

“STATISTICAL DUELING” WITH UNCONVENTIONAL WEAPONS: WHAT COURTS SHOULD KNOW ABOUT EXPERTS IN EMPLOYMENT DISCRIMINATION CLASS ACTIONS

William T. Bielby^{*}
Pamela Coukos^{**}

ABSTRACT

When statistical evidence is offered in a litigation context, the result can be bad law and bad statistics. In recent high-profile, high-stakes employment discrimination class actions against large multinationals like UPS, Wal-Mart, and Marriott, plaintiffs have claimed that decentralized and highly discretionary management practices result in systematic gender or racial disparities in pay and promotion. At class certification, plaintiffs have relied in part on statistical analyses of the company’s workforce showing companywide inequality. Defendants have responded with statistical presentations of their own, which frequently demonstrate widely varying outcomes for members of protected groups in different geographic areas of the company. These expert submissions usually suggest either that no problems exist, or that any discrimination is isolated and not attributable to institutional-level bias. In adjudicating between these competing visions, courts must referee what the Second Circuit terms “statistical dueling.” As we show in this Article, sometimes at least one of the parties is dueling with unconventional weapons. Using simulated

^{*} Professor of Sociology, University of Pennsylvania; Ph.D., Sociology, University of Wisconsin, Madison (1976).

^{**} Ph.D. Candidate, University of California, Berkeley, Boalt Hall School of Law Jurisprudence and Social Policy Program; J.D., Harvard Law School (1994).

This research benefited from participation in a research working group entitled “Social Science Perspectives on Employment Discrimination in Organizations (SSPEDO),” which is part of the Discrimination Research Group. The Discrimination Research Group is a joint effort funded by the American Bar Foundation, the Center for Advanced Study in the Behavioral Sciences, and the Ford Foundation (grant no. 1045-0189). The research reported in this Article was also conducted with the support of the College of Arts and Sciences of the University of Pennsylvania. The authors wish to thank all the members of SSPEDO and to also acknowledge KT Albiston, Hannah Brueckner, Lauren Edelman, Melissa Hart, Linda Krieger, Robert L. Nelson, Laura Beth Nielsen, Vicki Schultz, and Donald Tomaskovic-Devey for their contributions to this project.

data, we show why courts should become more critical of statistical expertise purporting to test for subunit differences, particularly when offered at the class certification phase of the case. Under some circumstances, the statistical approach often used to oppose class certification in employment discrimination litigation is guaranteed to support the defendant's position, regardless of the actual facts of the case. Furthermore, some courts have improperly or unwittingly legitimized the use of this approach, even when it is demonstrably nonprobative of the issues before the court. Courts need new ways to think about these problems—approaches that better reflect the relevant legal framework and statistical principles.

INTRODUCTION

When statistical evidence is offered in a litigation context, the result can be bad law and bad statistics. Recent high-profile, high-stakes employment discrimination class actions bear this out.¹ A series of similar cases litigated over the past several years involve potentially misleading statistical testimony purporting to show an absence of any pattern of discrimination.² Courts in these cases have not always understood the limitations of the statistical evidence before them or properly weighed its relevance to a ruling on class certification. Because the decision about whether to certify a class is critical to both sides, these errors may generate substantial costs.

In recent cases against large multinationals like UPS,³ Wal-Mart,⁴ and Marriott,⁵ plaintiffs have claimed that decentralized and highly discretionary management practices result in systematic gender or racial disparities in pay and promotion. At class certification, plaintiffs have relied in part on statistical analyses of the company's workforce showing companywide inequality. Defendants have responded with statistical presentations of their own, which frequently demonstrate widely varying outcomes for members of protected

¹ See *infra* Part I.

² See, e.g., *Abram v. UPS of Am., Inc.*, 200 F.R.D. 424 (E.D. Wis. 2001); *infra* text accompanying notes 72–74.

³ *Abram*, 200 F.R.D. at 433 (declining to certify nationwide class of African-American employees).

⁴ *Dukes v. Wal-Mart Stores, Inc.*, 222 F.R.D. 137, 187–88 (N.D. Cal. 2004) (certifying nationwide class of female employees).

⁵ *McReynolds v. Sodexo Marriott Servs., Inc.*, 208 F.R.D. 428, 448 (D.D.C. 2002) (certifying nationwide class of African-American employees).

groups in different geographic areas of the company.⁶ These expert submissions usually suggest either that no problems exist or that any discrimination is isolated and not attributable to institutional-level bias. In adjudicating between these competing visions, courts must referee what the Second Circuit terms “statistical dueling.”⁷

As we show in this Article, sometimes at least one of the parties is dueling with unconventional weapons. For-profit consulting companies and large defense firms are eagerly marketing unorthodox and unreliable statistical methods to employers anxious about how class actions multiply their potential liability.⁸ Of course, plaintiffs can, and sometimes do, submit statistical evidence that is inconsistent with good social science practice and biased in favor of their position. However, a bias in the opposite direction is often lurking within the statistical approach of experts working for defendants, and courts and litigators have largely ignored this bias. Moreover, while the issue as outlined below will be immediately apparent to social statisticians, it has received little attention from either academic statisticians or consulting experts.

Using simulated data, we show why courts should become more critical of statistical expertise purporting to test for subunit differences, particularly when offered at the class certification phase of the case. Under some circumstances, the statistical approach often used to oppose class certification in employment discrimination litigation is guaranteed to support the defendant’s position, regardless of the actual facts of the case. Furthermore, some courts have improperly or unwittingly legitimized the use of this approach, even when it is demonstrably nonprobative of the issues before the court. Courts need new ways to think about these problems—approaches that better reflect the relevant legal framework and statistical principles.

The conflicting expert submissions typical of contemporary Title VII class litigation reflect each side’s distinct litigation strategy, as framed by the legal requirements for class certification. Rule 23 of the Federal Rules of Civil Procedure requires that courts facing class certification motions determine the existence and extent of common factual and legal issues.⁹ The more that individual complaints of discrimination are part and parcel of a challenge to

⁶ *E.g.*, *Caridad v. Metro-North Commuter R.R.*, 191 F.3d 283, 288 (2d Cir. 1999); *Dukes*, 222 F.R.D. at 154.

⁷ *Caridad*, 191 F.3d at 292.

⁸ *See infra* text accompanying notes 112–18.

⁹ *FED. R. CIV. P. 23*; *Gen. Tel. Co. of the Sw. v. Falcon*, 457 U.S 147, 156 (1982).

larger common institutional practices, the more appropriate it would be to certify a class.¹⁰ The more idiosyncratic the claims are, the less reasonable class treatment becomes. Thus, when it comes to statistical evidence, plaintiffs and their experts focus on similarities, while defendants and their experts highlight differences. This debate over similarities and differences focuses on two frequently litigated and interrelated legal issues. The first is the use of aggregated versus disaggregated analyses, and the second is the potential commonality of subjective practices.

Disputes over aggregation played a central role in a number of recent class action decisions involving multi-facility classes.¹¹ Where a case challenges company policies or practices across geographically dispersed worksites, be they divisions, regions, or individual retail stores, defendants typically claim that any statistical evidence must account for potential subunit differences.¹² Mainly because of idiosyncrasies in the way case law has evolved, courts place more weight on “statistical significance” than on the magnitude of disparities between groups.¹³ As a result, plaintiffs’ statistical experts typically assess disparities with organization-wide data, pooled across geographic subunits.¹⁴ In contrast, defendants’ experts take the stance that where the company employs a decentralized management structure, a unique employment system exists within each organizational subunit.¹⁵ Therefore, statistical estimates of disparities must be computed separately by subunit. These different approaches frequently generate quite different results.

The issue of aggregation becomes more important in cases involving highly discretionary management practices. Increasingly, plaintiffs alleging systematic discrimination claim that the company-wide mechanism creating

¹⁰ See generally *infra* Part I (discussing the historical and legal development of class action in the employment discrimination context).

¹¹ See, e.g., *Caridad*, 191 F.3d at 292–93; *Anderson v. Boeing Co.*, 222 F.R.D. 521, 536–37 (N.D. Okla. 2004); *Dukes*, 222 F.R.D. at 156–59; *McReynolds v. Sodexo Marriott Servs., Inc.*, 208 F.R.D. 428, 435 (D.D.C. 2002); *Abram v. UPS of Am., Inc.*, 200 F.R.D. 424, 430–31 (E.D. Wis. 2001). Reviewing reported cases likely understates the prevalence of this defense, as many district court rulings are not published, and this may be particularly true of a procedural ruling such as a class certification motion. Further, the issue may not be discussed in a written ruling even if it arises in litigation, or the case may settle before a contested battle over class certification. The prominence of this issue at seminars, in practitioner papers, and in informal discussion among the plaintiff and defense attorneys who are most frequently counsel in these types of cases suggest that it is an increasingly common defense. See *infra* text accompanying notes 102–09.

¹² See, e.g., *Dukes*, 222 F.R.D. at 156–58.

¹³ See *infra* note 130 and accompanying text.

¹⁴ See *Dukes*, 222 F.R.D. at 156.

¹⁵ *Id.* at 156–57.

bias is discretionary and subjective decision making, which is implemented in the context of a decentralized personnel system. Plaintiffs further charge that this system lacks sufficient monitoring and oversight regarding the process and criteria used for making decisions about pay, promotion, and other conditions of employment.¹⁶ While longstanding legal doctrine clearly permits plaintiffs to challenge these mechanisms as a common practice applicable to a class of employees,¹⁷ defendants frequently use the plaintiffs’ framing of the problem as a basis to argue against class certification. They maintain that if decision making is truly discretionary, then by definition there cannot be a common policy or practice causing the alleged bias. Further, they point to statistical evidence of subunit differences, generated through a disaggregated analysis, as support for that position.¹⁸

Many of the employment discrimination class actions currently being litigated in federal court exemplify what some legal scholars call “second-generation” employment discrimination.¹⁹ While class action claims brought in the early days of Title VII often featured extreme levels of job segregation and anecdotal evidence of overtly racist and sexist management decisions, their progeny tell a more complex story. Today’s major corporate targets typically have at least a token representation of men and women of color and white women at senior levels. They avidly describe their good faith steps to combat discrimination within the corporation and appear to lack significant documented instances of explicit discriminatory animus.²⁰ Many scholars and advocates on both sides of the civil rights enforcement debate openly question whether the existing Title VII legal regime is adequate or appropriate for dealing with these cases.²¹

In this context, one might expect heightened judicial receptivity to arguments that discretionary decision making is not amenable to class

¹⁶ See, e.g., *Cooper v. Southern Co.*, 390 F.3d 695, 703 (11th Cir. 2004); *Anderson*, 222 F.R.D. at 536; *Dukes*, 222 F.R.D. at 145; *McReynolds*, 208 F.R.D. at 441; *Abram*, 200 F.R.D. at 425; *Ingram v. Coca-Cola Co.*, 200 F.R.D. 685, 687 (N.D. Ga. 2001).

¹⁷ *Watson v. Fort Worth Bank & Trust*, 487 U.S. 977, 991–92 (1988); *Gen. Tel. Co. of the Sw. v. Falcon*, 457 U.S. 147, 159 n.15 (1982).

¹⁸ See, e.g., Principal Brief for Wal-Mart Stores, Inc. at 23–28, *Dukes v. Wal-Mart Stores, Inc.*, 474 F.3d 1214 (9th Cir. 2007) (Nos. 04-16688 & 04-16720).

¹⁹ Susan Sturm, *Second Generation Employment Discrimination: A Structural Approach*, 101 COLUM. L. REV. 458, 468–74 (2001); see *infra* Part I.C.

²⁰ E.g. Wal-Mart Facts: Employment & Diversity, <http://www.walmartfacts.com/featuredtopics/?id=3> (last visited Mar. 22, 2007); UPS Community: Diversity, <http://www.community.ups.com/diversity/main.html> (last visited Mar. 22, 2007).

²¹ See *infra* text accompanying notes 86–92; see also *infra* notes 77, 93–96 and accompanying text.

treatment, especially when elaborate statistical presentations accompany that argument. In the context of a move to “second-generation” claims, defendants may be more successfully framing discrimination as individual deviance, rather than as stemming from structural or institutional factors. Thus, while in earlier cases a number of courts viewed “excessive subjectivity” as highly suspicious and easily linked to systemic practices of discrimination,²² more recent cases seem to treat such practices as neutral until proven otherwise.²³ An argument that a particular highly discretionary negative employment decision is either a reasonable management determination or unrelated to any larger system or pattern of bias may be an easier sell in an environment where second-generation style class action claims are more common.

Regardless of the reason, in increasing numbers of multi-facility class actions involving less obvious forms of discrimination, courts are reaching radically different conclusions about the relevance of statistical proof of subunit differences to class certification. One set of decisions views these cases as clearly following from prior Title VII class action precedents and grants little credence to claims that decentralized and discretionary management practices cannot be considered institutional practices for the purposes of class certification.²⁴ Under this approach, courts give less weight to statistical evidence of subunit differences. Another line of decisions views these kinds of cases as well outside traditional class action claims and inconsistent with an efficient group resolution. These decisions generally place great store in a showing of statistical differences from location to

²² See, e.g., *Pettway v. Am. Cast Iron Pipe Co.*, 494 F.2d 211, 231–32 (5th Cir. 1974) (quoting *Rowe v. Gen. Motors Corp.*, 457 F.2d 348, 359 (5th Cir. 1972)); *Rowe*, 457 F.2d at 359 (“[P]romotion/transfer procedures which depend almost entirely upon the subjective evaluation and favorable recommendation of the immediate foreman are a ready mechanism for discrimination against Blacks”); *Brown v. Gaston County Dyeing Mach. Co.*, 457 F.2d 1377, 1383 (4th Cir. 1972) (“[T]he lack of objective guidelines for hiring and promotion and the failure to post notices of job vacancies are badges of discrimination that serve to corroborate, not to rebut, the racial bias pictured by the statistical pattern of the company’s work force.”).

²³ See, e.g., *Denney v. City of Albany*, 247 F.3d 1172, 1186 (11th Cir. 2001) (“[A]n employer’s use of subjective factors in making a hiring or promotion decision does not raise a red flag.”); *Coleman v. Quaker Oats, Co.*, 232 F.3d 1271, 1285 (9th Cir. 2000) (“[S]ubjective evaluations are not unlawful per se and ‘their relevance to proof of a discriminatory intent is weak.’” (quoting *Sengupta v. Morrison-Knudsen Co.*, 804 F.2d 1072, 1075 (9th Cir. 1986))).

²⁴ See, e.g., *Caridad v. Metro-North Commuter R.R.*, 191 F.3d 283, 291–92 (2d Cir. 1999); *Ellis v. Costco Wholesale Corp.*, 2007 WL 127800 (N.D. Cal. Jan. 11, 2007); *Smith v. Nike*, 234 F.R.D. 698 (N.D. Ill. 2006); *Hnot v. Willis Group Holdings, Ltd.*, 228 F.R.D. 476 (S.D.N.Y. 2005); *Dukes v. Wal-Mart Stores, Inc.*, 222 F.R.D. 137, 157–61 (N.D. Cal. 2004), *aff’d*, 474 F.3d 1214 (9th Cir. 2007); *McReynolds v. Sodexho Marriott Servs., Inc.*, 208 F.R.D. 428, 442–43 (D.D.C. 2002).

location.²⁵ Thus, disputes over the application of statistical evidence in these cases are deeply intertwined in the larger theoretical and legal debate about decentralized decision making and subtle bias as sources of common “systemic” injury.

The rapid spread of this strategy and the typical approaches of each side reflect the politics of expertise in employment discrimination litigation. This statistical evidence develops in an institutional context in which a relatively small number of players—large law firms, plaintiffs attorneys, and sophisticated experts—come together repeatedly and enact strategies that become increasingly ritualized over time. Class certification is a critical moment in the life of these cases. If the court certifies the class, a defendant faces vastly increased potential financial liability as compared with an individual case, vulnerability to large-scale injunctive relief, and much higher litigation costs.²⁶ If the plaintiffs lose at class certification, they will be forced to proceed individually. The individual cases may not be financially viable, and denial of class certification may prevent the plaintiffs from adequately addressing any institutional level problems that lead to discrimination. The interests of plaintiffs in bringing forward as much affirmative proof of discrimination as early as possible to bolster their case collides with the interests of defendants in painting a picture of utterly random and unsystematic outcomes, resulting in the statistical dueling that is creating such a challenge for courts to adequately comprehend.

In setting forth this dynamic, this Article begins with the relevant legal and factual framework, presents the statistical simulation, and concludes with some recommendations for courts faced with these dilemmas. Part I sets out the historical evolution of Title VII class action doctrine and some background on the issues that animate much of the current systemic employment discrimination litigation in U.S. federal courts. It also considers the specific question of statistical proof of discrimination in the context of class certification and the current legal debates over that evidence. Part II presents and explains the statistical simulation that documents why some types of statistical results in these cases may be misleading. The Conclusion provides some recommendations for how courts can do better: procedural and

²⁵ See, e.g., *Abram v. UPS of Am., Inc.*, 200 F.R.D. 424, 429–31 (E.D. Wis. 2001); *Grosz v. Boeing Co.*, 2003 U.S. Dist. LEXIS 25341, at *15 (C.D. Cal. Nov. 7, 2003); *Reid v. Lockheed Martin Aeronautics Co.*, 205 F.R.D. 655, 672–73 (N.D. Ga. 2001).

²⁶ 2 NEWBERG ON CLASS ACTIONS § 5:44 (4th ed. 2007).

evidentiary tools that will provide fairness to both sides and that increase the likelihood that courts will reach the correct result when faced with these cases.

As the class action device increasingly comes under fire,²⁷ and corporate interests raise the specter of frivolous legal claims that detract from efficiency and violate norms of fair play,²⁸ this seemingly arcane technical dispute over proper statistical modeling takes on much greater urgency. Not only is the typical defense position in these cases statistically unreliable and in conflict with existing legal doctrine, but it also has substantial implications for Title VII enforcement. If implementing decentralized and highly discretionary management practices provides effective immunity from class action liability, corporations will not hesitate to exploit this regulatory loophole. To the extent that discrimination is less explicit and is related to a particular set of potentially problematic institutional practices, class remedies become more, rather than less, important. Courts should be aware of the potential for a statistically derived end-run around antidiscrimination law and ensure that Title VII class actions remain viable enforcement vehicles for twenty-first century discrimination claims.

I. STATISTICAL EVIDENCE IN SECOND-GENERATION CLASS ACTIONS

Much contemporary class action litigation under Title VII challenges a particular employment model—decentralized and highly discretionary management practices—as leading to unequal outcomes that violate antidiscrimination law.²⁹ Many of these cases directly or indirectly invoke “second generation” theories about the nature of bias in the workplace and represent a shift from single-site classes to classes that span dozens or hundreds of facilities. In this context, disputes over statistical evidence become nested in disputes over the applicability of existing legal doctrine to

²⁷ See John C. Coffee, Jr., *Class Action Accountability: Reconciling Exit, Voice, and Loyalty in Representative Litigation*, 100 COLUM. L. REV. 370 (2000); Richard A. Epstein, *Class Actions: Aggregation, Amplification, and Distortion*, 2003 U. CHI. LEGAL F. 475 (2003); Myriam Gilles, *Opting Out of Liability: The Forthcoming, Near-Total Demise of the Modern Class Action*, 104 MICH. L. REV. 373 (2005).

²⁸ CHRISTOPHER ERATH, DISRAELI WOULD HAVE LOVED EMPLOYMENT DISCRIMINATION CASES (2000); Zachary D. Fasman, *The Use and Misuse of Expert Witnesses in Employment Litigation*, in PRACTICING LAW INST., 34TH ANNUAL INSTITUTE ON EMPLOYMENT LAW 841, 873–76 (Reginald C. Govan et al. eds., 2005); Leslie M. Turner & Paul F. White, *The Numbers Game: Statistics Offered to Show Discrimination May Promise More Than They Prove*, LEGAL TIMES, Apr. 19, 2004, at 1, 1–2.

²⁹ Tristin K. Green, *Targeting Workplace Context: Title VII as a Tool for Institutional Reform*, 72 FORDHAM L. REV. 659, 683 (2003); Melissa Hart, *Subjective Decisionmaking and Unconscious Discrimination*, 56 ALA. L. REV. 741, 778 (2005); see *infra* note 82 and accompanying text.

changing fact patterns. When considering class certification motions in these cases, courts frequently fail to appreciate the limits of statistical evidence in resolving Rule 23 issues.

The doctrinal relationship between statistical evidence of discrimination and proof of “commonality” for purposes of class certification is becoming increasingly incoherent. Courts disagree strongly about whether these cases involve core and relatively straightforward applications of prior case law certifying discrimination class actions or whether they instead test the limits of the legal concepts of what an appropriate class action is.³⁰ That tension extends to their view of the statistical evidence submitted at class certification. Courts that place undue weight on the results of potentially unreliable statistical tests may be improperly circumscribing an important component of the Title VII remedial scheme.

A. *The Historical Doctrinal Approach to Discrimination Class Actions*

The underlying legal standards applicable to these cases frame the statistical issues in a variety of litigation contexts. Class action discrimination cases proceed in stages—a class certification decision, then a trial or summary adjudication “on the merits,” and finally a remedial phase.³¹ If the class is certified, the merits phase is on behalf of the class; if not, only the individually named plaintiffs try their claims.³² If the class wins at trial, then usually remedial proceedings are necessary to allocate relief to individual employees.³³ Statistical evidence may be relevant at any phase of the case, but the problems identified in this Article are most significant at the class certification phase.

Discrimination lawsuits seeking class action status frequently rely on both intentional discrimination (“pattern or practice”) and disparate impact legal theories.³⁴ Although the applicable standards and doctrines are distinct, under either approach, proof of liability requires establishing a discriminatory pattern, usually through statistical proof.³⁵ Under the Supreme Court’s

³⁰ See *infra* Part I.B.

³¹ E.g., *Int’l Bhd. of Teamsters v. United States*, 431 U.S. 324 (1977).

³² See JAMES WILLIAM MOORE ET AL., *MOORE’S MANUAL: FEDERAL PRACTICE & PROCEDURE* §§ 14A.40–43 (2006).

³³ BARBARA LINDEMANN SCHLEI & PAUL GROSSMAN, *EMPLOYMENT DISCRIMINATION LAW* 1244 (2d ed. 1983).

³⁴ 42 U.S.C. § 2000e-6(a) (2000); *Hazelwood Sch. Dist. v. United States*, 433 U.S. 299 (1977).

³⁵ SCHLEI & GROSSMAN, *supra* note 33, at 1331–32.

framework in *International Brotherhood of Teamsters v. United States*,³⁶ plaintiffs may prove pattern and practice claims using a combination of statistical and anecdotal evidence to show discrimination is part of the company's "standard operating procedure."³⁷ Disparate impact theory permits a remedy against facially neutral employment practices that nevertheless operate to discriminate and requires no proof of discriminatory intent.³⁸ Plaintiffs merely must show that a particular employment practice has a "significant adverse effect[]" based on race, gender, or other protected group status.³⁹ These theories govern the merits phase of the action and are not the doctrinal basis for class certification.

Rule 23 of the Federal Rules of Civil Procedure governs certification of a case as a class action. Rule 23(a) sets out the prerequisites to a class action, including that the class is too numerous to make joinder practicable ("numerosity"), the presence of common questions of law or fact ("commonality"), the claims or defenses of the named plaintiff representatives are typical of the claims or defenses of the proposed class ("typicality"), and the representative plaintiffs will fairly and adequately protect the interests of the class ("adequacy").⁴⁰ In addition, any class action must satisfy at least one provision of Rule 23(b).⁴¹

Recent disputes over the proper modeling of statistical evidence at the class certification phase generally arise in the context of determining commonality. Although there is no legal requirement that plaintiffs submit statistical

³⁶ 431 U.S. at 336–38.

³⁷ In *Teamsters*, the Court reaffirmed that class-wide discrimination cases may rely primarily on statistical proof, especially where anecdotal evidence "bolsters" the data and serves to bring "the cold numbers convincingly to life." *Id.* at 339.

³⁸ *Watson v. Fort Worth Bank & Trust*, 487 U.S. 977, 986–87 (1988).

³⁹ *Id.* at 986.

⁴⁰ FED. R. CIV. P. 23(a).

⁴¹ Rule 23(b) sets forth three different circumstances where a class action is appropriate if the factors in 23(a) are satisfied. Rule 23(b)(1) governs situations where there is a risk of inconsistent adjudications or where ruling on an individual case would be dispositive of the interests of other class members not parties to the suit. FED. R. CIV. P. 23(b)(1). Claims against a limited fund are one example of a situation that falls under the (b)(1) class. *See Ortiz v. Fibreboard Corp.*, 527 U.S. 815, 835 (1999). Rule 23(b)(2) applies to cases where the defendant "has acted or refused to act on grounds generally applicable to the class, thereby making appropriate final injunctive or declaratory relief." FED. R. CIV. P. 23(b)(2). The Advisory Committee Notes to the Rule point out that a civil rights class action is "illustrative" of the types of actions that would fit under this provision of the rule. FED. R. CIV. P. 23(b)(2) (advisory committee notes). Rule 23(b)(3) applies to cases where common questions predominate over individual issues, and a class action is the "superior" method for the "fair and efficient adjudication" of the case. FED. R. CIV. P. 23(b)(3). This provision covers a variety of actions, where class action treatment is appropriate because of basic and overarching similarities among the individual claims. FED. R. CIV. P. 23(b)(3) (advisory committee notes).

evidence at the class certification phase, it has become common practice for plaintiffs to include an expert report with a statistical analysis in support of their motion for class certification.⁴² Frequently, plaintiffs seek to present statistical or other evidence of a pattern of decisions pursuant to corporate policy that are adverse to the class in support of their argument for class treatment. Under the Supreme Court’s decision in *Eisen v. Carlisle & Jacquelin*,⁴³ a court may not resolve disputes going to the merits of the claims when ruling on class certification.⁴⁴ This complicates significantly the question of how statistical evidence should influence class certification because these expert reports usually make claims about the existence and extent of discrimination, seemingly a merits issue. Courts have dealt with this problem in conflicting ways: some have limited their review of the statistical evidence⁴⁵ while others ignored the *Eisen* rule and addressed the underlying merits at class certification.⁴⁶

Because these cases also involve decentralized management structures and challenges to discretionary employment practices, how courts treat the statistical evidence also tends to dovetail with the judge’s view on the commonality of subjective decision making. Traditionally, legal doctrine has treated a defendant’s policy or practice that delegates excessively subjective authority to local decision makers as a common practice that may present common factual and legal issues satisfying the commonality requirement.

⁴² *E.g.*, *Dukes v. Wal-Mart Stores, Inc.*, 222 F.R.D. 137, 155–66 (N.D. Cal. 2004), *aff’d*, 474 F.3d 1214 (9th Cir. 2007); *McReynolds v. Sodexo Marriott Servs., Inc.*, 208 F.R.D. 428, 434–36 (D.D.C. 2002).

⁴³ 417 U.S. 156 (1974).

⁴⁴ At the class certification stage, a court is barred from resolving disputed factual issues that go to the merits of the case. *Id.* at 177–78. In practice, however, this boundary is inconsistently enforced. *See, e.g.*, *Cooper v. Southern Co.*, 390 F.3d 695, 716–17 (11th Cir. 2004).

⁴⁵ Some courts view their review as limited to assessing whether plaintiffs have presented competent, admissible statistical evidence or other evidence that can support a reasonable inference of discrimination or both. *See, e.g.*, *Staton v. Boeing Co.*, 327 F.3d 938, 954 (9th Cir. 2003) (anecdotal evidence); *Caridad v. Metro-North Commuter R.R.*, 191 F.3d 283, 292–93 (2d Cir. 1999) (statistical evidence); *McReynolds*, 208 F.R.D. at 441 (combination of statistical and anecdotal evidence). In other words, have the plaintiffs created at least a reasonable question as to the existence of class-wide discrimination? Further litigation may ultimately confirm or disprove the inference of class-wide discrimination, a resolution that is left to the merits phase of the case. *Caridad*, 191 F.3d at 293 (statistical showing at class certification sufficient to establish commonality although “[m]ore detailed statistics might be required to sustain the Plaintiffs’ burden of persuasion”). The Second Circuit has recently limited certain aspects of its ruling in *Caridad*, *see In re IPO*, 471 F.3d 24 (2d Cir. 2007), although an expert dispute over whether discrimination occurred is “not a valid basis for denying class certification.” *Id.* at 35; *see also Hnot v. Willis Group Holdings Ltd.*, 241 F.R.D. 204 (S.D.N.Y. 2007).

⁴⁶ *See, e.g.*, *Cooper*, 390 F.3d at 716–17. These decisions frequently suggest that the presence of subjective decision making creates a need to delve into the merits of the statistical evidence. *See infra* Part I.B.

Although some courts have considered only the question of the existence of a common policy or practice of excessive subjectivity,⁴⁷ others, particularly more recently, have considered statistical evidence as also relevant to class certification, and to commonality in particular, reaching differing conclusions on the implications for certification.⁴⁸

In cases alleging bias due to discretionary and subjective decision making, the allegation usually is that decision makers have a high degree of discretion. Plaintiffs focus on the absence of sufficiently clear, concrete, and relevant criteria for evaluating individuals' skills, abilities, and interest relevant to a specific job, promotion to a higher-paying position, or other career opportunity.⁴⁹ Plaintiffs also typically allege that little exists in the way of organization-wide monitoring and oversight over the process and criteria used to make personnel decisions.⁵⁰ It is this discretionary decision making that is identified as the particular employment practice that results in bias.⁵¹ While a practice of granting wide discretion to lower-level decision makers might seem to undermine the claim that a specific practice has generated bias, the Supreme Court has ruled otherwise in two separate contexts. The first involves how to assess subjective practices at class certification,⁵² and the second involves disparate impact challenges.⁵³

In *General Telephone Co. of the Southwest v. Falcon*,⁵⁴ the Supreme Court recognized that excessive subjectivity can be a common practice amenable to challenge by a class.⁵⁵ *Falcon's* oft-cited footnote 15, which deals with "entirely subjective" practices that operate to discriminate, created an

⁴⁷ *Shipes v. Trinity Indus.*, 987 F.2d 311, 316–17 (5th Cir. 1993); *Cox v. Am. Cast Iron Pipe Co.*, 784 F.2d 1546, 1557 (11th Cir. 1986); *Carpenter v. Stephen F. Austin State Univ.*, 706 F.2d 608, 616 (5th Cir. 1983); see also *Green v. USX Corp.*, 843 F.2d 1511, 1525–26 (3d Cir. 1988), *relevant portion reinstated by Green v. USX Corp.*, 896 F.2d 801, 807 (3d Cir. 1990) (similar analysis under typicality prong of Rule 23(a)); *Ingram v. Coca-Cola Co.*, 200 F.R.D. 685, 697–98 (N.D. Ga. 2001).

⁴⁸ See, e.g., *Caridad*, 191 F.3d at 291–93; *Anderson v. Boeing Co.*, 222 F.R.D. 521, 536–37 (N.D. Okla. 2004); *McReynolds*, 208 F.R.D. at 441, 443–44; *Abram v. UPS of Am., Inc.*, 200 F.R.D. 424, 429–31 (E.D. Wis. 2001); *Grosz v. Boeing Co.*, 2003 U.S. Dist. LEXIS 25341 (C.D. Cal. Nov. 7, 2003); *Reid v. Lockheed Martin Aeronautics Co.*, 205 F.R.D. 655, 672–73 (N.D. Ga. 2001); *Dukes v. Wal-Mart Stores, Inc.*, 222 F.R.D. 137, 157–61 (N.D. Cal. 2004).

⁴⁹ *Cox*, 784 F.2d at 1552 (11th Cir. 1986); see also *Carpenter*, 706 F.2d at 613 (5th Cir. 1983).

⁵⁰ *Cox*, 784 F.2d. at 1551–52.

⁵¹ See William T. Bielby, *Minimizing Workplace Gender and Racial Bias*, 29 CONTEMP. SOC. 120 (2000).

⁵² See *Gen. Tel. Co. of the Sw. v. Falcon*, 457 U.S. 147 (1982).

⁵³ See *Watson v. Fort Worth Bank & Trust*, 487 U.S. 977 (1988).

⁵⁴ 457 U.S. at 147.

⁵⁵ *Id.* at 159 n.15.

exception to the requirement that each particular employment practice challenged in a class action must have a class representative who was impacted by that specific practice.⁵⁶ Under footnote 15, an employee can represent a class of applicants and employees where certain circumstances exist. If, under the facts of the case, “the discrimination manifested itself . . . in the same general fashion” among distinct employment practices, exact congruence between the practices that impacted the class representatives and the practices challenged by the class would not be required.⁵⁷

Falcon observed that an “entirely subjective decisionmaking process[]” could be an example of such a common practice, just as a common written test may be.⁵⁸ Thus, the Court recognized that excessive subjectivity can be a mechanism for discrimination to “manifest[] itself . . . in the same general fashion” from decision to decision and can be a concrete employment practice satisfying the commonality requirement like any objective criteria.⁵⁹ The Court’s subsequent ruling in *Watson v. Fort Worth Bank & Trust*⁶⁰ amplifies this understanding of excessive discretion as an employment practice that can have measurable and consistent effects across a class.

Although not a class action case, the Supreme Court’s *Watson* opinion provides an appropriate analytical framework for understanding the commonality issue. In *Watson*, the Supreme Court held that a plaintiff may challenge subjective employment practices using disparate impact theory.⁶¹ This critical doctrinal shift rejected the view that subjective decision making always constituted unique and autonomous incidents incapable of systematic analysis.⁶² In applying disparate impact theory, the Court recognized that under appropriate circumstances, subjective decision making constituted an identifiable and measurable practice:

We are also persuaded that disparate impact analysis is in principle *no less applicable to subjective employment criteria* than to objective or standardized tests. In either case, a facially neutral practice, adopted without discriminatory intent, may have effects that are

⁵⁶ *Id.*

⁵⁷ *Id.*

⁵⁸ *Id.*

⁵⁹ *Id.*

⁶⁰ 487 U.S. 977 (1988).

⁶¹ *Id.* at 990–91.

⁶² See generally *Lewis v. NLRB*, 750 F.2d 1266, 1271 n.3 (5th Cir. 1985); *Carroll v. Sears, Roebuck & Co.*, 708 F.2d 183, 188 (5th Cir. 1983); *Carpenter v. Stephen F. Austin State Univ.*, 706 F.2d 608, 620 (5th Cir. 1983); *Pouncy v. Prudential Ins. Co.*, 668 F.2d 795, 800 (5th Cir. 1982).

indistinguishable from intentionally discriminatory practices If an employer's *undisciplined system of subjective decisionmaking* has precisely the same effects as a system pervaded by impermissible intentional discrimination, it is difficult to see why Title VII's proscription against discriminatory actions should not apply.⁶³

One of the most critical concerns facing the justices as they ruled on *Watson* was the hotly debated question of whether one could apply any form of empirical analysis to subjective criteria.⁶⁴ By resolving that debate squarely in favor of applying disparate impact analysis, the Court reaffirmed that subjectivity can be a practice amenable to systematic empirical review. Subjective practices are not just an amalgamation of a series of autonomous decisions but instead are concrete employment practices and should be treated the same as any objective employment practice under Title VII.

B. The Current Conflict over Decentralization and Discretion

In recent class certification rulings, particularly at the district court level, this seemingly straightforward doctrine has been directly followed at times and ignored or rejected at others. For some courts, the entire assessment of commonality turns on the presence of allegations of subjective decision making.⁶⁵ For others, statistical evidence submitted at class certification helps “bridge the gap” between that allegation and whether an “aggrieved class exists.”⁶⁶ A third set of cases turns *Watson* on its head. Rather than considering excessive subjectivity on par with other discriminatory employment practices, these rulings require a heightened level of proof in cases challenging discretionary decision making and typically subject the plaintiffs' expert submissions to a heightened level of scrutiny.⁶⁷ In these

⁶³ *Watson*, 487 U.S. at 990–91 (emphasis added).

⁶⁴ This debate is clear from reviewing amicus briefs filed in that case. See, e.g., Brief for American Psychological Association as Amicus Curiae in Support of Petitioner, *Watson v. Fort Worth Bank & Trust*, 1987 WL 881423, at 7 & n.7 (No. 86-6139) (The “widely held view” that subjective practices are “less scientific, less quantifiable, less reliable, and less facially neutral than their objective counterparts” does not mean they cannot be systematically assessed. “It is not intrinsic to such devices to be unquantifiable.”).

⁶⁵ For some courts, those kinds of allegations are the basis for finding commonality. See, e.g., *Ingram v. Coca-Cola Co.*, 200 F.R.D. 685, 697–98 (N.D. Ga. 2001); *Beckman v. CBS, Inc.*, 192 F.R.D. 608 (D. Minn. 2000). For other courts, they are the reason commonality does not exist. See, e.g., *Reap v. Cont'l Cas. Co.*, 199 F.R.D. 536, 544–45 (D.N.J. 2001).

⁶⁶ *McReynolds v. Sodexo Marriott Servs., Inc.*, 208 F.R.D. 428, 441 (D.D.C. 2002); *Dukes v. Wal-Mart Stores, Inc.*, 222 F.R.D. 137, 154–58 (N.D. Cal. 2004), *aff'd*, 474 F.3d 1214 (9th Cir. 2007).

⁶⁷ *Love v. Veneman*, 224 F.R.D. 240, 243–44 (D.D.C. 2004); *Cooper v. Southern Co.*, 390 F.3d 695, 715–16 (11th Cir. 2004).

cases, defense evidence of subunit differences appears to carry substantial weight.

Courts considering commonality in cases involving multi-facility classes and decentralized management structures are also coming to different conclusions, especially with respect to the relevance and weight attached to aggregated versus disaggregated analyses. In many of these cases, plaintiffs, who argue that there is commonality in highly discretionary personnel policy and practice, support that claim with evidence of statistically significant disparities adverse to the class.⁶⁸ The defendant, claiming that personnel policy, practice, and implementation differs across decision-making units precisely *because* of subjective decision-making practices, supports that position with statistics purporting to demonstrate conclusively that disparities are neither consistent nor significant across units. For example, in opposing class certification in *Dukes v. Wal-Mart Stores, Inc.*, the defendant presented its position succinctly:

Managers employ distinct criteria in making their individualized decisions. The criteria differ store-by-store, manager-by-manager. Thus, if a discriminatory decision has been made at a particular store, the decision was made by the individual Store Manager. And that fact alone destroys commonality Wide differences exist in the *nondiscriminatory* pay criteria used. Hourly pay therefore cannot be analyzed in the aggregate. The analysis must be store-by-store.⁶⁹

Correspondingly, the company’s statistical expert submitted the results of thousands of regression analyses that showed no consistent pattern across geographical subunits.⁷⁰

Many courts are relatively untroubled by the presence of decentralized management schemes and tend to defer examination of the statistical evidence to the merits.⁷¹ Other courts have used evidence of subunit differences to rule

⁶⁸ See, e.g., *McReynolds*, 208 F.R.D. at 435.

⁶⁹ Defendant Wal-Mart Stores, Inc.’s Opposition to Plaintiffs’ Motion for Class Certification, *Dukes v. Wal-Mart Stores, Inc.*, 222 F.R.D. 137 (N.D. Cal. June 12, 2003).

⁷⁰ *Dukes*, 222 F.R.D. at 156 (defense performed approximately 7,500 separate regressions).

⁷¹ *Staton v. Boeing, Co.*, 327 F.3d 938, 954–56 (9th Cir. 2003); *Bates v. UPS*, 204 F.R.D. 440, 446 (N.D. Cal. 2001); *Beckman v. CBS, Inc.*, 192 F.R.D. 608, 613–14 (D. Minn. 2000); *Orlowski v. Dominick’s Finer Foods*, 172 F.R.D. 370, 373 (N.D. Ill. 1997); *Thomas v. Christopher*, 169 F.R.D. 224, 237 (D.D.C. 1996), *aff’d in part, rev’d in part on other grounds sub nom.* *Thomas v. Albright*, 139 F.3d 227 (D.C. Cir. 1998); *Shores v. Publix Super Mkts., Inc.*, 1996 WL 407850, at *6–7 (M.D. Fla. Mar. 12, 1996); *Butler v. Home Depot Inc.*, 1996 WL 421436, at **1, 3 (N.D. Cal. Jan. 25, 1996); *Morgan v. UPS of Am. Inc.*, 169 F.R.D. 349, 356 (E.D. Mo. 1996); *Cook v. Billington*, 1992 WL 276936, at *4 (D.D.C. Aug. 14, 1992).

that no commonality exists. One example of the latter approach is *Abram v. UPS of America, Inc.*, where evidence of subunit differences played a substantial role in the court's decision not to certify the class:

The parties agree that there is a statistically significant gap in compensation between African-Americans and whites when the data are considered *in the aggregate*. However, closer examination reveals differences that undermine commonality. For example, even the plaintiffs' expert admits that in a majority of the districts he could find no statistically significant difference in pay between African-American and white supervisors The lack of any *consistent* pattern belies the notion that class members have been affected in common ways by the supposed "practice" of "subjective decisionmaking."⁷²

The Court went on to quote the defense expert's explanation that "[w]hen pay is analyzed district by district . . . there are almost as many districts of UPS in which African-American supervisors make *more* than white supervisors, as there are districts where they make less."⁷³ The opinion draws an explicit connection between the idea of subjective decision making as inherently individualized and the supposed absence of any statistical "pattern of disparities." In this case, the disaggregated analysis carried the day, as it has in other similar cases.⁷⁴

Although it is difficult to articulate the precise relevance of statistical evidence to an assessment of commonality, one can at least stake out some theoretical and doctrinal parameters. While courts diverge on how to treat decentralized and discretionary employment practices, the most persuasive and articulate renderings of this test appear to balance competing considerations. Perhaps the purest reading of *Eisen* would exclude all statistical evidence at the class certification phase, but it seems reasonable to favor certification when plaintiffs have made some threshold showing of common injury. However, consistent with maintaining a distinction between the class certification and merits inquiries, only some preliminary showing should be required. Establishing precise uniformity of actual class-wide injury in every geographic subunit may be more than plaintiffs must do to win the case outright, let alone certify the class. In any event, the purpose of statistical evidence at class

⁷² 200 F.R.D. 424, 431 (E.D. Wis. 2001) (internal citations omitted).

⁷³ *Id.*

⁷⁴ *See, e.g., Grosz v. Boeing Co.*, 2003 U.S. Dist. LEXIS 25341, at *15 (C.D. Cal. Nov. 7, 2003).

certification is not to show commonality or the lack thereof but to show that if certified, the class could at least muster a prima facie case of discrimination.⁷⁵

Commonality turns on the existence of common questions of fact or law, and the statistical evidence merely provides a court comfort that all the effort of proceeding to class-wide trial on the merits will not be completely wasted. The decisions in *Abram* and other cases that have precluded an aggregated analysis at class certification are therefore inconsistent with the theory of commonality under Rule 23. As we show below, they also may be relying on an unreliable depiction of subunit variation.

C. *Second-Generation Classes and Shifting Fact Patterns*

These disputes over the meaning and proper application of statistical evidence in civil rights class actions are connected to a larger dispute over the prevalence and nature of employment discrimination in modern American corporations—a move from “first” to “second” generation claims. Shifts in the kinds of cases litigated as class actions have created an opening for statistical dueling over subunit differences. In the face of an unambiguous pronouncement of the Supreme Court that excessive subjectivity is an identifiable and empirically measurable employment practice, some courts nevertheless display marked skepticism that discretionary management practices could be anything but a series of sui generis individual decisions.⁷⁶ Both the defense bar and scholarly commentary on Title VII class action cases such as *Wal-Mart* challenge the notion that excessive managerial discretion leads to widespread and institutional-level bias and question the very existence of less overt forms of bias in the workplace.⁷⁷ These dynamics have led to an increasingly divergent set of court decisions about class certification in these

⁷⁵ As one court recently observed in declining to strike a plaintiff expert report presenting an aggregated analysis where the defendant claimed this analysis was not relevant to determining commonality: “Plaintiffs’ Motion for Class Certification does not rely on the Madden Report to show common practice but to show common effect.” *Gutierrez v. Johnson & Johnson*, 2006 WL 3246605, at *4 (D.N.J. Nov. 6, 2006) (mem.). The court, in applying this approach to class certification, ultimately declined to certify the class based on issues other than the statistical evidence. *Gutierrez v. Johnson & Johnson*, 467 F. Supp. 2d 403 (D.N.J. 2006).

⁷⁶ See *Abram*, 200 F.R.D. at 429–31; *Cooper v. Southern Co.*, 390 F.3d 695, 715–16 (11th Cir. 2004); *Reap v. Cont’l Cas. Co.*, 199 F.R.D. 536 (D.N.J. 2001).

⁷⁷ ERATH, *supra* note 28; Epstein, *supra* note 27, at 517–18; Ronald M. Green, *Class Actions in Equal Employment Matters*, 657 PLI/Lit 303 (2001); Daniel S. Klein, *Bridging the Falcon Gap: Do Claims of Subjective Decisionmaking in Employment Discrimination Class Actions Satisfy the Rule 23(a) Commonality and Typicality Requirements?*, 25 REV. LITIG. 131 (2006); Amy Wax & Philip E. Tetlock, *We Are All Racists at Heart*, WALL ST. J., Dec. 1, 2005, at A16.

cases and opened the door to the increasing frequency of the kind of statistical submissions analyzed here.

A quarter century ago, in *International Brotherhood of Teamsters v. United States*,⁷⁸ the Supreme Court considered a classic first generation discrimination case. The defendant company had virtually shut out African-American and Hispanic employees from the more desirable and higher-paying positions, relegating them to jobs that paid less and offered little future.⁷⁹ Indeed, the statistical picture of segregation presented in that case was so overwhelming the Court used the phrase “inexorable zero” referring to a situation of almost total exclusion.⁸⁰ A virtually all-white contingent occupied the coveted line-driver positions, and the handful of black line drivers had all been hired *after* the lawsuit.⁸¹

Individual testimony in these first generation cases made clear that this extreme degree of exclusion was no accident but was a direct result of white supervisors making repeated intentionally discriminatory hiring and promotion decisions.⁸² For example, one African-American worker seeking a line-driver post was told by a manager “that there would be ‘a lot of problems on the road . . . with different people, Caucasian, et cetera.’” The manager then stated, “I don’t feel that the company is ready for this right now Give us a little time. It will come around, you know.”⁸³ A personnel officer made an even more blunt comment to a Hispanic applicant, stating “that he had one strike against him—‘You’re a Chicano, and as far as we know, there isn’t a Chicano driver in the system.’”⁸⁴ The government, which brought the case against the defendant company and union on behalf of a class, apparently had little trouble convincing the court that its statistics and its stories met the legal test of discrimination as “standard operating procedure.”⁸⁵

Pattern or practice cases brought today seem to tell a more complicated story than the tale of *Teamsters*. Plaintiffs in these cases typically lack the

⁷⁸ 431 U.S. 324 (1977).

⁷⁹ *Id.*

⁸⁰ *Id.* at 342 n.23. The Supreme Court noted wryly, in rejecting the defense proposition that the government’s statistical analysis was inadequate, that “fine tuning of the statistics could not have obscured the glaring absence of minority line drivers [T]he company’s inability to rebut the inference of discrimination came not from a misuse of statistics but from ‘the inexorable zero.’” *Id.*

⁸¹ *Id.* at 337.

⁸² The Court cited to “over 40 specific instances of discrimination” in the record. *Id.* at 338.

⁸³ *Id.* at 338 n.19.

⁸⁴ *Id.*

⁸⁵ *See id.* at 336.

kind of smoking gun statements or anecdotes frequently found in early Title VII cases.⁸⁶ In place of “the inexorable zero,”⁸⁷ plaintiffs and defendants argue over seemingly more ambiguous or disputable race or gender disparities. Today’s defendants often successfully defend these lawsuits by trumpeting their antidiscrimination policies, diversity awards, and highly placed minority and female employees, arguing that any disparities stem from circumstances outside their control.⁸⁸

Many recent court decisions stress the challenge of identifying and proving discrimination and, in particular, the less overt nature of contemporary forms of bias. The First Circuit, for example, has observed that “discrimination tends more and more to operate in subtle ways.”⁸⁹ Since the early days of Title VII, inferential proof has become more significant “since ‘smoking gun’ evidence is ‘rarely found in today’s sophisticated employment world.’”⁹⁰ Indeed, although the principle that discrimination may be well hidden was established early on, it seems even more likely today that an employer “is unlikely to leave a well-marked trail” pointing to job bias as the reason for an action, or any “notation to that effect in the employee’s personnel file.”⁹¹

The views of a federal judge, who recently denied a group of African-American employees the ability to proceed as a class on claims of systemic discrimination, exemplify this position:

In contrast to the early days of Title VII, it is now more uncommon to find an employer that overtly encourages wholesale discrimination on the basis of race; race discrimination today usually comes in more

⁸⁶ For example, many of these cases explicitly or implicitly rely on a “glass ceiling” theory of discrimination, where hard to pinpoint barriers nonetheless seem to stall the careers of minority and female employees part way up the ladder. *See, e.g.,* Ingram v. Coca-Cola Co., 200 F.R.D. 685, 697–98 (N.D. Ga. 2001); McReynolds v. Sodexho Marriott Servs., Inc., 208 F.R.D. 428, 442–44 (D.D.C. 2002).

⁸⁷ *Teamsters*, 431 U.S. at 342 n.23.

⁸⁸ *Wal-Mart* is a textbook example of this approach, where the company pointed to diversity awards as well as EEO training and policies in its defense. *Dukes v. Wal-Mart Stores, Inc.*, 222 F.R.D. 137, 154 (N.D. Cal. 2004), *aff’d*, 474 F.3d 1214 (9th Cir. 2007).

⁸⁹ *Fernandes v. Costa Bros. Masonry, Inc.*, 199 F.3d 572, 580 (1st Cir. 1999).

⁹⁰ *Thomas v. Eastman Kodak Co.*, 183 F.3d 38, 58 n.12 (1st Cir. 1999) (citing *Hodgens v. Gen. Dynamics Corp.*, 144 F.3d 151, 171 n. 13 (1st Cir. 1998)).

⁹¹ *Carlton v. Mystic Transp., Inc.*, 202 F.3d 129, 135 (2d Cir. 2000). In *McDonnell Douglas Corp. v. Green*, the Supreme Court set forth the burden-shifting method of proof precisely because discrimination is often lurking in the background without being overt. 411 U.S. 792, 802–05 (1973); *see also* *Robin v. Espo Eng’g Corp.*, 200 F.3d 1081 (7th Cir. 2000); *Hasham v. Calif. State Bd. of Equalization*, 200 F.3d 1035 (7th Cir. 2000); *Bickerstaff v. Vassar Coll.*, 196 F.3d 435 (2d Cir. 1999); *Iadimarco v. Runyon*, 190 F.3d 151 (3d Cir. 1999); *Rutherford v. Harris County*, 197 F.3d 173 (5th Cir. 1999); *Smith v. Borough of Wilksburg*, 147 F.3d 272 (3d Cir. 1998); *Scott v. Univ. of Miss.*, 148 F.3d 493 (5th Cir. 1998).

subtle forms. It is perhaps more unusual still to find an employer such as a federal defense contractor—required . . . to create and implement affirmative action programs, and whose employees are represented by a number of different unions—that can manage to engage in discrimination on a class-wide basis in the face of executive branch oversight and collectively-bargained grievance procedures through which issues of discrimination can be brought to light.⁹²

The courts' tightening of legal standards applicable to these cases over time appears to be guided by a shared social narrative that our nation's workplaces have, by and large, put the days of overt bias and deliberate segregation behind them—a narrative that suggests *Teamsters*-style fact patterns unlikely, and by extension, *Teamsters*-style collective action inappropriate.

Simultaneously, an explosion of legal scholarship grounded in empirical analysis, and largely seeking to expand the universe of actionable discrimination claims, contends that existing legal doctrine fails to account for most kinds of discriminatory harm in today's workplace. These legal scholars draw on a large and widely accepted body of social science research on implicit bias.⁹³ This empirical work demonstrates the importance of understanding discrimination not just as overt animus, but also as implicit attitudes that may lead to biased decision making.⁹⁴ These legal scholars share the view that changing social conditions have created a disconnect between the doctrine of the past and the problems of the present but are far less skeptical than some courts that discrimination continues to exist. While a handful of legal scholars have applied this thinking to institutional and organizational

⁹² Reid v. Lockheed Martin Aeronautics Co., 205 F.R.D. 655, 660 (N.D. Ga. 2001).

⁹³ See, e.g., Gary Blasi, *Advocacy Against the Stereotype: Lessons from Cognitive Social Psychology*, 49 UCLA L. REV. 1241 (2002); Linda Hamilton Krieger, *The Content of Our Categories: A Cognitive Bias Approach to Discrimination and Equal Employment Opportunity*, 47 STAN. L. REV. 1161 (1995); Charles R. Lawrence III, *The Id, the Ego and Equal Protection: Reckoning with Unconscious Racism*, 39 STAN. L. REV. 317 (1987); Ann C. McGinley, 'Viva La Evolucion!: Recognizing Unconscious Motive in Title VII', 9 CORNELL J.L. & PUB. POL'Y 415 (2000); David Benjamin Oppenheimer, *Negligent Discrimination*, 141 U. PA. L. REV. 899 (1993); Rebecca Hanner White, *De Minimis Discrimination*, 47 EMORY L.J. 1121 (1998).

⁹⁴ See John F. Dovidio & Samuel L. Gaertner, *Aversive Racism and Selection Decisions: 1989 and 1999*, 11 PSYCHOL. SCI. 315 (2000); Susan T. Fiske, *Controlling Other People: The Impact of Power on Stereotyping*, 48 AM. PSYCHOL. 621 (1993); Samuel L. Gaertner & John F. Dovidio, *The Subtlety of White Racism, Arousal, and Helping Behavior*, 35 J. PERSONALITY & SOC. PSYCHOL. 691 (1977); Anthony G. Greenwald & Mahzarin R. Banaji, *Implicit Social Cognition: Attitudes, Self-Esteem, and Stereotypes*, 102 PSYCHOL. REV. 4 (1995); Barbara F. Reskin, *Including Mechanisms in Our Models of Ascriptive Inequality*, 68 AM. SOC. REV. 1 (2003).

behaviors and dynamics,⁹⁵ most have focused on interpersonal, “one-on-one” forms of discriminatory conduct. Almost all conclude that existing law cannot be harmonized with this knowledge and that new legal theories must be developed or doctrine changed to accommodate them.⁹⁶

Changes in Title VII class litigation trends mirror this move from first to second generation discrimination. Research increasingly recognizes that bias can infect the workplace through both implicit and explicit means, and expert testimony about implicit bias and stereotyping evidence is finding its way into class action cases.⁹⁷ Cases are presenting more complex factual issues and less clear-cut statistical disparities.⁹⁸ A change from government to private enforcement also appears to be occurring. This shifting factual ground is making a defendant’s statistical evidence about the absence of any clear pattern of discrimination more plausible.

Following the passage of Title VII of the Civil Rights Act of 1964 banning employment discrimination based on race, gender, and certain other classifications, class-wide legal challenges to entrenched racial and gender-based hierarchies and job segregation proliferated. These cases focused on problems in policies and practices rather than individual claims.⁹⁹ Federal judges desegregated entire industries and held the power to carry out sweeping social and organizational change through the class action device. While similar kinds of pattern or practice claims continued throughout the 1980s and 1990s, their numbers plummeted from the thousands to the dozens,¹⁰⁰ shifting the focus from institutional to personal accountability. Originally, the government, through the Department of Justice (DOJ) and the Equal Employment Opportunity Commission (EEOC), played a visible role in

⁹⁵ See, e.g., Green, *supra* note 29; Hart, *supra* note 29; Ian F. Haney-Lopez, *Institutional Racism, Judicial Conduct and a New Theory of Racial Discrimination*, 109 YALE L.J. 1717 (2000).

⁹⁶ But see Hart, *supra* note 29; Michael Selmi, *Subtle Discrimination: A Matter of Perspective Rather Than Intent*, 34 COLUM. HUM. RTS. L. REV. 657 (2003).

⁹⁷ See, e.g., *Dukes v. Wal-Mart Stores, Inc.*, 222 F.R.D. 137, 152–53 (N.D. Cal. 2004), *aff’d*, 474 F.3d 1214 (9th Cir.).

⁹⁸ See Green, *supra* note 29; Melissa Hart, *Skepticism and Expertise: The Supreme Court and the EEOC*, 74 FORDHAM L. REV. 1937 (2006); Susan Sturm, *Second Generation Employment Discrimination: A Structural Approach*, 101 COLUM. L. REV. 458 (2001).

⁹⁹ See, e.g., *Int’l Bhd. of Teamsters v. United States*, 431 U.S. 324 (1977); *Rowe v. Gen. Motors Corp.*, 457 F.2d 348 (5th Cir. 1972); *Pettway v. Am. Cast Iron Pipe Co.*, 494 F.2d 211 (5th Cir. 1974).

¹⁰⁰ Current class action filings for employment discrimination cases are currently a fraction of their number from the days of the *Teamsters* ruling, while the federal government currently pursues even fewer than its usual handful of systemic cases. John J. Donohue III & Peter Siegelman, *The Changing Nature of Employment Discrimination Litigation*, 43 STAN. L. REV. 983 (1991).

bringing class action lawsuits, but today the private plaintiffs' bar dominates systemic discrimination litigation.¹⁰¹ Virtually all employment discrimination cases in American courts today proceed largely as competing narratives about deviant individuals—the racist supervisor or the incompetent worker with a frivolous claim—and rarely engage questions of policy or practice.

During the 1990s, a small resurgence in private class actions followed the passage of the Civil Rights Act of 1991,¹⁰² which made damages and jury trials available under Title VII for the first time.¹⁰³ It is in these later cases that disputes over aggregation begin to dominate class certification rulings. The passage of the Civil Rights Act of 1991 raised the stakes for both sides. The private bar had greater incentives to develop and bring class action cases, and defendants became more concerned over their potential financial exposure. Private firms obtained some extraordinarily large class action settlements in the 1990s, with headline-making race discrimination lawsuits against Texaco¹⁰⁴ and Coca-Cola,¹⁰⁵ and dramatic gender and race discrimination claims against a series of retail chains including Shoney's,¹⁰⁶ Lucky Stores,¹⁰⁷ Home Depot,¹⁰⁸ Publix,¹⁰⁹ and many others.

¹⁰¹ Title VII grants the Department of Justice and the EEOC authority to bring pattern or practice cases against public and private employers, respectively. See 42 U.S.C. § 2000e-6 (1964) (Prior to 1972, the Department of Justice also had authority to bring pattern or practice claims against private employers.). During the 1980s and 1990s, the DOJ filed an average of thirteen cases per year, mostly pattern or practice cases. Doug Huron, *No More Enforcers?*, LEGAL TIMES, May 19, 2003. However, under the George W. Bush administration, this relatively low number dropped even further. As of May of 2003, the DOJ was averaging about three new filings a year, only one of which was a pattern or practice case. *Id.* The EEOC today files less than 2% of all discrimination cases filed in federal court. *EEOC v. Waffle House, Inc.*, 534 U.S. 279, 290 n.7 (2002). Some evidence suggests that government enforcement has never been very significant compared with the work of the private bar, even in the early days of Title VII. *Developments in the Law – Employment Discrimination and Title VII of the Civil Rights Act of 1964*, 84 HARV. L. REV. 1195 (1971). The EEOC recently announced an intention to revive its systemic enforcement, acknowledging that its systemic discrimination office was “virtually dormant in recent years, bringing no cases.” Mary Swanton, *Offensive Measure: EEOC Shifts Its Focus from Individual Complaints to Cases of Systemic Discrimination*, INSIDE COUNSEL, Mar. 2007.

¹⁰² Civil Rights Act of 1991, Pub. L. No. 102-166 (codified as amended in scattered sections of 2 U.S.C. 29 U.S.C., and 42 U.S.C.).

¹⁰³ 42 U.S.C. § 1981a (1991).

¹⁰⁴ Thomas S. Mulligan & Chris Kraul, *Texaco Settles Race Bias Suit for \$176 Million*, L.A. TIMES, Nov. 16, 1996, at A1.

¹⁰⁵ Henry Unger, *Judge Approves Coke Bias Settlement*, ATLANTA J.-CONST., May 30, 2001, at C1.

¹⁰⁶ *Haynes v. Shoney's, Inc.*, 1992 WL 752127 (N.D. Fla. June 22, 1992).

¹⁰⁷ *Stender v. Lucky Stores, Inc.*, 803 F. Supp. 259 (N.D. Cal. 1992).

¹⁰⁸ *Butler v. Home Depot, Inc.*, 1996 WL 421436 (N.D. Cal. Jan. 25, 1996).

¹⁰⁹ *Shores v. Publix Super Mkts., Inc.*, 1996 WL 407850 (M.D. Fla. Mar. 12, 1996).

An increase in multi-facility classes is also likely a factor in the increasing disputes over aggregation. The most oft-cited class action cases from the early days of Title VII primarily involved race discrimination in single-site manufacturing facilities¹¹⁰ and often focused on claims of discrimination in hiring.¹¹¹ While these cases frequently involved disputes over statistical evidence, doctrinal issues focused on what variables should be included in the analysis and the proper modeling of applicant pools. Reported cases increasingly deviate from the single-facility model. This change could be the result of shifts in the labor force, national economy, and increasingly globalized corporations. Alternatively or additionally, it could stem from changes in the behavior of plaintiffs and their lawyers in viewing claims against these companies as presenting potential class issues. Either way, these larger, multi-facility companies and retail chains raised the prominence of decentralized management because their corporate structure itself suggested decentralization in employment practices.

Finally, an increasingly organized defense bar and a sophisticated industry of consultants catering to corporate defendants have generated even greater attention to this issue. Expert consulting firms that mainly represent defendants in complex employment litigation actively promote their critiques of plaintiffs' expert approaches and their own strategies for avoiding an unfavorable ruling on class certification.¹¹² In particular, they stress particular approaches to modeling statistical evidence to hone an effective class certification defense.¹¹³ One of the largest consulting firms, National Economic Research Associates (NERA), is particularly aggressive in this regard, promoting in-house papers with titles such as *Common Statistical Fallacies in Pattern-and-Practice Employment Discrimination Cases: A User's Guide for Defense Attorneys*; *Disraeli Would Have Loved Employment Discrimination Cases*; *Elvis Told Me It Was Statistically Significant: Keeping Expert Testimony Based on Junk Statistical Science Out of the Courtroom and Responding When It Gets in*; and, most clearly on point, *Class Dismissed: Using Economic and Statistical Evidence to Defeat Class Certification*.¹¹⁴

¹¹⁰ *Pettway v. Am. Cast Iron Pipe Co.*, 494 F.2d 211 (5th Cir. 1974); *Rowe v. Gen. Motors Corp.*, 457 F.2d 348 (5th Cir. 1972).

¹¹¹ *See, e.g., Int'l Bhd. of Teamsters v. United States*, 431 U.S. 324 (1977).

¹¹² *See, e.g.,* NERA Economic Consulting, *Discrimination/Class Certification*, http://www.nera.com/FocusArea.asp?PA_ID=5&FA_ID=73 (last visited Mar. 22, 2007).

¹¹³ *Id.*

¹¹⁴ *See* ERATH, *supra* note 28.

Among the capabilities in the area of workforce analysis promoted at the NERA website is the following statement of the strategy implied in the paper titles cited above:

OUR CAPABILITIES

Liabilities/Class Certification

Our work in these areas involves performing statistical analyses of employment discrimination claims both in a pattern and practice and in a class certification context In a class context, we have participated extensively at the certification phase, performing statistical analyses to determine whether the commonality and typicality prongs of Rule 23 are met and going beyond the broad-brush types of analyses frequently presented to learn whether any alleged disparity pervades the proposed class, is limited to some subgroup, or does not exist at all.¹¹⁵

This is, essentially, the kind of “heterogeneity” defense—promoting a disaggregated analysis to attack the notion of any pattern or any possible finding of commonality that increasingly appears in second-generation Title VII class certification proceedings.¹¹⁶

NERA was one of the first to pursue this approach successfully, and the results it achieved already appear in at least one reported case. In *Abram v. UPS of America, Inc.*, the trial judge relied heavily on an NERA statistical expert in criticizing the statistics offered by the plaintiffs, explaining that “[t]he lack of any *consistent* pattern belies the notion that class members have been affected in common ways by the supposed ‘practice’ of ‘subjective decisionmaking.’”¹¹⁷ The Court goes on to quote UPS’s expert:

¹¹⁵ NERA Economic Consulting, *Our Practice, At a Glance, Employment & Labor*, http://www.nera.com/image/AAG_EmplLabor_3.2005.pdf; *see also* *At a Glance: Class Certification*, http://www.nera.com/image/AAG_Class%20Certification_9.2004.pdf (touting NERA’s expertise in providing “extensive expert testimony on statistical issues related to Rule 23 and whether any statistical claims are artifacts of the statistical model an opposing expert has chosen”).

¹¹⁶ In addition to the frequent mentions of this approach in defense bar papers, the plaintiffs bar is also organizing to counter disaggregated statistical models and promote the more traditional pooled analyses in the context of class certification. *See, e.g.*, Jocelyn Larkin & Christine E. Webber, *Challenging Subjective Criteria in Employment Class Actions* (Apr. 20, 2005), <http://www.impactfund.org/pdfs/Subjective%20Criteria.pdf>. Indeed, that this issue matters appears to be one place where plaintiff and defense counsel can reach agreement. Jocelyn D. Larkin & Fred W. Alvarez, *Ten Reasons Why Discrimination Class Actions Do or Do Not Get Certified*, <http://www.impactfund.org/pdfs/Ten%20Reasons%20Final.pdf>.

¹¹⁷ *Abram v. UPS of Am., Inc.*, 200 F.R.D. 424, 431 (E.D. Wis. 2001).

In some of the districts, the advantage to African-Americans is statistically significant and in others the disadvantage is statistically significant, while in most of the districts the difference is insignificant—*exactly what would be expected if these differences were unsystematic and random* The plaintiffs’ somewhat tendentious reliance on aggregate data illustrates the perils and misuses of statistical analysis In short, the numbers do not reveal a “pattern or practice” of discrimination (intentional or otherwise) that unites the class.¹¹⁸

In sum, when an allegation is brought regarding systematic discrimination due to discretionary and subjective decision making, defendants are able to respond with a ready-made argument about decentralization and organizational heterogeneity, confirmed by a seemingly sophisticated statistical analysis showing that disparities are neither systematic, pervasive, uniform, nor statistically significant. The fact that at least some courts have accepted this approach, coupled with the active promotion by large consulting firms of the “heterogeneity defense” and accompanying statistical analysis, has guaranteed that this strategy has diffused rapidly. But does the statistical analysis upon which the heterogeneity defense rests really enlighten the court on issues of typicality and commonality? Not necessarily.

II. THE MEANING OF STATISTICAL “PATTERNS”: A DEMONSTRATION OF THE STATISTICAL POWER PROBLEM

The “heterogeneity defense” promoted by consulting firms like NERA exploits doctrinal ambiguities in the relevance of statistics to class certification, seeking to generate favorable results for corporate defendants. This defense essentially argues that the *uniformity* of the subunit disparities is the most relevant consideration for class certification.¹¹⁹ Under this framework, if different subunits of a company exhibit differing levels of racial, gender, or other disparities, then it would be inappropriate to treat employees of all those units as a single class.¹²⁰ While it is intuitively appealing to draw this conclusion, this “uniformity” test does not necessarily track Rule 23’s commonality requirement, as explained further in the Conclusion. Further, as

¹¹⁸ *Id.*

¹¹⁹ *See, e.g., id.* at 432 (agreeing with defendant’s argument that diversity of class membership “weighs heavily against certification”).

¹²⁰ *See, e.g., id.* (pointing out disparities in the proposed class with respect to training, prior experience, educational background, geographic locations, and salaries).

the simulation will show, this approach may exaggerate any differences that truly exist. It can even create the appearance of subunit differences when any racial or gender-based disparities in the separate business units are in fact uniform.

Although statistical power has been the core issue underlying legal disputes over aggregation for decades, these recent approaches place that problem at the forefront. While courts have taken different views on the propriety of separate subunit analyses, plaintiffs typically argue that courts should disregard defense submissions featuring separate subunit regressions because they lack statistical power—that is, they are unlikely to detect discrimination when it really exists.¹²¹ Usually, plaintiffs can only posit the hypothetical implications of defendants' low-power analyses. The simulation below provides a clear example of why statistical power really matters and how an analysis with low statistical power can yield a misleading result.

Using a hypothetical example based on real data, we show how each side's statistical experts would likely approach the typicality and commonality issues central to class certification in a lawsuit alleging discrimination due to discretionary and subjective decision making. The data upon which the example is based were not used in litigating a class certification issue in an actual discrimination case. Using these data, we present regression results using the same kinds of models that plaintiffs and defendants in recent similar class actions tend to utilize. As we will show, in some cases, the regression results are substantially inconsistent with the real patterns in the underlying data.

A. *Smith v. UFS: A Hypothetical Gender Discrimination Class Action*

We will refer to our hypothetical company as Universal Financial Services (“We’re Everywhere”). It has 181 offices nationwide and nearly 8,000 financial service agents. The agent workforce is over 86% male, and a lawsuit has been filed by Maggie Smith, a female UFS agent, and several of her colleagues. The *Smith* case alleges that due to a lack of specificity and oversight, and a high degree of managerial discretion, the company favors men in assigning agents to accounts. Plaintiffs claim that women receive smaller and less profitable accounts to manage, which reduces how much compensation they can earn as agents. The female agents who filed the suit

¹²¹ McReynolds v. Sodexo Marriott Servs., Inc., 349 F. Supp. 2d 1, 15 (D.D.C. 2004).

seek to represent a class of nearly 1,100 female agents employed by the company nationwide. Table 1 describes the size and gender composition of the UFS offices, which are 13% female on average.¹²²

Table 1. Gender Composition of UFS

Universal Financial Services

Number of Offices **181**

Number of Agents Per Office

Median # Agents	41
Mean # Agents	43.9
Range	2 to 122
IQR (25th – 75th percentile)	31 to 54

Office Gender Composition - % Female

Median % Female	13.0%
Mean % Female	13.8%
Range	3.4 to 50.0%
IQR (25th – 75th percentile)	9.5 to 17.8%

We assume that the experts for each side conduct regression analysis to determine if men and women have relevant differences in compensation. In the typical analyses used in employment discrimination cases, regression techniques calculate the mathematical relationship between belonging to a protected class and an outcome, such as the amount of compensation (the “dependent” variable).¹²³ A regression yields what statisticians refer to as an “estimate” of the effect of the variable of interest, such as race, ethnicity, or

¹²² Some offices are as low as 3.4% female and a few small offices are as high as 50% female.

¹²³ For general reference texts on regression analysis, see WILLIAM H. GREENE, *ECONOMETRIC ANALYSIS* (3d ed. 1997); ERIC A. HANUSHEK & JOHN E. JACKSON, *STATISTICAL METHODS FOR SOCIAL SCIENTISTS* (1977); JAMES H. STOCK & MARK W. WATSON, *INTRODUCTION TO ECONOMETRICS* (2003).

gender, on the dependent variable while controlling for other variables that might also influence that outcome.¹²⁴

These potential relationships between outcome, membership in a protected class, and potential controls can be formulated into mathematical equations.¹²⁵ An analysis will yield a regression coefficient for each independent variable. That coefficient is a specific number that represents the best estimate of the impact of that independent variable on the dependent variable. In our example, regression analysis will yield a numerical result that represents the salary difference for women as compared to men.¹²⁶ Regression also generates a “standard error”—how much that estimate might vary across different samples

¹²⁴ See generally HANUSHEK & JACKSON, *supra* note 123, at 25–28, 35–37. For example, statistical analyses frequently control for what economists term “human capital”—factors such as education and experience. To the extent members of a protected class who have the same qualifications as the majority group employees still receive different lower compensation on average, or have different promotion rates, social scientists generally attribute this difference to discrimination. See, e.g., JOYCE P. JACOBSEN, *THE ECONOMICS OF GENDER* 289–317 (2d ed. 1998); Orley Ashenfelter, *Changes in Labor Market Discrimination over Time*, 5 J. HUM. RESOURCES 403 (1970); Francine D. Blau & Marianne A. Ferber, *Discrimination: Empirical Evidence from the United States*, 77 AM. ECON. REV. 316 (1987); William A. Darity, Jr. & Patrick L. Mason, *Evidence on Discrimination in Employment: Codes of Color, Codes of Gender*, 12 J. ECON. PERSP. 63, 67 (1998).

¹²⁵ The examples discussed in this section use an “additive” model, as distinguished from an “interacted” model. WILLIAM D. BERRY, *UNDERSTANDING REGRESSION ASSUMPTIONS* 5 (1993). Interaction terms are sometimes utilized to generate more refined estimates where an expert believes that the relationship between the dependent and independent variables is more complex.

¹²⁶ A sample regression equation might look like this:

$$\text{Salary} = A + B * \text{Education} - C * \text{Female} + D * \text{Experience} + e$$

The letter attached to each variable such as education or experience represents a “regression coefficient.” Looking at each of the coefficients, one can draw certain conclusions about their effect on the dependent variable; for example, which variables are positive (or negative)? How large is the effect? This model proposes to explain salary as a function of an individual’s education, experience, and gender. Besides these measured factors, the regression model includes a residual factor, “e,” that represents the impact of unmeasured factors on salary. Using the available data, a regression would generate actual numbers for each regression coefficient B, C, and D, as well as for A, a constant. It also generates a number for the impact of unmeasured factors, the so-called “unexplained variance” in a regression model.

Suppose that a linear regression yielded the following results:

$$\text{Predicted Salary (\$1000's)} = 10 + 2.1 * \text{Educ} - 3.4 * \text{Female} + 1.7 * \text{Experience}$$

This equation shows the “estimated effect” of each of the variables on the average salary, based on a particular set of data. The regression coefficients in this example represent the average predicted or estimated change in the outcome (in this case salary) when that variable increases by one unit. See, e.g., HANUSHEK & JACKSON, *supra* note 123, at 35–37. Thus, each additional year of education increases salary by \$2,100, everything else being equal, and each additional year of experience for otherwise similarly situated employees is worth an additional \$1,700. This particular sample equation also uses a binary variable (sometimes called “dummy variable”) to capture the effect of belonging to a protected class on salary. MELISSA A. HARDY, *REGRESSION WITH DUMMY VARIABLES* 7–9, 18–21 (1993). That variable is set to equal one for all women in the dataset, and zero for all men. Thus, the estimated effect of being female on an individual’s salary is to decrease it by \$3,400 according to this hypothetical example.

of data. With those two pieces of information, one can calculate the “statistical significance” of the result or how likely it might be to occur by chance.¹²⁷

To understand statistical significance, it is helpful to consider a Groundhog Day analogy¹²⁸—starting over, repeating the process, but getting a slightly different outcome. In other words, if one were to repeat the salary-setting process being studied with the same individuals, one would not get the exact same dollar amounts for each person—a certain amount of random variation would affect the outcome. This is why statisticians refer to the results of an analysis as an “estimate.” If there is no discrimination present, one would expect that, for example, salaries for females would not be either lower or higher, on average, than those of similarly qualified males: There would be zero difference between them. One might also not expect the difference to be exactly zero in every calculation. A certain amount of random variation might mean that any particular analysis generates some positive or negative differential.¹²⁹

Statistical significance allows social scientists to draw certain conclusions about whether the difference recorded in an analysis is likely to be a real difference or simply the result of this kind of random variation. Generally, a test of whether an estimated regression coefficient for an independent variable is “statistically significant” asks whether it falls more than two standard deviations from zero, a result that would be expected just one time in twenty (5% of the time) when the variable truly has no impact on the dependent variable. When the estimated coefficient falls outside that range, it is reasonable to conclude that the true effect differs from zero. Courts, following social science practice, have typically concluded a statistically significant coefficient for a binary variable representing protected class membership reflects discrimination.¹³⁰ As explained below, whether one conducts an aggregated or disaggregated analysis may influence whether a particular difference is statistically significant.

¹²⁷ See generally Daniel L. Rubinfeld, *Reference Guide on Multiple Regression*, in REFERENCE MANUAL ON SCIENTIFIC EVIDENCE 179 (2d ed. 2000).

¹²⁸ GROUNDHOG DAY (Columbia Pictures Corp. 1993). In the film, Bill Murray played a weather reporter forced to repeat a single day over and over until he got it “right.” Certain things happened exactly the same way each day, but there was also variation each time, an analogy that captures in certain respects the theory of estimation described here.

¹²⁹ The “margin of error” reported for an opinion poll reflects a similar concept.

¹³⁰ See, e.g., *Hazelwood Sch. Dist. v. United States*, 433 U.S. 299, 311 n.17 (1977). See generally RAMONA L. PAETZOLD & STEVEN L. WILLBORN, *THE STATISTICS OF DISCRIMINATION: USING STATISTICAL EVIDENCE IN DISCRIMINATION CASES* § 4.11 (2002).

In a typical case, the experts would likely disagree on the controls to be used in the equation, but for purposes of this analysis, we can ignore that complication. In our example, each side agrees that the amount of time spent in the industry affects an agent's productivity and, therefore, should be included as a control in the regression analysis. They also agree that any other factors, like education, that may influence how much an agent earns are unrelated to either gender or length of service in the industry. Therefore, no other control variables need to be included in the analysis.¹³¹

In arguing for class certification, plaintiffs seek to show the Court that there have been large, consistent, and statistically significant disparities in compensation favoring men. The plaintiffs' expert, Dr. P, offers a regression analysis testing the effect of gender on earnings while taking only years of service into account. That analysis would generate a single regression coefficient that represents the average effect of gender on compensation throughout UFS. These results suggest that there is a large and statistically significant gender gap in compensation of about 13%, which is consistent from year to year. The first line of Table 2 shows the results for each of five years.¹³²

The defense expert, Dr. D, mercilessly attacks plaintiffs' statistical expert, suggesting that this analysis was performed by someone who is either incompetent or unscrupulous—or possibly both. Dr. D raises two criticisms. The first is the “fallacy of composition” critique. If women who work at UFS live disproportionately in markets that generate little business, failing to take that into account in a statistical analysis would bias the results. The consequences of working in a low-compensation office might be the real reason for the difference, and not gender. Or both location and gender might matter, but failing to include location could exaggerate the gender differences. The second is the “decentralization” critique. Dr. D asserts that decisions about pay are delegated to management in each of the local offices, and as a result, the analysis must be done on an office-by-office basis.

¹³¹ BERRY, *supra* note 125, at 10–11 (appropriate to exclude variables if they have a weak impact, if data is lacking, or if there is no strong theoretical basis to include them).

¹³² Plaintiffs' expert would use the following regression specification (where “Female” is a binary variable and “Experience” is years of service in the industry):

$$(1) \text{Log Earnings} = a + b1*Female + b2*Experience + b3*Experience**2 + e$$

Here b1 is the effect of gender on earnings, in this case controlling for length of service. As is common in analyses of compensation, this expert would use the logarithm of earnings and would include years of service squared.

Dr. P responds to the criticisms by adding controls for office location.¹³³ The results, reported in the second line of Table 2, show that the gender disparities are indeed smaller after controlling for location,¹³⁴ although in this example they remain large, statistically significant,¹³⁵ and consistent from year to year.¹³⁶ In our hypothetical example, plaintiffs are hopeful that they have met the burden of supplying a statistical analysis sufficient to justify class certification.

¹³³ The new specification, with a set of binary variables for office location, is as follows:

$$(2) \text{ Log Earnings} = a + b1*\text{Female} + b2*\text{Experience} + b3*\text{Experience}**2 + \text{Office Location} + e.$$

¹³⁴ For each year, the regression coefficient for Female with controls for office (line 2) is smaller than the corresponding coefficient without those controls (line 1).

¹³⁵ Lines 6 and 7 show the level of statistical significance. In general t ratios greater than 2 represent p values of .05 or less—the level of significance courts typically rely upon.

¹³⁶ When the dependent variable is measured on a logarithmic scale, the coefficients for binary variables correspond approximately, but not exactly, to percentage disparities. Row 3 of Table 2 translates the regression coefficients in row 2 into exact estimates of percentage disparities. So, for example, the coefficient for Female for year 1998 of -.116 translates into a gender gap in earnings of 11.0% among men and women comparable with respect to years of experience and office location.

Table 2. Regression Results of Plaintiffs' Expert

	1994	1995	1996	1997	1998
Regression Results					
Gender coef., controlling experience	-0.137	-0.141	-0.144	-0.133	-0.123
Gender coef., controlling experience, office	-0.134	-0.139	-0.142	-0.130	-0.116
Percentage gender disparity, controlling experience, office	-12.5%	-13.0%	-13.2%	-12.2%	-11.0%
Number of offices	7300	7510	7707	7927	7943
Observations	170	180	181	187	181
Gender coef. t ratio, controlling experience	-6.65	-6.83	-7.28	-6.81	-5.98
Gender coef. t ratio controlling experience, office	-6.57	-6.88	-7.30	-6.79	-5.79
R-Square controlling experience, office	0.3136	0.3411	0.3454	0.3720	0.3827

Dr. D is not persuaded. She concedes that including control variables for offices eliminates the possibility of the fallacy of composition, but she insists it fails to address the decentralization critique. She offers an alternative analysis that seems to indicate that certification is not appropriate. According to Dr. D, the 11% to 13% differences in compensation represent an improper averaging across the 181 different offices in each year. Her alternative analysis begins with the same regression equation as Dr. P's original equation, estimating the effect of gender on compensation while controlling for length of service. However, she estimates the effect of gender *separately* for each of the 181 UFS offices that employ both male and female agents. She reports 181 gender coefficients for each year. In some offices the disparity favors men, in others it favors women, and in most instances, it is not statistically significant.

As defense experts have done in other cases, Dr. D presents her results graphically. The results for 1998 appear in Figures 1 and 2. The graphical presentation emphasizes the subunit differences, dramatically demonstrating in Figure 1 how the various separate estimates of the effect of gender on earnings vary widely from office to office. Figure 2 arrays the same 181 coefficients from lowest (most negative) to highest (most positive), again illustrating that in most instances, the disparities are not significant and in many offices they actually favor women—precisely the finding that the judge found so persuasive in *Abram v. UPS of America, Inc.*¹³⁷

¹³⁷ See 200 F.R.D. 424, 431 (E.D. Wis. 2001).

Figure 1. Estimated Gender Coefficient on Vertical Axis; Squares Denote Significant Effects.

Effects of Gender Vary Widely Across Offices, and with Just a Few Exceptions Are Not Significant

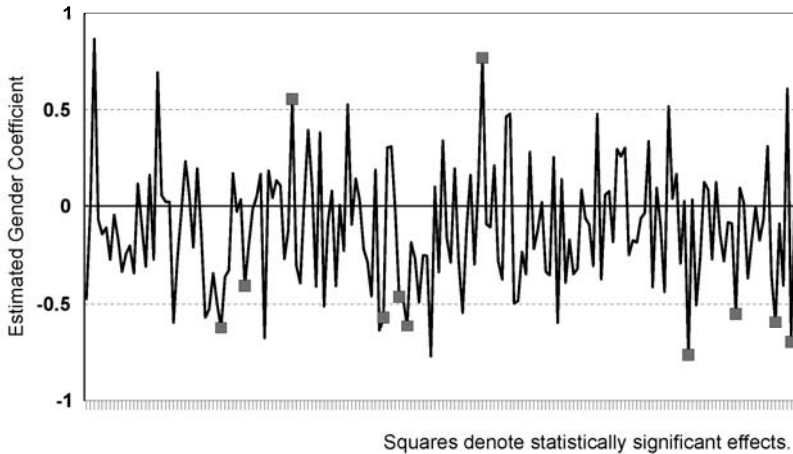
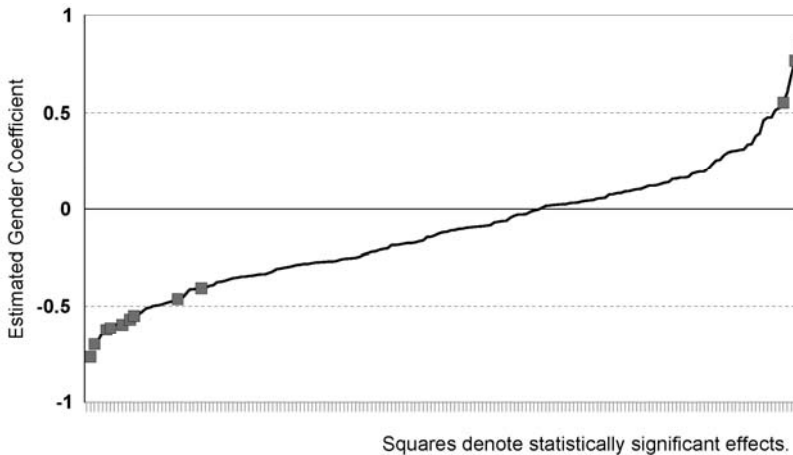


Figure 2. Estimated Gender Coefficient on Vertical Axis; Squares Denote Significant Effects (same results as Figure 1, but with estimated coefficients displayed from most negative to most positive).

Effects of Gender Vary Widely Across Offices, and with Just a Few Exceptions Are Not Significant

(offices displayed from most negative to least negative gender coefficient)



Armed with these findings, the defendant presents to the judge a bottom line that seems irrefutable: Plaintiffs achieved their “statistically significant” disparity, purportedly showing a “pattern and practice” resulting in an 11% gender gap in earnings, through *improper aggregation*. In fact, in the vast majority of UFS offices there is no significant earnings disparity between men and women, and there is no consistency in the size of those disparities from office to office. Thus, there is no “commonality” here and no basis for certifying a class.

In response, plaintiffs argue that by estimating separate regression coefficients for each of 181 offices, Dr. D has dramatically lowered the power of her tests for statistical significance. When studying discrimination, “power” is the probability of detecting discrimination when it really exists. In other words, a statistical test that lacks power may cause you to wrongly conclude that the effect of race, ethnicity, or gender on salary is zero. Ramona Paetzold and Steven Willborn explain: “Power is particularly important for correctly interpreting nonsignificant results,” and “if there is low power to detect important disparities, then *nonsignificant results could be due to chance and may not imply there is no discrimination.*”¹³⁸

In a regression model, several factors affect the power of a statistical test of the hypothesis that the effect of an independent variable is zero. First, power increases with the size of the effect to be detected.¹³⁹ In the context of testing for the impact of gender on earnings, this simply means that large disparities (e.g., a gender gap of 20%) are more likely to show up as “statistically significant” than small disparities (e.g., a gender gap of 5%). Second, power increases with overall “explained variance” of the regression model.¹⁴⁰ This simply means that when unmeasured factors have little impact on earnings, the “Groundhog Day” effect is smaller than when unmeasured factors have a big impact. Third, and most importantly, power increases with sample size. A gender gap of a specific magnitude (say 15%) is more likely to show up as statistically significant in a regression analysis based on 1,000 observations than in one based on 100 observations.¹⁴¹ Aggregation affects the number of observations and, therefore, the power of the test.

¹³⁸ PAETZOLD & WILLBORN, *supra* note 130, § 4.15 (emphasis added).

¹³⁹ William T. Bielby & James R. Kluegel, *Statistical Inference and Statistical Power in Applications of the General Linear Model*, 8 SOC. METHODOLOGY 283, 300 (1977).

¹⁴⁰ *Id.*

¹⁴¹ HERMAN AGUINIS, REGRESSION ANALYSIS FOR CATEGORICAL MODERATORS 69–82 (2004); Bielby & Kluegel, *supra* note 139, at 300; *see, e.g.*, HANUSHEK & JACKSON, *supra* note 123, at 54 (Increasing the

Some courts have recognized the effect that low statistical power and a disaggregated analysis of a smaller number of observations can have on the ability to detect discrimination. As Judge Posner explains, “It is an unacceptable statistical procedure to turn a large sample into a small one by arbitrarily excluding observations.”¹⁴² Some other Title VII cases have addressed statistical power and aggregation but not in the context presented here of testing for subunit differences.¹⁴³

From the plaintiffs’ perspective, the fact that few of the results in our hypothetical example are statistically significant is not surprising in light of the much smaller sample sizes in Dr. D’s approach. The pooled data represent a more powerful test, which is more likely to detect discrimination when it occurs. Plaintiffs contend that by controlling for office location, Dr. P addressed any potential relevance of the subunits. However, plaintiffs’ analysis as presented here does not provide any information on the extent to which gender disparities differ among subunits, an issue considered in more detail below.

The judge now faces the challenge of adjudicating between these dueling statistical claims within the framework of legal doctrine. Prior case law largely favors plaintiffs, but the defense argument is intuitively appealing. How can there be a class suffering a common injury if the office to office results are all over the map? Though rarely considered explicitly in employment discrimination litigation, these assertions are greatly influenced by statistical power and can actually be tested empirically.

number of observations “can improve the precision of our estimates (reduce their variance).”). Two other factors affect the power of a statistical test of a regression coefficient. First, power decreases to the extent that the factor being tested is highly correlated with other control variables in the regression model. For example, if nearly all men have extensive industry experience and nearly all women have little industry experience, then it is more difficult to precisely disentangle the relative contributions of gender and experience to earnings than when the ratio of men to women is similar at all levels of experience. Second, adding too many control variables relative to the number of observations can also decrease the power of the statistical test and make the results less precise. Mathematically, one extraneous control variable added to the model has the same effect as dropping one observation. *Id.* at 93–94.

¹⁴² *Washington v. Elec. Joint Apprenticeship & Training Comm.*, 845 F.2d 710, 713 (7th Cir. 1988) (rejecting plaintiffs’ statistical analysis of a single year in favor of defense analysis pooled across years).

¹⁴³ *Eldredge v. Carpenters* 46 N. Cal. Counties Joint Apprenticeship & Training Comm., 833 F.2d 1334, 1339 n.7 (9th Cir. 1987) (“[A]ggregated data presents a more complete and reliable picture”); *Cook v. Boorstin*, 763 F.2d 1462, 1468–69 (D.C. Cir. 1985) (rejecting defendant’s argument to restrict statistical analysis to particular job categories); *Ezell v. Mobile Housing Bd.*, 709 F.2d 1376, 1382 (11th Cir. 1983); *Capaci v. Katz & Besthoff, Inc.*, 711 F.2d 647, 654 (5th Cir. 1983) (rejecting defendant’s claim statistical analysis should be fragmented by city and year as “an unfair and obvious attempt to disaggregate the data to the point where it was difficult to demonstrate statistical significance”).

B. *Who’s Got the Power: A Simulation Exercise*

Suppose that in a parallel universe there was another UFS (*Sim1*)—one where we *know for certain* that the exact same gender disparity exists across offices. We assume that some discriminatory mechanism operates the same way in every office. We can assume the disparity corresponds to the same 11% gap Dr. P calculated for the most recent year in her analysis. We can now generate *simulated* earnings for each individual employed at UFS using the same data as before but manipulating it to reflect a consistent 11% gender gap.¹⁴⁴

Now, with the simulated data, we replicate Dr. P’s analysis—the analysis that assumes a uniform gender disparity. We also replicate Dr. D’s analysis, estimating a separate regression equation for each office—even though we have generated simulated earnings through a process that makes the gender disparity constant from office to office. The results from replicating Dr. P’s analysis appear in Table 3. Not surprisingly, the estimates from the simulated data (a coefficient of -.108, which is equivalent to a 10.2% gap) conform closely to the simulated gender difference built into the data (a coefficient of -.116, which is equivalent to a 10.9% gap).

¹⁴⁴ The following equation is used to generate simulated earnings:

$$(3) \text{ Sim Log Earnings} = a - .116 * \text{Female} + b2 * \text{Experience} + b3 * \text{Experience} ** 2 + \text{Office Location} + eDel$$

where a, b2, b3, and the coefficients for the office binary variables are also fixed (at values taken from the 1998 regression controlling for experience and office location) and e is generated from a random normal distribution, independently across observations. In other words, for two employees of the same sex in the same office with identical length of service, their *simulated* earnings will be identical, but for the random component introduced by the disturbance term. If one is a man and the other is a woman with the same experience, their simulated earnings will be identical but for the random component *and* a 11% penalty subtracted from the woman’s earnings.

Table 3: Comparing Results Based on Simulated Earnings and Actual Earnings

	Actual	Simulation with b set to -.116
Gender coef., controlling experience, office	-0.116	-0.108
Percentage gender disparity, controlling experience, office	-10.9%	-10.2%
Observations	7941	7941
Number of offices	181	181
Gender coef. t ratio controlling experience, office	-5.76	-5.33
Number of offices with no significant disparity	170	174
Number of offices with significant disparity favoring women	2	2
Number of offices with significant disparity favoring men	9	5

Replicating Dr. D's analysis—a search for differences in the gender gap in earnings—generates much more interesting results. Those results, in Figures 3 and 4, are strikingly similar to the original results. Even though the simulated data have a uniform disparity built in, the defendant's expert produces a result that supports her client's litigation position. According to this analysis, the disparities favor men in some offices, women in other offices, and in most instances are not significantly different from zero. The statistical picture is at odds with the reality of the data.

Figure 3. Replicating the Heterogeneity Analysis of the Defendant's Expert with Simulated Data Containing a Uniform Disparity.

***Estimated Effects of Gender Still Vary Widely Across Offices,
and with Just a Few Exceptions Are Not Significant***

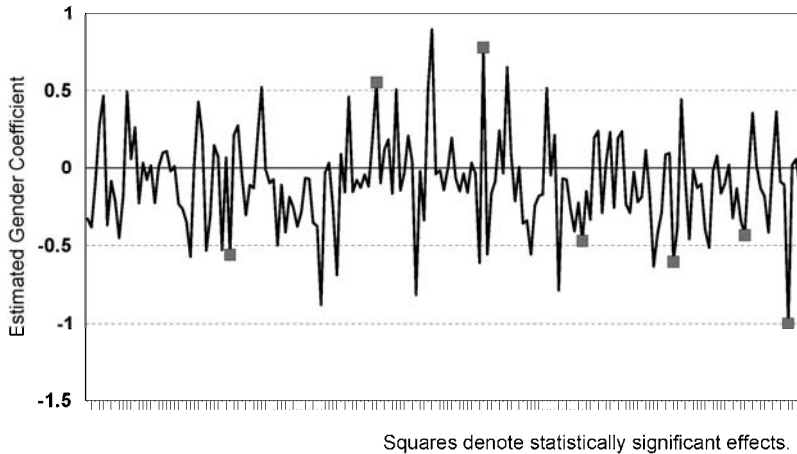
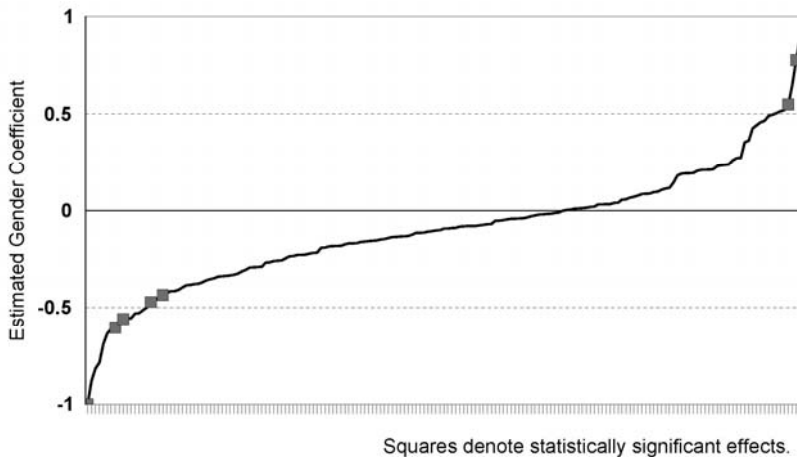


Figure 4. Replicating the Heterogeneity Analysis of the Defendant's Expert with Simulated Data Containing a Uniform Disparity (same results as Figure 3 but with estimated coefficients displayed from most negative to most positive).

***Estimated Effects of Gender Still Vary Widely Across Offices,
and with Just a Few Exceptions Are Not Significant***
(offices displayed from most negative to least negative gender coefficient)



Armed with these results, the plaintiffs present their argument against disaggregation: Defendant's expert achieved the finding of no consistent "statistically significant" disparity across offices by capitalizing on chance in a way that guaranteed a finding of no systematic discrimination, even if there was uniform discrimination across organizational units, an example of *improper disaggregation*. In fact, while *within* the vast majority of hypothetical UFS offices, there are no statistically significant earnings disparities between men and women, this neither proves nor disproves that there is a consistent, uniform disparity by gender.

The results of statistical analysis—seemingly random disparity—fail to line up with the empirical reality—uniform disparity—because of lowered statistical power. By breaking up the dataset into much smaller units, the amount of "noise," or random variation, is elevated over any pattern that exists. In this case, any particular individual man or woman might have especially high or low earnings because of factors not measured in the analysis, factors that vary randomly among employees irrespective of their gender.¹⁴⁵ Because individual salaries will tend to vary for a host of reasons, looking at smaller groups will tend to diminish evidence that women as a group are harmed more than men.

Further, the presence of this kind of random variation does not mean that the average is deceptive. A woman may be the highest earner in her particular office but may be paid far lower than a host of equally or lesser-qualified men in other offices. Isolating the comparison to only those in a particular office could unfairly exclude otherwise relevant evidence of gender-based compensation disparities. Conversely, a woman who is penalized because of her gender to the same extent as any other class member could still end up being more highly compensated than the vast majority of men or even be the highest-paid person in the entire company due to the impact of unmeasured factors. Absent discrimination, her already high salary would have been even higher.¹⁴⁶

Imagine performing 1000 coin flips. Given such a large sample, the result is likely to be very close to 500 heads and 500 tails. If the result were 600

¹⁴⁵ Indeed, because in this hypothetical the employees work in the financial services industry, and their compensation is determined in part by volatile market forces, the degree of "noise" in the dataset may be especially high.

¹⁴⁶ Again, this is not an unlikely occurrence in an industry in which earnings are highly volatile due to market factors that have nothing to do with gender or with other measurable determinants of earnings.

heads and 400 tails, one would suspect a loaded coin. Yet each group of ten tosses might be all over the map. Some would be 6 heads and 4 tails, but others might include more tails than heads. Looking at the smaller groups might not alert the coin tosser to the underlying problem. Statistical methods that focus solely on subgroup uniformity and do not consider the size or significance of average effect are similarly blind to disparities that would be immediately apparent in an organization-wide analysis.¹⁴⁷

But what happens when the world works the way the defendant maintains? That is, what if there really is considerable heterogeneity in gender disparities across offices, and what if it were truly due to distinctive features of the personnel system in each office?¹⁴⁸ Again, a simulation exercise proves illuminating.¹⁴⁹

In this iteration, we generate simulated earnings for a hypothetical UFS (*Sim2*) where we *know for certain* how disparities vary across offices.¹⁵⁰ This time, however, we generate two sets of simulated earnings reflecting different patterns of heterogeneity. In *Sim2*, the gender coefficient is set to vary uniformly across offices in the range of $-.1158$, plus or minus $.05$. We refer to this as “continuous heterogeneity”: Although the disparity is not identical across offices, women face a disadvantage in every office, but a bit more in some offices than in others (a gender gap in earnings ranging from 6% to 15%). In *Sim3*, the gender disparity is fixed at 0 for two thirds of the offices (gender parity in earnings) and at $-.3474$ (a gender gap of 29%) at the remaining third of offices. We refer to this as “radical heterogeneity,” a case

¹⁴⁷ A series of psychology experiments reveals that whether data is presented in an aggregated or disaggregated format has a substantial impact on the ability of subjects to perceive discrimination, a finding that is particularly relevant when the legal and policy environment disfavors looking at institutional patterns of bias. Linda Hamilton Krieger, *Civil Rights Perestroika: Intergroup Relations After Affirmative Action*, 86 CAL. L. REV. 1251, 1305–09 (1998).

¹⁴⁸ In establishing causation for purposes of Title VII liability between a particular set of personnel policies and practices and documented disparities, courts do not assume that relationship and would also look to evidence well beyond these kinds of statistical results to draw a conclusion.

¹⁴⁹ At this point, it is important to keep in mind that the simulation results are shaped not just by the magnitude of the gender disparities, but also by the overall size of the workforce, the size in each office, and the amount of variation due to unmeasured factors, a point we return to below.

¹⁵⁰ This time, the model used to generate simulated earnings is

$$(4) \text{ Sim Log Earnings} = a + b1*\text{Female} + b2*\text{Experience} + b3*\text{Experience}**2 + \text{Office Location} + e$$

where a , $b2$, $b3$, and the coefficients for the office binary variables are also fixed (at values taken from the 1998 regression controlling for experience and office location) and e is generated from a random normal distribution, independently across observations.

that raises a plausible question as to whether a common policy or practice is operating across the company.

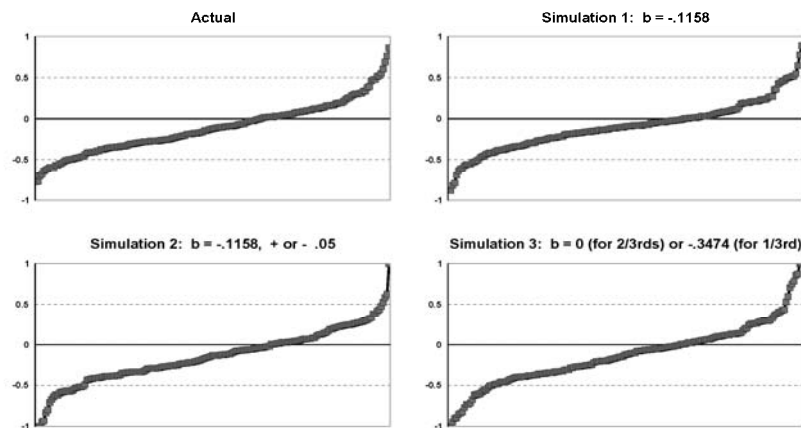
The results of the disaggregated heterogeneity analysis conducted by the defendant's expert under different scenarios appear in the bottom three rows of Table 4 and in Figure 5. No matter what the actual pattern of disparities across offices—uniform (no heterogeneity), continuous, or radical—given the sample sizes in the UFS example, the overall pattern of results produced by the defendant's expert are always the same. They always show a pattern of seemingly random results, favoring men in some offices, women in other offices, and no significant disparity in most offices. Indeed, a cynic might argue that the defendant's expert chose to conduct this kind of analysis in order to guarantee a result supporting her client's litigation position, regardless of the underlying facts.

Table 4. Testing for Heterogeneity in Gender Disparities Under Alternative Scenarios

	Actual	Sim 1	Sim 2	Sim 3
Type of heterogeneity across offices in gender effect		<i>uniform</i> at <i>-.116</i>	<i>continuous</i> - <i>.066</i> to <i>-.166</i>	<i>radical</i> 0 or <i>-.37</i>
Gender coef., Controlling experience, office	-0.116	-0.108	-0.136	-0.121
Percentage gender disparity, controlling experience, office	-10.9%	-10.2%	-12.7%	-11.4%
Observations	7941	7941	7941	7941
Number offices	181	181	181	181
Gender coef. t ratio controlling experience, office	-5.76	-5.33	-6.72	-5.95
Number of offices with no significant disparity	170	174	167	160
Number of offices with significant disparity favoring women	2	2	3	6
Number of offices with significant disparity favoring men	9	5	11	15
F-ratio for test of gender by office interaction	0.91	0.97	1.04	1.48
p-value for test of gender by office interaction	0.811	0.612	0.343	< .0001

Figure 5. Results of a Disaggregated Analysis Under Different Patterns of Heterogeneity

Gender Effects Estimated Separately by Office



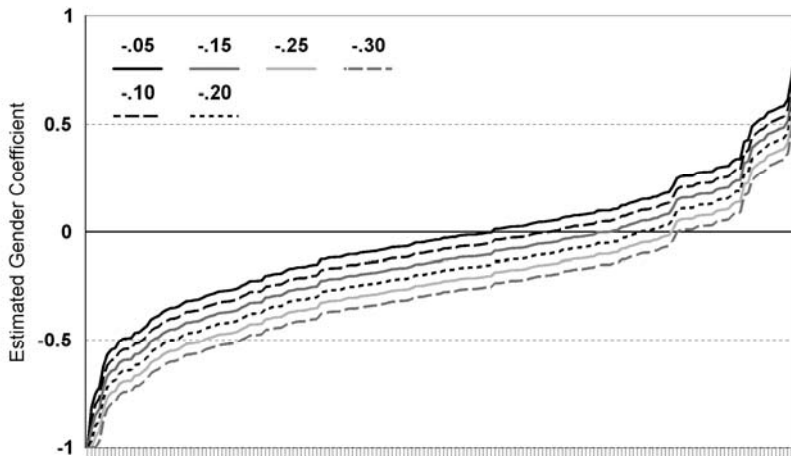
The figures above raise the question: how large would the disparity have to be before the analysis of the defendant's expert shows any evidence that women face a systematic disadvantage at UFS? The results appear in Table 5 and Figure 6. Here, we have repeated the *Sim1* exercise, with simulated compensation data containing a built-in uniform bias towards women, but under six different circumstances, allowing the amount of uniform bias to range from -.05 to -.30 for estimate of the gender difference (i.e., gender gaps ranging from 5% to about 25%).¹⁵¹

TABLE 5 GOES HERE

¹⁵¹ So, for instance, in the most extreme example with the gender effect set to -.30, data are generated such that for two employees of the same sex in the same office with identical lengths of service, their *simulated* earnings will be identical, but for the random component introduced by the disturbance term. However, if one is a man and the other is a woman with the same experience, their simulated earnings will be identical but for the random component *and* a 25% penalty subtracted from the woman's earnings.

Figure 6. Results of a Disaggregated Analysis Applied to Simulated Data Containing Uniform Disparity of Varying Magnitude

**Estimated Effects of Gender Vary Across Offices;
Underlying Gender Coefficient Ranges from -.05 to -.30**
(offices displayed from most negative to least negative gender coefficient)



As would be expected, the greater the amount of uniform disparity, the fewer instances when the *estimated* coefficient for a specific office shows a significant disparity favoring women, and once the disparity exceeds $-.15$, the analysis produces no results that are statistically significant in favor of women. Similarly, the number of offices where the estimated coefficient is significant and favors men increases as the underlying disparity increases, and when the underlying disparity gets as high as $-.3$, significant disparities are detected in 32 of the 181 offices. But as can be seen from Figure 6, even in the circumstance where the underlying gender bias is uniform and substantial, the disaggregated analysis yields a result consistent with the defendant's litigation position: Widely ranging disparities sometimes favoring women, other times favoring men, but in the vast majority of instances not statistically significant. Consistent with rulings such as *Abram v. UPS of America, Inc.* and the arguments set forth in the NERA publications, the defendant could argue that while there may be problems in a handful of offices, each due to unique

circumstances, the practice of decentralized decision making coupled with the statistical evidence convincingly supports a decision to deny class certification.

CONCLUSION: RECOMMENDATIONS FOR COURTS

This exercise demonstrates that evaluating the nature and character of organization-wide discrimination presents a potentially daunting technical challenge. Further, it implicitly raises a question that is rarely asked regarding statistical evidence used in determining the merits of certifying a class: how much heterogeneity across organizational subunits in the level of disparity between dominant and subordinate groups is sufficient to justify denying class certification? We are working on a future paper that considers what kinds of alternate statistical models might better address the relevant legal questions while ensuring sufficient power to conduct a reliable analysis. However, even if parties can bring a better statistical methodology to court, judges will still face the challenge of determining how it should be applied and what questions it can properly answer.

In other words, even if one can develop a more reliable statistical method for determining the extent of subunit variation, the legal relevance and broader meaning of those differences remain contested for at least two reasons. First, courts must be attentive to the relevant legal and procedural framework. Consideration of this level of factual nuance at the class certification stage of the case is inconsistent with preserving a real boundary between procedural hurdles and merits adjudication. That boundary serves important policy interests, both in terms of avoiding protracted and expensive litigation of a preliminary procedural issue and in a fair operation of the rules of preclusion.¹⁵² Second, it is clear that even the plaintiffs' burden of proof at trial falls far short of showing consistently equivalent levels of disparity across an entire company.¹⁵³ Simply put, the commonality test is not, and never has

¹⁵² The Supreme Court has explained that a key purpose of the Rule 23 amendments is to avoid so-called "one-way intervention" practices under the prior class action rule. One-way intervention allowed potential plaintiffs to sit on the sidelines and wait to see if the class claims succeeded or failed. If they succeeded, individuals could intervene and recover, but they took no risk of being bound by an adverse ruling against the class. In adopting the 1966 amendments, the Federal Rules made clear that class determination must precede merits, to ensure that all class members would be identified and bound prior to a merits ruling. *Am. Pipe & Constr. Co. v. Utah*, 414 U.S. 538, 547–48 (1974).

¹⁵³ A pooled analysis clearly suffices to carry the burden for plaintiffs on the merits. *See, e.g., Int'l Bhd. of Teamsters v. United States*, 431 U.S. 324, 337 (1977). Indeed, a federal district court recently denied summary judgment in a Title VII class case, ruling that aggregated data could be sufficient to support a pattern or practice claim. *McReynolds v. Sodexho Marriott Servs., Inc.*, 349 F. Supp. 2d 1, 15 (D.D.C. 2004).

been, a uniformity test. Thus, what variation among subunits might mean in terms of the legal questions implicated by either Rule 23 commonality or Title VII “pattern or practice” language is a critical issue that courts must consider in conjunction with the use of any statistical test in litigation.

Indeed, a rigorous analysis of this problem raises core questions regarding appropriate class-wide Title VII enforcement. At one extreme, if the *actual* (as opposed to estimated) level of disparity across organizational subunits truly does vary randomly from subunit to subunit, favoring males in some subunits, females in others, and neither group in still others, with no apparent basis in how the organization’s policies and practices are designed and implemented, then the statistical basis for certifying a class does not exist. On the other hand, should *any* level of heterogeneity be sufficient to defeat class certification? If a company as a matter of policy establishes a discretionary management regime and fails to ensure equal employment opportunity, and as a result some subunits discriminate just a little and others discriminate a lot, has the company insulated itself from being sued for systematic discrimination?

Put another way, suppose there is a discriminatory mechanism that operates company-wide and creates a uniform level of disparity across subunits, but in addition, there are subunit-specific practices that create additional barriers—and additional disparity—in some units but not others. There is clearly heterogeneity in the organization, yet this is surely a company that practices systematic discrimination of the type that the class action mechanism was meant to address. This kind of individual-level variation in harm is akin to the type of problem that remedial relief is designed to address, suggesting that much of the issue may be best addressed in the context of allocation.¹⁵⁴ At this point, the questions of just how much consistency or variation in the impact of a discriminatory practice is sufficient to establish liability and the ways that remedies can be most fairly tailored to the areas and individuals who actually suffered harm, remain open.

¹⁵⁴ Typically, these kinds of cases involve post-liability proceedings, either through mini-trials or formula-based distributions, to determine how to calculate and allocate class relief. Even if there is no statistically detectable discrimination in certain areas of the company, some individuals in those areas might still be able to prove actual harm at the relief phase of the case. It would be unfair to exclude them from the case at the outset under the theory that they are less likely to have been injured because of where they work. If ultimately no injury is shown at the remedial phase for those areas or individuals, the defendant will not have to pay their claims. Thus, a generous standard for class certification, including all parts of the company subject to a potentially discriminatory policy, regardless of demonstrated consistent impact, would be fair to both sides, and would best allocate the risk of error.

This Article's hypothetical example represents a real world problem for courts trying to make sense of their responsibilities under Rule 23. Given that today's large corporations frequently span many locations and jurisdictions—and increasingly adopt decentralized and highly discretionary management policies—the operation of this heterogeneity defense can preclude meaningful class relief. We can expect this issue to continue to confront courts considering class certification motions. Ultimately, our simulated results not only illustrate the limits of relying solely on statistical tests to make determinations about class treatment, but also show what kinds of questions courts should be asking at class certification and how to better understand the relationship between proof of class-wide harm and the threshold issue of commonality.

Courts should understand how statistical evidence is relevant to the particular questions before them, reserving battles over aggregation to the appropriate phase of the case. To avoid relying on potentially misleading statistical presentations, courts should carefully evaluate the relevance of statistical evidence to class certification issues, consider any such evidence in tandem with other available evidence, and return to the framework of a class–merits distinction. In cases where courts are relying on statistical proof of the nature or distribution of workforce disparities to rule on class certification, they are crossing the class–merits divide. Whether a pattern or practice of discrimination exists is a different question than whether class-wide evidence will allow an efficient litigation of this issue.

Imagine an alternative scenario—a classic disparate impact problem. Suppose instead of excessive managerial discretion, plaintiffs complained about gender bias in a written test used to differentiate employees for compensation purposes. Even if UFS administered an identical test throughout the company, outcomes would not be uniform. Some women would likely do extremely well on the test, and some men would do extremely poorly. Varying gender disparities across subunits would not show the absence of a uniform policy or the absence of discrimination.

At the merits phase of litigation, the kind of simulation demonstrated here provides courts with one tool to understand how statistical power may drive the outcomes in the cases they are considering. The analysis above shows that in some circumstances, a disaggregated approach can be guaranteed to support the defendant's litigation position even when the data upon which it is based come from a context where members of the protected group face a uniform,

consistent, and substantial disadvantage. In *any* organization-wide analysis of group differences on an outcome (whether that outcome is earnings, test scores, performance ratings, promotion rates, or something else), if there is variation within groups that cannot be fully explained by measured factors, disaggregation will necessarily result in variation across subunits in estimates of disparities. The smaller the size of the subunits and the larger the impact of unmeasured factors, the more certain it will be that low statistical power will guarantee that sampling variability—the luck of the draw—results in the appearance of heterogeneity, regardless of whether it in fact exists. In any specific litigation context, whether the kind of disaggregated analysis advocated by experts for defendants is at all probative of the issues facing the court can be determined relatively easily by doing the kind of simulation exercise reported here for the hypothetical UFS (which is roughly equivalent to doing a formal analysis of statistical power).

Moreover, undertaking such an exercise forces the parties to the litigation to be clear about the precision of their statistical tests and the kinds of discriminatory patterns their analyses are likely to detect and are likely to miss. In formal terms, it forces them not just to define the null hypothesis but substantively meaningful alternatives as well. Doing so also forces the statistical expert to link her analysis more closely to an understanding of the organizational mechanisms presumed to generate (from the plaintiffs’ perspective) or minimize (from the defendant’s perspective) disparities, at least when the expert has an incentive to devise a more powerful statistical test.

Such a linkage would involve examining statistical evidence in tandem with other evidence about the organization and its practices. Asking if apparently higher and lower levels of disparity can be reasonably attributable to relevant and specifically identified policy differences is a far more pertinent inquiry than simply testing subunit coefficients against a standard of uniformity. For example, with more complete information about how the company designed and implemented personnel policy, the statistical expert in our UFS case might have had reason to believe that barriers were concentrated in one specific area of the organization, for example, in a specific division where oversight was especially lax and criteria for personnel actions particularly vague. That type of information could guide statistical inquiry at the appropriate point in the litigation using methods that take proper account of statistical power concerns.

Ultimately, courts need better tools to determine what analyses are and are not probative of the issues being decided. They would avoid repeating the error of the *Abram vs. UPS of America, Inc.* decision, which encourages statistical analyses that have the appearance of scientific rigor but are constructed in a way that guarantees support of one side's litigation position regardless of the facts.