals need to be able to demonstrate that they are objective, valid, unbiased, and relevant. After that, it is still necessary to demonstrate the fact that the performance scores, as they exist in the database of the company, are not significantly different for those who do or do not belong to the protected minority group of the plaintiff. And, again, multiple regression with covariates is critical for doing that. In an interesting follow-up to the landmark litigation at Duke Power, it is instructive to note that the company has now redesigned its performance appraisal and uses it for top-level management. They currently use a 360-degree assessment where employees are evaluated by no fewer than eight of their peers, supervisors, subordinates, and customers—a comprehensive set that is supplemented by the employee's own self assessment. Scores from this appraisal (unlike the one contested by Griggs) predict objective performance scores such as annual growth in the size of the employee's merit bonus, salary, number of subordinate employees, and even high placement in their salary range—an attribute that only one of the employee's raters would have any way of knowing at the time of the assessment.

What Do Employees Say in Your Company Survey?

The last area of interest concerns *survey scores*. Usually HR departments have an annual survey that they administer which, like the performance appraisal, has already been vetted so that its reliability and validity are airtight. Because some litigation refers to working conditions or a hostile working environment, it is imperative that survey scores be examined vis-à-vis membership or non-membership in the protected class.

For example, a large multinational corporation recently used this approach to look for any evidence of discrimination in survey scores from its workforce of 160,000 employees living in 45 countries; neither the ratings from the 60,000 surveys nor their 45,000 written comments showed any evidence whatsoever of discrimination: Scores for all protected groups of respondents (classified by gender, self-reported race, citizenship, and even age) showed no statistically significant differences for any major section of the survey, proving the equivalence of scores for topics such as teamwork, communication, fairness, and even perceived pay.

As a counter example we'll only mention one case: An international vehicle manufacturer recently redesigned its employee survey because they wanted to include new items on respect and communication; a comprehensive statistical analysis showed that communication scores predicted low defect rates on its assembly lines, whereas high scores for respect and ethics predicted a low number of EEOC complaints and similarly low EEOC settlement costs during the next year. In both corporations (for different reasons) statistics played a cardinal role in discrimination complaints, for the first organization because they helped defend a blameless corporation during litigation, and for the second organization because they helped a less-than-perfect company measure, locate, and address an underlying problem that historically led to discrimination lawsuits.

In any such analysis of survey data it is usually incumbent on corporate counsel to show that working conditions are virtually identical for members of the protected class and nonmembers alike. The best way to do this is not to take subsets of questions nor individual questions one at a time, but to look at the overall averages for each major section of the survey as well as the survey's total score. Are the scores of men and women different for communication, or teamwork, or training, or any of the other major topics that most employee surveys include? And if they do vary, *why* do they vary? So, once again, it is critical to build an appropriate mix of predictor variables and variables to control for confounding. Once you control for variables like job type, employment tenure, and related factors, it is very rare to see differences in survey scores that are driven significantly by membership in a protected class.

Experienced Specialists Can Improve the Bottom Line

The court record of statistics in EEO litigation is voluminous. It is well-established that statistics can be used in EEO litigation, and the guidelines for using them are clear to most social psychologists. However, an effective statistical analysis of disparate impact requires an experienced specialist who is familiar with the literature in research psychology, statistics, and law. This multi-disciplinary approach is imperative in EEO litigation so that statistical analyses of disparate impact can be run properly, interpreted sensibly, and applied in a manner that promotes a fair, objective and honest resolution of the litigation under consideration. In the absence of solid experience, familiarity with the rules of evidence, knowledge of discrimination law, and a command of social psychological research, your statistical evidence will be little more than disconnected strings of unconvincing examples. And, as that wonderful old European saying so aptly tells us: "A mere example is not a proof."

Have a comment on this article? Email editorinchief@acc.com.

NOTES

1. Morrel-Samuels, P. (2002). Measuring illegal immigration at US border stations by sampling from a flow of 500 million travelers, *Population and Environment*, 23 (3), 285-302.
2. Morrel-Samuels, P. (2002). Getting the truth into workplace surveys, *Harvard Business Review*, 80 (2), 111-118.
3. Morrel-Samuels, P. (2003). Web Surveys' Hidden Hazards, *Harvard Business Review*, 81(7), 16-17.
4. Triplett, N. (1898). The dynamogenic factors in pacemaking and competition. *American Journal of Psychology,* (7), 507-533.
5. Wright, S., 1921, Correlation and causation. *Journal of Agricultural Research*. 20 (7), 557-585.
6. Cohen J. and Cohen, P., 1983. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates.
7. Mischel, W. 1968. *Personality and Assessment.* New York: John Wiley & Sons, Inc.
8. Quine, W.V.O. (1941) *Elementary Logic.* New York: Harper & Row.
9. Title VII of the Civil Rights Act of 1964 (42 U.S.C.A.2000e); the Equal Pay Act of 1963 (29 U.S.C.A. 206(d)) (EPA); the Age Discrimination and Employment Act of 1967 (ADEA) (29 U.S.C.A. 621); Title I and Title V of the Americans with Disabilities Act of 1990 (ADA) (42 U.S.C.A. 1201); Sections 501 and 505 of the Rehabilitation Act of 1973; the Civil Rights Act of 1991.

The authors would like to thank Diana Palmiere Hunt for her alert and dedicated assistance with this manuscript.