Toward a Coherent Test for Disparate Impact Discrimination

JENNIFER L. PERESIE[*]

Statistics are generally plaintiffs' primary evidence in establishing a prima facie case of disparate impact discrimination. Thus, the use, or misuse, of statistics dictates case outcomes. Lacking a coherent test for disparate impact, courts choose between the two prevailing tests, statistical significance and the four-fifths rule, in deciding cases, and these tests frequently produce opposite results. Litigants thus face considerable uncertainty and the risk that a judge's preferred outcome will dictate which test is applied. This Article recognizes that the two tests perform complementary functions that both play a useful role in determining whether liability should be imposed: statistical significance establishes that the challenged practice likely caused the disparity, and the four-fifths rule establishes that the disparity is large enough to matter. Rather than choose between the two tests, courts should use a uniform and coherent standard that combines both them. Determining the parameters of this standard involves difficult policy decisions about the purposes of the doctrine as well as who, and to what extent, should bear the risk of error.

INTRODUCTION

The famous adage "There are three kinds of lies: lies, damned lies, and statistics"[1] sheds light on the difficulties courts face when they rely on statistics in deciding cases.

[*] Associate, Kellogg, Huber, Hansen, Todd, Evans & Figel, P.L.L.C., Washington, D.C.; J.D., Yale Law School; M.A., B.A., Emory University. I am especially indebted to Brianne Gorod and Jim Metzger for their helpful suggestions on numerous drafts. I also wish to thank the Honorable Morris S. Arnold, Richard Brooks, Erin Morrow, Vicki Schulz, and my Statistics and the Law students at the University of Arkansas at Little Rock William H. Bowen School of Law for their inspiration and encouragement.

[1] 1 SAMUEL LANGHORNE CLEMENS, MARK TWAIN'S AUTOBIOGRAPHY 246 (1924) (attributing the remark to Benjamin Disraeli); *see also* DARRELL HUFF, HOW TO LIE WITH STATISTICS (1954).

Litigants and courts can and do cherry pick those statistics that produce the desired results.

This problem is particularly acute in Title VII disparate impact discrimination cases because statistics are plaintiffs' key evidence in establishing a prima facie case of disparate impact and because the two primary statistical tests—statistical significance and the four-fifths rule—often lead to contrary results. Briefly stated, under the four-fifths rule, a disparity is actionable when one group's pass rate is less than four-fifths (eighty percent) of another group's pass rate.[2] Under statistical significance tests, a disparity is actionable when we can be confident at a specified level—generally ninety-five percent—that the observed disparity is not due to random chance.[3]

Courts are forced to choose between the two tests. Not surprisingly, plaintiffs advocate whichever statistical test allows them to establish a prima facie case, while defendant-employers either urge the test that shows no significant disparity or suggests that the requirements of the chosen test be strengthened such that any disparity diminishes or disappears entirely.[4] Lacking a coherent metric, judges might use their choice between the two tests to advance their normative views of when liability should attach to employers for racial or gender disparities or even to justify a decision reached on some other grounds (e.g., dislike of a party).

This lack of coherency extends beyond Title VII to cases involving claims of disparate impact under the Age Discrimination in Employment Act (ADEA)[5] and the Equal Credit Opportunity Act (ECOA),[6] claims of disparate discriminatory effect under the Fair Housing Act (FHA),[7] and claims brought under state antidiscrimination laws.[8] The doctrinal confusion also has implications beyond litigated cases. Although plaintiffs' success in litigated disparate impact cases is relatively low,[9] it is likely that

[2] 29 C.F.R. § 1607.4(D) (2008).

[3] *See* RAMONA L. PAETZOLD, STEVEN L. WILLBORN & DAVID C. BALDUS, THE STATISTICS OF DISCRIMINATION: USING STATISTICAL EVIDENCE IN DISCRIMINATION CASES, § 2.04, at 2–13 (2006).

[4] *See, e.g.*, Groves v. Ala. State Bd. of Educ., 776 F. Supp. 1518, 1527 (M.D. Ala. 1991) ("From this welter of statistics, the [defendant] predictably fastens upon the result under the four-fifths standard, and concludes that the . . . requirement does not disproportionately exclude blacks . . . .").

[5] 29 U.S.C. § 623(a) (2000); *see* Smith v. City of Jackson, 544 U.S. 228, 240 (2005).

[6] 15 U.S.C. § 1691 (2006). Although courts have permitted disparate impact claims to proceed under the ECOA, it is still unsettled whether such claims are available under the statute. *See generally* Peter N. Cubita & Michelle Hartmann, *The ECOA Discrimination Proscription and Disparate Impact—Interpreting the Meaning of the Words That Actually Are There*, 61 BUS. LAW. 829 (2006).

[7] 42 U.S.C. § 3604 (2000); *see also* Graoch Assocs. #33, L.P. v. Louisville/Jefferson County Metro Human Relations Comm'n, 508 F.3d 366 (6th Cir. 2007) (holding that a plaintiff must present statistical evidence to make out a disparate impact claim under the FHA).

[8] *Compare* Kohn v. City of Minneapolis Fire Dep't, 583 N.W.2d 7, 13 (Minn. Ct. App. 1998) (applying the four-fifths rule to a claim brought under the Minnesota Human Rights Act), *with* Strand v. Interlachen Country Club, No. C0-01-1826, 2002 WL 1365637, at *6–*7 (Minn. Ct. App. June 25, 2002) (using statistical significance to evaluate a claim brought under the Minnesota Human Rights Act and the Minnesota Age Discrimination Act).

[9] *See* Michael Selmi, *Was the Disparate Impact Theory a Mistake?*, 53 UCLA L. REV. 701, 738–43 (2006) (presenting an empirical analysis of disparate impact claims).

plaintiffs are using these two statistical tests to gauge whether to bring (or threaten to bring) cases, and defendant-employers are using the tests to determine when to settle and how to structure testing and use the results to minimize potential liability.[10]

A recent Sixth Circuit decision, *Isabel v. City of Memphis*,[11] is emblematic of the difficulty of dueling statistical tests in disparate impact cases. In that case, the plaintiffs, black sergeants in the Memphis Police Department, sued the City of Memphis, alleging that the written exam given for promotion to lieutenant discriminated against minority candidates. The disparity between the selection rates of white and minority candidates was statistically significant, but it was not actionable under the four-fifths rule.[12] The plaintiffs, of course, advocated for a statistical significance test, while the City urged reliance on the four-fifths rule.

The district court applied the plaintiffs' chosen test and ruled for the plaintiffs. The court's decision required the Department to promote the minority sergeants who had failed the test to lieutenant, and it awarded them back pay and attorney fees. The Sixth Circuit affirmed and held that establishing statistical significance was sufficient to show disparate impact. The court declined, however, to adopt statistical significance as the standard test and instead embraced a case-by-case approach for identifying actionable disparate impact.[13] The Sixth Circuit decision appropriately recognizes that the two different statistical tests perform valuable functions. But in so doing, the decision furthers the existing doctrinal confusion and the arbitrary selection of the test applied in a given case. The dissent in *Isabel* properly criticized the majority for "provid[ing] absolutely no guidance as to which tests are to be used in assessing whether an employment practice results in an adverse impact."[14]

The Sixth Circuit is not an outlier—none of the circuits have a uniform standard for evaluating disparate impact cases.[15] Courts' unwillingness to adopt a standard might reflect the inadequacy of the two existing tests, standing alone, in evaluating whether a plaintiff establishes actionable disparate impact. Each test addresses a separate inquiry: statistical significance tests ask whether the plaintiff has established causation, that is

[10] *See, e.g.*, Cotter v. City of Boston, 193 F. Supp. 2d 323, 329–31 (D. Mass. 2002) (noting that the police department promoted additional black officers to comply with the four-fifths rule); Duane Bourne, *Virginia Beach Agrees to Change the Way It Scores Police Math Exams*, THE VIRGINIAN-PILOT (Norfolk, Va.), April 3, 2006, http://hamptonroads.com/node/86031 (detailing the decision of the Virginia Beach Police Department to eliminate its math exam for police officers based on the racial disparity in pass rates).

[11] 404 F.3d 404 (6th Cir. 2005).

[12] In *Isabel*, fifty-one of fifty-seven white candidates (89.5%) and forty-seven of sixty-three minority candidates (74.6%) passed the test. This difference is statistically significant at a 0.05 level of significance, but is not actionable under the four-fifths (eighty percent) rule because the minority passage rate is 83.4% of the white passage rate. *Id.* at 417.

[13] *See id.* at 412 (citing Int'l Bhd. of Teamsters v. United States, 431 U.S. 324, 339–40 (1977)).

[14] *Id.* at 418 (Batchelder, J., dissenting).

[15] *See, e.g.*, Smith v. Xerox Corp., 196 F.3d 358, 366 (2d Cir. 1999) ("Although courts have considered both the four-fifths rule and standard deviation calculations in deciding whether a disparity is sufficiently substantial to establish a prima facie case of disparate impact, there is no one test that always answers the question. Instead, the substantiality of a disparity is judged on a case-by-case basis.").

whether the disparity is statistically significant; and the four-fifths rule asks whether the plaintiff has shown that the law should be concerned, that is whether the disparity is practically significant.

This Article argues that the two tests fulfill complementary roles and thus should be viewed not as alternatives, but as components of a new and coherent two-part test for establishing a prima facie case of disparate impact. Requiring statistical significance ensures that courts can be sufficiently certain (at the specified statistical level) that a given disparity is not caused by chance. The four-fifths rule ensures that the disparity is large enough to matter. We can and should quibble over the purposes of the doctrine and the specific parameters of both tests, but we should acknowledge the need for plaintiffs to satisfy both inquiries.

Part I of the Article describes the disparate impact doctrine and the two primary tests for disparate impact, statistical significance and the four-fifths rule. Part II evaluates the two tests—first, showing how mathematically they yield different results, second, explaining how the tests perform different yet complementary functions in evaluating disparate impact, thus warranting courts' use of a coherent test that combines the tests. Part III identifies the factors that policy makers should consider in establishing the appropriate level for each test and suggest a starting point for that debate.

I. Understanding Disparate Impact

A. The Doctrine

Title VII prohibits employment discrimination based on race, color, religion, sex, or national origin.[16] At one time, it was thought that the statute only applied to direct acts of discrimination, for example, an employer who refuses to hire blacks, but the Court established in the 1971 case of *Griggs v. Duke Power Co*.[17] that employers could be held liable under Title VII where a test or facially neutral employment practice disproportionately impacts a protected group.[18] Under this doctrine, employers are legally responsible when their selection practices create a nondiverse workforce unless they can show a business justification for those practices.[19]

Griggs involved a challenge by black employees to Duke Power's high school education and testing requirements for promotion. The plaintiffs showed that large disparities existed in the statewide graduation rates of blacks and whites and in each race's pass rate on the written test, but presented no evidence that Duke Power harbored any malicious intent in adopting the requirements, which resulted in the promotion of significantly more whites than blacks. The Supreme Court ruled that Duke Power was liable for disparate impact discrimination, concluding that an employer's "good intent or absence of discriminatory intent does not redeem employment procedures or testing mechanisms that operate as 'built-in headwinds' for minority groups and are unrelated to measuring job capability."[20] As the Court explained in a subsequent

[16] Civil Rights Act of 1964, 42 U.S.C. § 2000e-1 to -15 (2000).

[17] 401 U.S. 424 (1971).

[18] *Id.* at 432.

[19] Wards Cove Packing Co. v. Atonio, 490 U.S. 642, 659–60 (1989).

[20] *Griggs*, 401 U.S. at 432.

decision, "the necessary premise of the disparate impact approach is that some employment practices, adopted without a deliberately discriminatory motive, may in operation be functionally equivalent to intentional discrimination."[21] Though commentators long debated whether *Griggs* was rightly decided,[22] Congress essentially mooted that debate when it codified *Griggs* in the 1991 Civil Rights Act.[23]

Post-*Griggs* Supreme Court cases have strengthened the disparate impact doctrine. The Court held that employers could be found liable where an individual practice had a disparate impact, but the overall selection process had no disparate impact.[24] Even more significantly, the Court held that courts could analyze subjective employment policies, for example, interviews or job evaluations, under the disparate impact framework.[25] Under this decision, employers without strict hiring guidelines who hire those they, or individual supervisors, deem most qualified can be held liable for disparate impact discrimination if the end result of their hiring processes is not a representative workforce.[26]

Under the disparate impact doctrine, courts have invalidated numerous employment practices, including written tests,[27] physical tests,[28] height and weight requirements,[29] and subjective evaluation processes,[30] for having a disparate impact on a protected class without a business justification.

To establish a prima facie case of disparate impact, plaintiffs must show that a particular employment practice caused an adverse impact on the basis of a protected status, such as race.[31] Plaintiffs generally prove such causation by comparing the selection rates of majority and minority applicants for a position and then showing that the disparity is statistically significant or that it violates the four-fifths rule. The Supreme Court has rejected a "rigid mathematical formula" for disparate impact, providing instead the ambiguous guidance to lower courts that "statistical disparities

[21] Watson v. Fort Worth Bank & Trust, 487 U.S. 977, 987 (1988).

[22] *See, e.g.*, Hugh S. Wilson, *A Second Look at* Griggs v. Duke Power Company*: Ruminations on Job Testing, Discrimination, and the Role of the Federal Courts*, 58 VA. L. REV. 844, 844 (1972) (criticizing *Griggs* for "provid[ing] a strong imprimatur for a freewheeling use of Title VII by the lower courts").

[23] Civil Rights Act of 1991, Pub. L. No. 102-166, 105 Stat. 1071 (codified as amended in scattered sections of 2, 16, 29, and 42 U.S.C. (2000 & 2006)).

[24] Connecticut v. Teal, 457 U.S. 440, 442 (1982).

[25] *Watson*, 487 U.S. 977. The Court's decision resolved a circuit split on this issue.

[26] *See, e.g.*, McClain v. Lufkin Indus., 187 F.R.D. 267, 273–77 (E.D. Tex. 1999) (certifying a class action in a disparate impact case challenging the employer's subjective hiring practices).

[27] *See, e.g.*, Fickling v. N.Y. State Dep't of Civil Serv., 909 F. Supp. 185, 193 (S.D.N.Y. 1995) (concluding that written examination for welfare eligibility examiners had a racially disparate impact).

[28] *See, e.g.*, Brunet v. City of Columbus, 642 F. Supp. 1214 (S.D. Ohio 1986).

[29] Dothard v. Rawlinson, 433 U.S. 321 (1977) (holding that a height and weight requirement for prison guards had a disparate impact on female applicants).

[30] *See, e.g.*, Stender v. Lucky Stores, Inc., 803 F. Supp. 259, 335–36 (N.D. Cal. 1992) (concluding that the employer's "standard policy of discretionary, subjective and frequently unreviewed decision making with respect to initial placement, promotion and training" had a disparate impact on women).

[31] *See* 42 U.S.C. § 2000e-2(k)(1)(A) (2000).

must be sufficiently substantial that they raise . . . an inference of causation."[32] Once a plaintiff establishes a "sufficiently substantial" disparity, however defined, the burden passes to the employer-defendant to rebut the plaintiff's statistics[33] or to show that the challenged practice is "job related" and "consistent with business necessity."[34] If the defendant establishes business necessity, a standard which is rather deferential to employers, the plaintiff can still prevail by proving that an alternative practice has less discriminatory effect and that the defendant failed to adopt it.[35] This standard is difficult to meet, however, because "[f]actors such as the cost or other burdens of proposed alternative selection devices are relevant in determining whether [the alternative test] would be equally as effective as the challenged practice in serving the employer's legitimate business goals."[36]

Plaintiffs must establish disparate impact with respect to the pool of qualified persons in the relevant labor market for the given position.[37] Most often, plaintiffs present statistics from the actual applicant pool for the position.[38] Plaintiffs might also choose to use national population statistics;[39] state data, as in *Griggs*;[40] or data from a smaller geographic area.[41]

Because neither the doctrine nor the statutes specify the statistical showing required to establish disparate impact, courts make that decision within the context of particular cases. Even the most ardent judicial idealist will recognize that this creates the potential for judges to choose whatever test allows their preferred party to prevail. Although such potential certainly exists wherever there are conflicting precedents, this problem is particularly acute in disparate impact cases given the entrenchment of the two tests as valid alternatives and the almost complete lack of guidance in the case law about which test to apply. Well-meaning judges must play amateur statisticians in order

[32] Watson v. Fort Worth Bank & Trust, 487 U.S. 977, 995 (1988).

[33] For instance, the defendant could show that the test statistic was calculated incorrectly.

[34] 42 U.S.C. § 2000e-2(k)(1)(A). The Supreme Court has not clearly defined these terms. *See* Ernest F. Lidge III, *Financial Costs as a Defense to an Employment Discrimination Claim*, 58 ARK. L. REV. 1, 27–29 (2005). For a thorough discussion of how employers validate tests or other standards, see generally LEX K. LARSON, EMPLOYMENT DISCRIMINATION §§ 27.00–27.12 (2d ed. 2006).

[35] 42 U.S.C. § 2000e-2(k)(1)(A), (C).

[36] *Watson*, 487 U.S. at 988; *see also* N.Y. City Transit Auth. v. Beazer, 440 U.S. 568, 590 (1979) (noting that "any special rule . . . [that the defendant] might adopt is likely to be less precise—and will assuredly be more costly—that the one it currently enforces" (footnote omitted)); *id.* at 590 n.33, 591–92.

[37] Hazelwood Sch. Dist. v. United States, 433 U.S. 299, 308 (1977); *see also* Albemarle Paper Co. v. Moody, 422 U.S. 405, 425 (1975) (stating that the plaintiff must "show that other tests or selection devices, without a similarly undesirable racial effect, would also serve the employer's legitimate interest in 'efficient and trustworthy workmanship'" (quoting McDonnell Douglas Corp. v. Green, 411 U.S. 792, 801 (1973))).

[38] *See, e.g.*, Waisome v. Port Auth., 948 F.2d 1370, 1372 (2d Cir. 1991).

[39] *See* Dothard v. Rawlinson, 433 U.S. 321, 329–30 (1977).

[40] Griggs v. Duke Power Co., 401 U.S. 424, 430 (1971).

[41] *See, e.g.*, EEOC v. Joe's Stone Crab, Inc., 220 F.3d 1263, 1277 (11th Cir. 2000) (finding the applicant data unreliable and instead using data from "those local food servers who were theoretically 'available' and 'qualified' to work at [the employer]").

to determine the proper outcome.[42] Moreover, the use of statistics in disparate impact cases creates a false, and highly problematic, sense of objectivity.

The lively debate in the literature about the proper function of disparate impact doctrine further compounds the difficulty judges face in choosing the appropriate test. Professor Richard Primus has aptly summarized the tension between the two primary competing views of the disparate impact doctrine as follows: one view sees disparate impact as an "evidentiary dragnet designed to discover hidden instances of intentional discrimination," while the other views the doctrine as a "more aggressive attempt to dismantle racial [and other] hierarchies."[43] Where there is no evidence of bad intent on the part of the employer, judges who characterize disparate impact as a means of smoking out employers with animus toward a protected class (Primus's first view) might be more willing to choose whatever statistical test favors the defendant. Conversely, judges who see the doctrine as a grand way of leveling the playing field between different groups of people (Primus's second view) might err on the side of penalizing employers and be willing to impose liability whenever the plaintiff can satisfy either of the tests.

We need look no further than *Isabel* to see evidence that judges may be using the statistical tests to further their normative views. In rejecting the use of the four-fifths rule, the district court noted that the rule should be used only "to the extent that [it is] useful . . . for advancing the basic purposes of Title VII."[44] This raises the question of what these "purposes" are and whether the judge chose the statistical significance test, which favored the plaintiff, to further the "purpose" of promoting equality—Primus's dismantling view of the doctrine. In contrast, Judge Batchelder's dissent might have been motivated by Primus's dragnet view of the doctrine. In dissent, Judge Batchelder pointed to the absence of anything "in the record to suggest that the City purposefully discriminated against African-Americans" as a reason to find for the defendants.[45] This conclusion might have motivated her insistence on the four-fifths rule, which favored the employer.

[42] *Groves v. Alabama State Board of Education*, 776 F. Supp. 1518 (M.D. Ala. 1991), is illustrative of the difficulty judges face in evaluating statistics. The judge noted that neither party "was able to fashion a perfect statistical picture" and ultimately concluded that "this is one of those rare cases where if one stands back and applies reason and common sense the answer is apparent." *Id*. at 1529.

[43] Richard A. Primus, *Equal Protection and Disparate Impact: Round Three*, 117 HARV. L. REV. 494, 518 (2003); *see also* Michael Perry, *A Brief Comment on Motivation and Impact*, 15 SAN DIEGO L. REV. 1173, 1178–81 (1978) (arguing that laws or policies with a disparate impact must be subject to "an unusually heavy burden of justification . . . for the simple reason that the disproportionate character of the impact is not ethically neutral but is a function of prior massive societal discrimination against blacks"). The Eleventh Circuit has characterized disparate impact as a "doctrinal surrogate for eliminating unprovable acts of intentional discrimination hidden innocuously behind facially-neutral policies or practices." *Joe's Stone Crab, Inc.*, 220 F.3d at 1274.

[44] Isabel v. City of Memphis, No. 01-2533 ML/BRE, 2003 WL 23849732, at *3 n.5 (W.D. Tenn. Feb. 21, 2003), *aff'd*, 404 F.3d 404 (6th Cir. 2005).

[45] Isabel v. City of Memphis, 404 F.3d 404, 418 (6th Cir. 2005) (Batchelder, J., dissenting).

B. The Statistical Tests

Below, this Part details the two primary statistical tests for establishing a disparate impact—the four-fifths rule and statistical significance. But first a brief statistical primer on false positives, false negatives, and causation is necessary.

False positives and false negatives are the two types of "error" in statistical analysis. As the name indicates, false positives are false findings of discrimination—that is, the erroneous conclusion that a fair practice, which would have no disparate impact in the relevant labor market, is discriminatory. False negatives are erroneous findings that a discriminatory practice has no disparate impact. False positives are onerous for employers who must bear the burden of justifying their practices or facing liability, while false negatives undermine the role of Title VII in identifying and eliminating discrimination. There is an unavoidable trade-off between false positives and false negatives. Choosing the appropriate statistical test involves policy considerations about how much error of each type is acceptable.

False positives and false negatives are inevitable in disparate impact cases because of the use of samples. "Where only sample data is available, the disparate impact observed in a single sample of individuals drawn from the relevant population . . . may not justify the conclusion that the [challenged practice] has a discriminatory impact upon the population as a whole."[46] Said otherwise, the observed disparity might be a false positive, which is caused by something other than the challenged practice. Readers who have taken statistics might remember this problem by the shorthand: correlation does not equal causation. The sample might produce a false negative if other factors are masking a real disparity.

Causation is important because employers are only liable for disparities where plaintiffs establish that the challenged employment practice "causes a disparate impact."[47] It is not enough for a plaintiff to show that an observed disparity exists. The plaintiff must establish with some certainty that the challenged practice, not chance or some other factor, is the cause of the disparity. Because courts can never know the counterfactual (what the selection rates would be in the absence of the challenged practice), they instead impose liability when the plaintiff satisfies one of the two prevailing tests, thus satisfying the burden of proof by a "preponderance of the evidence."[48] Both the four-fifths rule and statistical significance tests, however, are notably imperfect indicators of causation.[49] To avoid false positives, the tests build in

[46] Fudge v. City of Providence Fire Dep't, 766 F.2d 650, 658 (1st Cir. 1985); *see also* United States v. Lansdowne Swim Club, 713 F. Supp. 785, 809 (E.D. Pa. 1989) ("The danger posed by small samples is that they may produce short-term results that would not hold over the long run, and thus erroneously may be attributed to discriminatory practices rather than to chance."), *aff'd*, 894 F.2d 83 (3d Cir. 1990).

[47] 42 U.S.C. § 2000e-2(k)(1)(A)(i) (2000).

[48] Bazemore v. Friday, 478 U.S. 385, 400 (1986) ("A plaintiff in a Title VII suit need not prove discrimination with scientific certainty; rather, his or her burden is to prove discrimination by a preponderance of the evidence.").

[49] *See* Kingsley R. Browne, *Statistical Proof of Discrimination: Beyond "Damned Lies"*, 68 WASH. L. REV. 477 (1993) (arguing for reduced reliance on statistics in discrimination cases because statistical analyses lead courts to exclude chance as a cause of disparities).

room for the possibility that random differences between applicants or employees—what statisticians term sampling error—might be causing the observed disparity.

Ideally, courts could assess causation by looking at whether the challenged practice would have a disparate impact if implemented on the relevant population. Indeed, this is how the Supreme Court determined *Dothard v. Rawlinson*, where plaintiffs claimed that a height and weight requirement created a disparate impact on female applicants (who tended to be shorter and weigh less than male applicants).[50] But establishing causation in this way is rarely feasible because of the nature of the challenged practices. A court cannot, for instance, make every person in the relevant labor market take a written test in order to determine whether that test is fair. Thus, plaintiffs in disparate impact cases typically use applicant data as a proxy for the overall pool.[51]

1. The Four-Fifths Rule

The Equal Employment Opportunity Commission (EEOC), the federal agency charged with enforcing federal civil rights laws,[52] adopted the four-fifths rule in the wake of the *Griggs* decision.[53] Since then, the Department of Labor,[54] the Department of Justice, and the Office of Personnel Management (formerly known as the Civil Service Commission),[55] have also adopted the four-fifths rule for measuring for disparate impact.[56] Under the rule,

> A selection rate for any race, sex, or ethnic group which is less than four-fifths (4/5) (or eighty percent) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact, while a greater than four-fifths rate will generally not be regarded by Federal enforcement agencies as evidence of adverse impact.[57]

To apply the rule, litigants calculate the selection rate for each group and then divide the selection rate of the minority group (SR_{min}) by the selection rate of the majority group (SR_{maj}): Selection ratio = SR_{min}/SR_{maj}.

If the ratio is lower than four-fifths or eighty percent, then the court can, but is not required to, find a prima facie case of disparate impact.

[50] 433 U.S. 321 (1977). The Court determined that the plaintiffs, who were challenging Alabama's height and weight requirement for prison guards, could use national height and weight data because this data was unlikely to "differ markedly" from the Alabama data. *Id*. at 330.

[51] A significant problem with using the applicant pool for analysis is that knowledge of the challenged practice might weed out prospective applicants. For instance, a short female is unlikely to apply for a job at a company with a minimum height requirement that she does not meet. The result is that the observed disparity in the pool is smaller than the actual disparity in the labor market.

[52] *See* The U.S. Equal Employment Opportunity Commission, The Commission, http://www.eeoc.gov/abouteeoc/commission.html.

[53] Uniform Guidelines on Employee Selection Procedures, 41 C.F.R. § 60-3.4(D) (2008).

[54] *Id.*

[55] 5 C.F.R. § 300.103(c) (2008).

[56] 28 C.F.R. § 50.14(4)(D) (2008).

[57] *Id.*; 29 C.F.R. § 1607.4(D) (2008).

To illustrate: suppose that female applicants allege that a requirement that all police officers be able to lift 100 pounds has a disparate impact. Sixty percent of females in the applicant pool met this requirement compared to eighty percent of men. Dividing the two selection rates (0.6/0.8), we conclude that the selection ratio of females to males is seventy-five percent. Because this ratio is less than eighty percent, the disparity is actionable under the four-fifths rule.

Despite being the EEOC-endorsed standard for disparate impact, the four-fifths rule does a remarkably poor job of evaluating whether a plaintiff has established that the challenged employment practice "*causes* a disparate impact," as required under the statute.[58] Rather than addressing the causation question directly, the four-fifths rule sets a high bar for plaintiffs to meet to establish a prima facie case. The logic is that by requiring one group's selection rate to be less than eighty percent of the other group's rate, instead of just requiring that one rate be less than another, the test leaves enough room for other factors to operate. If the disparity is large enough to be actionable under the rule, then, as the logic goes, the jury reasonably can infer causation. As will be elaborated upon further below, the four-fifths rule is, at best, a proxy for what the statute mandates it measure—causation.

This problem is magnified by the fact that the EEOC's decision to adopt the four-fifths rule as the appropriate ratio, rather than, for example, the five-sixths or six-sevenths rule, seems to be completely arbitrary.[59] The EEOC itself cautions that the rule is "not a legal definition of discrimination, rather it is a practical device to keep the attention of enforcement agencies on serious discrepancies."[60] The four-fifths (eighty percent) rule is certainly preferable to the ninety-nine percent rule, under which almost any disparity would be actionable, because it leaves some room for the operation of chance or other factors (e.g., education, work experience) that might be causing the observed disparity. But it is uncertain how the proper amount of leeway was determined.[61]

Although the EEOC's guidelines are not "controlling," the Supreme Court consistently has recognized that they constitute "a body of experience and informed judgment to which courts and litigants may properly resort for guidance." As such, they

[58] 42 U.S.C. § 2000e-2(k)(1)(A)(i) (2000) (emphasis added).

[59] *See, e.g.*, Anthony E. Boardman, *Another Analysis of the EEOCC 'Four-Fifths' Rule*, 25 MGMT. SCI. 770 (1979); Elaine W. Shoben, *Differential Pass-Fail Rates in Employment Testing: Statistical Proof Under Title VII*, 91 HARV. L. REV. 793, 805–811 (1978). *But see* Paul Meier, Jerome Sacks & Sandy L. Zabell, *What Happened in* Hazelwood*: Statistics, Employment Discrimination, and the 80% Rule*, 1984 AM. B. FOUND. RES. J. 139 (praising the four-fifths rule).

[60] *See* Adoption of Questions and Answers to Clarify and Provide a Common Interpretation of the Uniform Guidelines on Employee Selection Procedures, 44 Fed. Reg. 11,996 (Mar. 2, 1979) [hereinafter Questions and Answers].

[61] One of the committee members described the eighty percent test as "born out of two compromises: (1) a desire expressed by those writing and having input into the Guidelines to include a statistical test as the primary step but knowing from an administrative point of view a statistical test was not possible for the FEPC consultants who had to work the enforcement of the Guidelines, and (2) a way to split the middle between two camps, the 70% camp and the 90% camp." DAN BIDDLE, ADVERSE IMPACT AND TEST VALIDATION: A PRACTITIONER'S GUIDE TO VALID AND DEFENSIBLE EMPLOYMENT TESTING 3 (2005).

are entitled to a "measure of respect."[62] However, neither the Supreme Court nor any federal circuit has adopted the four-fifths rule. Although courts often use the rule in disparate impact cases,[63] other courts have criticized or outright rejected it.[64]

The primary advantage of the four-fifths rule is its simplicity. Unlike statistical significance tests, which require sophisticated mathematical analysis, the four-fifths rule is easy to calculate. Further, it puts employers on notice of the relative balance an employer must achieve in its workforce to avoid litigation. Employers who use written tests might choose to recalculate the passing score based on test performance. Or they might even use affirmative action to avoid liability by promoting lower-scoring members of an underrepresented group. The Boston Police Department did just that in 1996 when twenty-nine nonblack officers, but only one black officer, passed a written exam. To bring the ratio into compliance with the four-fifths rule, the Department promoted three additional black officers instead of three nonblack officers with higher scores on the exam.[65] Those who view disparate impact as a means of achieving workforce diversity may see such deliberate maneuvers as positive evidence that employers are responding appropriately to Title VII. Others, including opponents of affirmative action, will consider these actions unwarranted impositions on an employer's ability to hire the most qualified workers. Those who prefer a different bar than the one imposed by the four-fifths rule may resent the use of the rule as a mandate for a specified level of diversity. It is also concerning that in specifying the desired ratio, the four-fifths rule institutes a permissible level of discrimination.

Moreover, even cheerleaders of the four-fifths rule must acknowledge that the rule itself creates a disparate impact: it burdens small employers more harshly than large employers because the addition or subtraction of as few as one employee will have a larger impact on the selection ratio and expose a small employer to liability but will have no noticeable effect on a large employer.[66] Consider for instance two employers—one with ten female applicants and one with 10,000 female applicants. For that first employer, every additional female applicant hired represents a ten percent difference in the female selection rate; for the second employer, in contrast, each additional female hired represents only a 0.01% difference.

[62] *E.g.*, Fed. Express Corp. v. Holowecki, 128 S.Ct. 1147 (2008) (quoting Skidmore v. Swift & Co., 323 U.S. 134 (1944)).

[63] *See, e.g.*, Clady v. County of L.A., 770 F.2d 1421, 1428 (9th Cir. 1985) (characterizing the four-fifths rule as a "rule of thumb").

[64] *See, e.g.*, Eubanks v. Pickens-Bond Constr. Co., 635 F.2d 1341 (8th Cir. 1980); Cormier v. P.P.G. Indus., Inc., 519 F. Supp. 211 (W.D. La. 1981), *aff'd*, 702 F.2d 567 (5th Cir. 1983).

[65] *See* Cotter v. City of Boston, 193 F. Supp. 2d 323, 330–31 (D. Mass. 2002), *aff'd in part and rev'd in part*, 323 F.3d 160 (1st Cir. 2003). The court held that the city's promotion of black officers instead of white officers was narrowly tailored to the compelling interest of remedying past discrimination, *Cotter*, 193 F. Supp. 2d at 357, and the appellate court affirmed, *Cotter*, 323 F.3d at 169–72. The disparity in the promotion rates would not have been actionable under a statistical significance test with a ninety-five percent significance level.

[66] *See* Shoben, *supra* note 59, at 806–10; *see also* Fudge v. City of Providence Fire Dep't, 766 F.2d 650, 658 n.10 (1st Cir. 1985) ("Where the size of the sample is small, . . . the 'four-fifths rule' is not an accurate test of discriminatory impact.").

The problem with the four-fifths rule is that a small employer with a small absolute disparity between male and female applicants might face liability under the rule, while a large employer can have a much greater disparity and still comply with the four-fifths rule.[67] Consider again our two employers and further assume that they have equal numbers of male and female applicants; our hypothetical large employer thus has 20,000 applicants, 10,000 men and 10,000 women, and the small employer has twenty applicants, ten men and ten women. In our hypothetical large employer, 3200 (or thirty-two percent) of the female applicants meet a hiring requirement, while 4000 (or forty percent) of the male applicants do so. In our hypothetical small employer, three female applicants (thirty percent) and four male applicants (forty percent) satisfy the requirement. The selection ratio for the large employer is eighty percent (0.32/0.40), so the disparity is not actionable under the four-fifths rule, even though it is highly statistically significant[68] and 800 more male applicants met the requirement than female applicants. The selection ratio for the small employer is seventy-five percent (0.3/0.4) so the disparity is actionable under the rule, even though it is far from statistically significant and only one more male applicant than female applicant met the requirement.

My guess is that if a lawsuit was brought against either of our hypothetical employers, the court would choose not to adhere to the four-fifths rule.[69] The court might conclude, as others have,[70] that the ten applicants at the small employer is too small a number to analyze using the four-fifths rule. Or it might conclude that the observed disparity in the large employer is large enough to matter, notwithstanding the employer's compliance with the four-fifths rule. But in cases less extreme than my hypothetical, small employers face liability under the rule, while similarly situated large employers avoid liability.[71] Indeed, inequities like the one described above are surely part of the reason that courts have failed to adopt a single uniform test for disparate impact, but have instead employed a flexible (and sometimes fickle) approach.

[67] To be sure, the small employer often can avoid liability relatively easily by hiring one or two more women. But this too highlights the weaknesses of the four-fifths rule or an arbitrary measure or disparate impact.

[68] The difference is significant at the 0.05 significance level.

[69] *See, e.g.*, Mems v. City of St. Paul, 224 F.3d 735 (8th Cir. 2000) (ruling for defendants even though the disparity violated the four-fifths rule because the sample size was too small to rule out chance).

[70] *See, e.g.*, EEOC v. Joint Apprenticeship Comm., 186 F.3d 110, 119 (2d Cir. 1999) ("Under these circumstances, we find that the application of the four-fifths rule to this particular fail ratio was inappropriate, because such a small sample would tend to produce inherently unreliable results."); Black v. City of Akron, 831 F.2d 131, 134–35 (6th Cir. 1987) ("Plaintiffs cannot allege or infer any adverse impact suggestive of discrimination from these figures.").

[71] *See* Shoben, *supra* note 59, at 810 (noting "the relatively favorable position of large employers compared to small employers under the Agency Guidelines").

2. Statistical Significance

Courts frequently use statistical significance tests instead of the four-fifths rule to evaluate disparate impact cases.[72] Unlike the four-fifths rule, statistical significance tests are a direct means of evaluating whether the plaintiff has established causation, as required under the statute. Statistical significance tests produce a test statistic that indicates at what level of mathematical certainty we can conclude that the practice causes a real disparity in the relevant labor market.

Statistical significance tests come in various technical forms, including multiple regressions, t-tests, Z-tests, the chi-square test, and the Fisher exact test, but they all calculate the probability that the observed disparity is due to chance. Where that probability is less than a specified value—most often five percent[73]—a disparity is said to be "statistically significant." For an observed disparity to be statistically significant it must be sufficiently large, given the sample size, to outweigh the possibility of random fluctuation.[74]

To illustrate, consider statistical significance testing using a Z-test. Although the specifics differ somewhat depending on the particular test selected, the general methodology is the same. The Z-test is based on the fact that the observed applicants are a sample (approaching a randomly and independently selected one[75]) of the universe of potential applicants. In conducting a Z-test, the statistician starts with the null hypothesis that there is no disparity between the selection rates of the two groups in the overall population. The alternate hypothesis is that *some* disparity exists. The researcher calculates the Z-score for the observed disparity.[76] That score yields a corresponding value reflecting the probability that the null hypothesis (of no disparity) cannot be rejected. If this probability is lower than a specified level (usually five percent), then the researcher can conclude that a disparity likely exists in the population. In sum, a statistically significant disparity exists when the observed difference between two groups varies from the expectation of no difference by a defined amount.

Researchers most commonly use the ninety-five percent confidence level, which is also termed the five percent (0.05) level of significance.[77] This level corresponds to a

[72] *See, e.g.*, Clark v. Pennsylvania, 885 F. Supp. 694, 707–08 (E.D. Pa. 1995); Reynolds v. Sheet Metal Workers Local 102, 498 F. Supp. 952, 966 (D.D.C. 1980), *aff'd*, 702 F.2d 221 (D.C. Cir. 1981).

[73] *See* WILLIAM L. HAYS, STATISTICS 267–82 (5th ed. 1994).

[74] *See* Shoben, *supra* note 59, at 798–800.

[75] *See id*. at 801. For a critique of these assumptions, see Meier et al., *supra* note 59.

[76] The Z-score is computed as follows, where SRT is the total selection rate, SR_{min} and SR_{maj} are the rates for the minority and majority groups respectively, and N1 and N2 are the number of applicants in each group:

$$Z_D = \frac{SR_{\min} - SR_{maj}}{\sqrt{SR_T(1 - SR_T)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}$$

See OFFICE OF FED. CONTRACT COMPLIANCE PROGRAMS, U.S. DEP'T OF LABOR, FEDERAL CONTRACT COMPLIANCE MANUAL app. 3A-1 (1993).

[77] *See* PAETZOLD ET AL., *supra* note 3, § 2.04, at 2–13.

Z-score of -1.96 or 1.96. At the ninety-five percent level, we can be ninety-five percent certain that the observed disparity in the applicant pool reflects a real disparity in the relevant labor market with respect to the challenged practice. There is still, however, a one in twenty possibility that there is no disparity in the overall population. It warrants emphasis that the Z-test, or any other statistical significance test, can only tell us that it is statistically unlikely that chance is responsible for a disparity, not that the challenged practice is causing the disparity.[78] As the Tenth Circuit explained, a statistically significant difference "strongly indicates *some* influence on the results other than the operation of pure chance,"[79] but that influence may be another factor that is correlated with the challenged practice.

Courts' reliance on statistical significance tests in disparate impact cases was prompted largely by the Supreme Court's holding in *Hazelwood School District v. United States*[80] that such tests were proper means of establishing discrimination in disparate treatment (intentional discrimination) cases. The *Hazelwood* Court described statistical significance testing as a "precise method of measuring the significance of . . . disparities."[81] The EEOC's explicit recognition of the value of statistical significance tests in its commentary establishing the four-fifths rule,[82] likely contributes to the use of such tests.

Statistical significance tests, however, are subject to the same criticism for arbitrariness as the four-fifths rule. There is no established level for statistical significance. The *Hazelwood* Court said that a disparity is statistically significant where it is more than two or three standard deviations from the expected rates,[83] but this standard is far from precise. Two standard deviations roughly corresponds to a ninety-five percent confidence interval, meaning that there is a five percent chance that the disparity is random, while three standard deviations corresponds to roughly a 99.75% confidence interval, meaning that there is a 0.25% chance that the disparity is random.[84] Although the most common significance level is five percent, nothing restricts courts to this level. There is considerable disagreement among statisticians and litigants on what level should be required before we can conclude a real difference

[78] Daniel L. Rubinfeld, *Reference Guide on Multiple Regression*, *in* REFERENCE MANUAL ON SCIENTIFIC EVIDENCE 183–84 (2d ed. 2000); *see also* Smith v. Xerox Corp., 196 F.3d 358, 366 (2d Cir. 1999), *overruled in part by* Meacham v. Knolls Atomic Power Lab., 461 F.3d 134, 141 (2d Cir. 2006); Browne, *supra* note 49, at 491 (criticizing statistical significance testing for yielding "meaningless results" because it convinces courts that plaintiffs have established a prima facie case of causation).

[79] Carpenter v. Boeing Co., 456 F.3d 1183, 1202 (10th Cir. 2006) (emphasis in original).

[80] 433 U.S. 299 (1977).

[81] *Id.* at 308 n.14.

[82] *See* Questions and Answers, *supra* note 60, at 11,998.

[83] *Hazelwood*, 433 U.S. at 311 n.17; *see also* Castaneda v. Partida, 430 U.S. 482, 496 n.17 (1977) (adopting a two or three standard deviation standard for a case involving racial discrimination in jury selection).

[84] Many courts present the results of their analysis in standard deviations from the expected mean rather than statistically significant at a specified level. But the two techniques are equivalent; the observed disparity is compared to the theoretical expected disparity of no difference. *See* THOMAS H. WONNACOTT & RONALD J. WONNACOTT, INTRODUCTORY STATISTICS FOR BUSINESS AND ECONOMICS § 9-1 (4th ed. 1990). For that reason, combining both types of analyses in this Part is appropriate.

exists in the population.[85] Further, individual litigants argue for whatever level leads to them prevailing.

Statistical significance is also tremendously sensitive to sample size. Simply put, the larger the number of applicants, the smaller the magnitude of difference that will be statistically significant (at whatever level is selected).[86] To illustrate, we return to our hypothetical large employer, where 3200 of 10,000 female applicants and 4000 of 10,000 male applicants met the hiring requirement. The disparity is not actionable under the four-fifths rule, but is actionable under statistical significance testing.[87] We can narrow the observed disparity significantly and still achieve statistical significance and thus potential liability: keeping the number of male applicants who meet the standard constant at 4000 (forty percent), if 3886 female applicants (38.86% of the total) or fewer meet the requirement, the disparity will be statistically significant at the ninety-five percent confidence level.[88] Thus, a net difference in the selection rates of 1.14 percentage points or greater is sufficient to establish disparate impact. In sum, whereas the four-fifths rule could be said to itself have a disparate impact on small employers, the statistical significance rule could be said to have a disparate impact on large employers because even a small disparity may achieve statistical significance.

In contrast, much larger disparities are necessary for statistical significance to be achieved in smaller samples.[89] For instance, in a court requiring statistical significance at the ninety-five confidence level, our hypothetical small employer will avoid liability unless the disparity in the hiring rates is at least twenty percent. Because small samples are likely in disparate impact cases, statistical significance tests can be criticized for being too harsh on plaintiffs and creating an unacceptable level of false negatives.

II. EVALUATING THE TWO TESTS

As we saw in *Isabel*, the consequence of having two different tests for disparate impact is that the choice of test often dictates the victor. This section begins by illustrating this problem mathematically. It then shows how the two tests work at cross-purposes and why courts should require plaintiffs to use a statistical significance test to

[85] *See* discussion *infra* text accompanying notes 101–104.

[86] *See* DAVID C. BALDUS & JAMES W.L. COLE, STATISTICAL PROOF OF DISCRIMINATION § 9.221, at 309–10 (1980).

[87] Indeed this disparity would be actionable up to the 99.9% confidence level.

[88] This number was calculated using a one-tailed test and z= -1.645. The equivalent value with a two-tailed test and z= -1.96 is 3864 female applicants.

[89] Recognizing that disparities are less likely to achieve statistical significance in small samples, employers often urge courts to disaggregate the plaintiffs' statistics by store, region, or some other grouping to decrease the likelihood a court will impose liability. *See, e.g.*, Segar v. Smith, 738 F.2d 1249, 1285–86 (D.C. Cir. 1984) (criticizing the defendant's disaggregation of the data in a disparate treatment case); Capaci v. Katz & Besthoff, Inc., 711 F.2d 647, 654 & n.4 (5th Cir. 1983) (calling disaggregation a "divide and conquer" technique because "[b]y fragmenting the data into small sample groups, the statistical tests become less probative."). Two commentators have suggested that disaggregation of data should only be justified where the employer can demonstrate that "the stratification is appropriate, and that the stratifying variable is business justified." PAETZOLD ET AL., *supra* note 3, § 5.08, at 35–36.

establish causation and a variation of the four-fifths rule to establish practical significance.

A. Different Results

By now it is hopefully clear that the four-fifths rule and statistical significance tests often lead to different results. There are two primary reasons for this. First, the four-fifths rule is highly sensitive to the magnitude of the selection rates being compared. Second, statistical significance tests are highly sensitive to the size of the sample being analyzed. Table 1 illustrates how disparities are actionable under the four-fifths rule and the *minimum* sample size necessary for that disparity to be statistically significant.[90]

Table 1. Actionable Disparities

Nonminority Selection Rate	Minority Selection Rate	Difference in Selection Rates	Minimum Sample Size Necessary for Statistical Significance
90%	72%	18	22
80%	64%	16	50
70%	56%	14	84
60%	48%	12	128
50%	40%	10	266
40%	32%	8	388
30%	24%	6	590
20%	16%	4	996
10%	8%	2	2214

First, we can observe that the differences actionable under the four-fifths rule fall within a rather wide range because the rule only examines the ratio of rates and does

[90] These numbers were calculated by using a one-tailed t-test with p=0.05 (ninety-five percent confidence level). The sample sizes assume that an equal number of persons are in each group, for example, eleven nonminorities and eleven whites for the first row to make up a sample of twenty-two applicants. For the first two rows, the sample sizes are approximate because the sample sizes are too small for the normal approximation to the binomial distribution to be valid.

Note that these sample sizes are actually artificially low for racial discrimination cases because they presume equal size groups. But if minorities are applying in a rate that matches their representation in the general population, of every ten applicants, fewer than three are minorities, see U.S. CENSUS BUREAU, TABLE DP-1. PROFILE OF GENERAL DEMOGRAPHIC CHARACTERISTICS: 2000 (2000), http://censtats.census.gov/data/US/01000.pdf (indicating that 75.1% of Americans self-reported as white), so much higher overall sample sizes are needed to have a sufficient number of minorities. The assumption of equal majority and minority group sizes is also problematic in sex discrimination cases involving sex segregated jobs (e.g., construction or nursing).

not consider the magnitude of the differences.[91] Liability under the rule is equally established when the selection ratio is ninety percent to seventy-two percent (a difference of eighteen percentage points) and when the selection ratio is ten percent to eight percent (a difference of two percentage points). In both instances, the selection ratio is eighty percent—although a court might choose to ignore the latter difference. Difference in the selection rates lower than the differences specified in table 1 are not actionable under the rule. Consider an extreme example: if the selection rate is ninety percent for whites and seventy-three percent for blacks (a difference of seventeen percentage points), reliance on the four-fifths rule would lead the court to conclude that plaintiffs had not shown disparate impact.

Courts who wish to impose their own normative views in disparate impact discrimination cases can take a lesson from this. If the court wants to side with the defendant, it will prefer the four-fifths rule where the selection rates at issue are high (because a significant disparity will not be actionable), but not where the selection rates are low. The opposite is true if the court favors the plaintiff.

Second, we can see that, as discussed in the previous Part, the number of employees in the sample matters. The minimum number of total employees required for a given disparity to be statistically significant increases dramatically as the magnitude of the observed disparity increases. A difference of eighteen percentage points is statistically significant in a sample as small as twenty-two employees, but a difference of two percentage points will not achieve statistical significance unless the company has at least 2214 employees. The result is that plaintiffs in small companies generally lose where the court favors statistical significance tests.

When we consider the problem of false negatives (findings of no discrimination when it actually exists), the outlook gets even more dire for employees in small companies. The sample sizes in table 1 represent the minimum sample size at which it is possible to find a statistically significant difference. At those levels, however, there is a rather significant likelihood of finding a false positive because the statistical power of the test is low—that is, the odds are low that we will observe a statistically significant disparity if one indeed exists in the population.[92]

Table 2 shows the sample sizes necessary to avoid false negatives at a statistical power level of eighty percent, the commonly accepted level for statistical power.[93] At this level, eighty percent of the time when there is an effect, the statistical test will conclude correctly that an effect exists; twenty percent of the time, it will conclude falsely that no effect exists. Table 2 confirms that large sample sizes are necessary to achieve statistical significance. As explained above, a company with a small number of employees is far more likely to face liability under the four-fifths rule than under a statistical significance test, while the opposite is true for a company with many employees.[94] Consequently, where the sample size is small, defendants will favor statistical significance and plaintiffs will favor the four-fifths rule. As the sample size increases, that preference flips.

[91] *See* Shoben, *supra* note 59, at 810–11.

[92] For an overview of statistical power analysis, see JACOB COHEN, STATISTICAL POWER ANALYSIS FOR THE BEHAVIORAL SCIENCES 1–17 (2d ed. 1988).

[93] KEVIN R. MURPHY & BRETT MYORS, STATISTICAL POWER ANALYSIS: A SIMPLE AND GENERAL MODEL FOR TRADITIONAL AND MODERN HYPOTHESIS TESTS 18 (2004).

[94] *See supra* text accompanying notes 87–90.

Table 2. Minimum Sample Sizes Needed to Avoid False Positives and False Negatives

Nonminority Selection Rate	Minority Selection Rate	Minimum Sample Necessary to Achieve Statistical Significance and Avoid False Positives	Minimum Number Necessary to Achieve 80% Statistical Power and Avoid False Negatives
90%	72%	22	70
80%	64%	50	148
70%	56%	84	246
60%	48%	128	372
50%	40%	266	540
40%	32%	388	774
30%	24%	590	1128
20%	16%	996	1716
10%	8%	2214	2890

B. Different Functions

Given all the criticism of the current tests, we might be tempted to discard the tests entirely and instead defer to judges' subjective views of whether liability is warranted. But we need not adopt some sort of "I know it when I see it" test for disparate impact.[95] The reason the tests are problematic when used individually is that they only answer one part of the relevant inquiry under the doctrine. This inquiry consists of two sub-inquiries: (1) does the practice cause a disparity?; and (2) is that disparity large enough to matter? The first inquiry involves establishing statistical significance, and the second inquiry involves establishing what statisticians term practical significance—the subjective determination of the size of a disparity where "it is sufficiently important substantively for the court to be concerned."[96]

Neither of the two tests can answer both of the relevant inquiries, but they actually each work quite well at answering one. Statistical significance answers the first inquiry—it tells us the mathematical certainty with which we can know that the relationship is not due to chance. The four-fifths rule answers the second inquiry—it

[95] *Cf.* Jacobellis v. Ohio, 378 U.S. 184, 197 (1964) (Stewart, J., concurring) (adopting this test for obscenity cases). There is some appeal to such an approach, particularly among those who view disparate impact as a way of identifying intentional discrimination, because it allows judges to impose liability on bad defendants. But the approach is unsatisfying because it is standardless and provides no guidance to employers seeking to avoid liability.

[96] Rubinfeld, *supra* note 78, at 179, 191–92. The EEOC actually suggests "practical significance" as a potential third alternative to statistical significance tests and the four-fifths rule "under which the court evaluates whether findings of statistical significance are 'practically' sound, rather than just 'barely significant.'" Questions and Answers, *supra* note 60. The suggestion of practical significance as a third alternative overlooks the fact that the four-fifths rule already provides a means at assessing whether findings are practically sound and creates the potential for a relative free-for-all among litigants and courts in defining disparate impact, but my review of the case law indicates that this fortunately has not yet occurred.

tells us whether a disparity is sufficiently large to warrant imposing liability. Rather than using one test to perform both inquiries, courts and litigants should use the two tests together such that each test can perform the function to which it is best suited.

1. Evaluating Causation

Statistical significance tests allow courts to determine whether they can be sufficiently confident (at whatever level specified) that the observed disparity reflects a real disparity in the relevant labor market such that they can infer that the challenged practice probably *caused* the disparity—the showing required under the statute. Plaintiffs' statistics will generally show a disparity in the selection rates for each group. Absent statistical analysis, however, it is uncertain whether this disparity is a fluke in the sample selected or is indicative of an overall disparity in the population.

In contrast, the four-fifths rule does not aid courts in evaluating causation.[97] The rule instead attempts to sidestep the causation problem by creating a rather high threshold (the four-fifths ratio) necessary to establish disparate impact in order to provide for the possibility that other factors are causing the disparity. But this at most indirectly evaluates causation and results in a significant false negatives problem.

2. Evaluating Practical Significance

Although the four-fifths rule is ill-suited for analyzing causation, it is well-suited for aiding courts in determining whether a disparity is sufficiently large to matter—that is, whether it has practical significance.[98] Indeed, the four-fifths rule is best understood as a policy determination by the EEOC of the level at which courts should get involved. As the Fourth Circuit has said, "It offers a definition of what is a serious difference in the passing rates for protected classes."[99]

Statistical significance tests, in contrast, provide no insight into whether a disparity is practically significant. In fact, the inclusion of "significance" is a bit of a misnomer because we generally understand "significant" to mean "important," while in this context "significant" means "probably true." A finding can be probably true without being important. As shown above, statistical significance often indicates a difference of only a few percentage points.[100]

3. Furthering the Doctrine

The different functions of the two tests are important not just mathematically, but doctrinally.

[97] *See* Rich v. Martin Marietta Corp., 467 F. Supp. 587, 612 (D. Colo. 1979) ("[W]hile it may supply an inference of discrimination, the inference of the 4/5's Rule (resting, as it does, upon untested intuition) is not a strong one.").

[98] *See* Meier et al., *supra* note 59, at 163–64.

[99] Chisholm v. U.S. Postal Serv., 665 F.2d 482, 495 n.22 (4th Cir. 1981).

[100] *See supra* text accompanying note 87–89.

The first function, establishing causation, relates to the smoking-out-discrimination view of the disparate impact doctrine, which posits that employers adopt practices with the intent of causing disparate impact. If this view is correct, then defendant-employers know or should know that the challenged practice is causing the disparity. Establishing causation through statistical significance tests thus suggests that the observed results were intended by the employer, not due to chance.

The second function, establishing practical significance, relates to the promoting equality view of the doctrine, which aims to eliminate disparities in the workforce regardless of why those disparities exist. Its concern is not causation but results. Although proponents of this view seek to eliminate all disparities, no matter how small, establishing practical significance helps target resources to the elimination of significant disparities.

C. Using the Tests Together

Because the two tests perform complementary functions, both mathematically and doctrinally, courts should use them together. To establish a prima facie case, the plaintiff should be required to show that the observed disparity is statistically significant (i.e., likely caused the observed disparity) and practically significant at the level established by Congress or the EEOC. If the plaintiff makes both of these showings, the court should be *required* to conclude that the plaintiff has established a prima facie case, thus shifting the burden to the defendant to rebut that case.

Although it is certainly possible to evaluate a disparate impact case without using the two mathematical tests together, doing so promotes transparency and predictability both in court and before litigation is brought. Indeed, a mathematical test is necessary to assess statistical significance; adequately eyeballing the result to make a conclusion is not acceptable. Although practical significance arguably can be eyeballed, a mathematical test reduces the impact of whose eyes are doing the assessing by developing a uniform standard, which is not affected by an individual litigant's or judge's sympathies, preferred outcome, or past experiences.

My argument that we should require plaintiffs to meet some version of both tests should not be read as an argument for increasing the burdens on plaintiffs, although it would certainly lead to that result should courts apply both tests in their current forms. Requiring plaintiffs to meet the four-fifths rule and establish statistical significance at the five percent level would make it almost impossible for plaintiffs to prevail, thus undermining the goals of antidiscrimination statutes. Congress or the EEOC instead should adopt modified versions of both tests with reduced burdens on plaintiffs. Under the combined tests, some plaintiffs would fare better and some would fare worse, but the new standard would more accurately measure whether a disparity is statistically and practically significant.

To be sure, the argument that courts should use both tests rightly could be criticized for being too harsh on plaintiffs applying to or working for small employers. It certainly will be somewhat difficult for such plaintiffs to establish a disparate impact regardless of what level of statistical significance is applied. This difficulty, however, should not lead courts to ignore the clear statutory requirement that the plaintiff establish causation.[101] We might make a policy determination to relax that causation showing and indeed might even consider different standards for small employers, particularly if we had some reason to believe that disparities were more problematic when they occurred in small employers. But courts should not ignore the statutory requirement of causation because the EEOC has made a policy determination that a ratio of less than eighty percent is problematic.

Further, plaintiffs have other ways of establishing causation—just as they do under the current doctrine. They can argue for aggregation of plaintiffs across workplaces either within the same employer (e.g., different factories) or across different employers (e.g., automobile factories in Detroit). Plaintiffs could also point to evidence that the same or similar test being used elsewhere also had a disparate impact. Furthermore, a finding of causation in one case may be persuasive in a subsequent case.

III. IDENTIFYING THE APPROPRIATE TEST

In recognizing the value of both statistical significance testing and the four-fifths rule, we should not feel constrained to accept the rules as they currently stand—particularly because that would be highly onerous for plaintiffs. Instead, we should begin a normative debate about how to weigh the two types of potential errors—false positives and false negatives—and what magnitude of disparity is sufficient to be actionable. In choosing the appropriate test, we decide how much to tip the balance in favor of plaintiffs or defendants.

A. Choosing the Statistical Significance Level

Selecting a statistical significance level involves an assessment of how to weigh the likelihood of false negatives and false positives. Because we can never be certain whether causation exists, we have to decide how much uncertainty we are comfortable with and who should bear the risks of that uncertainty.[102] Determining the appropriate level for statistical significance is in part a policy decision of how comfortable we are with letting discriminators off as compared to punishing nondiscriminating employers.

In evaluating the considerations for choosing the appropriate level, it might be helpful to start with a more familiar example—William Blackstone's oft-repeated maxim that it is "better that ten guilty persons escape, than that one innocent suffer."[103] Blackstone never stated a justification for this number; instead he relied on an intuitive sense of right and wrong. But with this standard comes an associated risk of false positives and false negatives. The implication of the ten guilty men standard is that we want to be approximately ninety-one percent certain before we convict. We might prefer a "three guilty men" standard—if we are seventy-five percent certain a criminal is guilty, then we convict. If we want to be ninety-five percent certain (the most

[101] *See* 42 U.S.C. § 2000e-2(k)(1)(A) (2000).

[102] Allocating the risks of uncertainty is likewise central to tort law. Courts balance the expected risks and benefits of finding liability in the face of uncertainty and consider the equitable treatment of the parties and other normative considerations (e.g., utility, fairness, and economic efficiency). *See generally* ARIEL PORAT & ALEX STEIN, TORT LIABILITY UNDER UNCERTAINTY 16–56 (2001) (evaluating various decision rules for allocating uncertainty).

[103] WILLIAM BLACKSTONE, 4 COMMENTARIES *358.

frequently used level for statistical significance), then we must adopt a "nineteen guilty men standard."[104]

Ultimately, statistical significance is an "arbitrary convention."[105] As amicus American Psychological Association observed in a disparate impact case,

> Under some circumstances . . . one might need a prediction so badly or have such severely restricted range of scores on prediction or criterion that he will decide to accept any finding significant at the 10% level; under other circumstances, the cost of using a test that might later prove invalid may be so great that the investigator will insist on at least a 1% level.[106]

One's predilections dictate his preferred level.

Some will take the position that we should never be truly satisfied as long as one company is able to discriminate (i.e., use a practice that causes a disparity); they want to avoid all false negatives. Some number might even believe that every observed disparity is conclusive evidence of discriminatory intent. Others might take the position that we should never be satisfied as long as one nondiscriminating company faces liability; they want to avoid all false positives.

But such absolute standards are impossible because judges and juries cannot determine causation conclusively. Even if we set the confidence level at 50.1% (and thus created a 49.9% chance of false positives), such that all that was required was that it was (slightly) more likely than not that an observed disparity was not due to chance, we would still have a problem of false negatives.[107] If we set the confidence level at 99.9% (a level common in the social sciences),[108] there would still be a 1-in-1000 chance that an observed disparity was a false positive in addition to a rather significant problem of false negatives. Moreover, adopting either extreme ignores the fact that under the statute both false positives and false negatives are problematic.

[104] For a historical discussion of opinions on the "correct" aspirational ratio, see Alexander Volokh, n *Guilty Men*, 146 U. PA. L. REV. 173 (1997). The Blackstone parallel is admittedly somewhat awkward because in the criminal context, the jury is trying to determine an objective fact (e.g., did the defendant kill another man?). In the discrimination context, there is no objective fact to discover. Instead, the tests themselves define what constitutes discrimination. If the disparity is sufficiently large, then the challenged practice discriminatory.

[105] United States v. Georgia Power Co., 474 F.2d 906, 915 n.11 (5th Cir. 1973); *see, e.g.*, Neil B. Cohen, *Confidence in Probability: Burdens of Persuasion in a World of Imperfect Knowledge*, 60 N.Y.U. L. REV. 385, 412 (1985) ("This convention reflects nothing more than an arbitrary balancing of the disutilities, or 'regrets,' of [false positives and false negatives].").

[106] *Georgia Power Co.*, 474 F.2 at 915 n.11.

[107] The extent of that problem would depend on the statistical power, which is determined by the sample size and the relevant selection rates.

[108] *See, e.g.*, Yung-An Hu & Day-Yang Liu, *Altruism Versus Egoism in Human Behavior of Mixed Motives: An Experimental Study*, 62 AM. J. ECON. & SOC. 677 (2003) (using 0.001 level of significance); Craig Volden, *States as Policy Laboratories: Emulating Success in the Children's Health Insurance Program*, 50 AM. J. POL. SCI. 294 (2006) (same); Kemal Yildirim, Aysu Akalin-Baskaya & Mine Celebi, *The Effects of Window Proximity, Partition Height, and Gender on Perceptions of Open-Plan Offices*, 27 J. ENVTL. PSYCHOL. 154 (2007) (same).

In determining the appropriate statistical significance level, we should use the current standard, the ninety-five percent confidence level, as the starting point. We must then consider the benefits and the costs of lowering or raising that level.

1. The Case for Requiring a Lower Level

There is good reason to believe that the current level for statistical significance is too high because it establishes too great a risk of false negatives in order to minimize the risk of false positives. As Professor Neil Cohen argues, "Although this conservative balancing of risks may be appropriate for deciding when to accept a scientific hypothesis, it is not necessarily appropriate within the legal context."[109] Cohen estimates that the risk of a false negative at the ninety-five percent level is as high as fifty percent.[110] A high level of certainty (and the concomitant high risk of false negatives) is necessary for scientific testing. For example, a company would not tout a new drug for treatment of a disease until it was rather certain that drug worked. Such a high level of certainty is less important, however, in disparate impact cases where statistical significance is only the first step of the analysis because defendants have ample opportunity to avoid liability through the rest of the proof structure by justifying their practices or rebutting the plaintiff's statistics.

Readers who believe it is okay to disregard a greater risk of error than five percent will favor a lower statistical significance level. There is extensive support in the statistics literature for a ninety percent level.[111] And legal scholars regularly report effects at a ninety percent confidence level as statistically significant.[112] Lowering the level required for statistical significance would also bring disparate impact litigation more in line with the preponderance of the evidence, or "more likely than not," standard frequent in civil litigation.[113] John Kaplan has identified a probability of 0.5

[109] Cohen, *supra* note 105, at 412.

[110] *See id*. at 413.

[111] *See, e.g.*, THOMAS R. DYCKMAN & L. JOSEPH THOMAS, FUNDAMENTAL STATISTICS FOR BUSINESS AND ECONOMICS 342–47 (1977) (regarding ten percent chance of error as acceptable); WILLIAM MENDENHALL & JAMES E. REINMUTH, STATISTICS FOR MANAGEMENT AND ECONOMICS 263–64 & tbl. 7.1 (4th ed. 1982) (showing the ninety percent, ninety-five percent, and ninety-nine percent significance levels as valid alternatives, the choice of which depends on "the degree of confidence the [researcher] wishes to place in the estimate"). *But cf.* Flue-Cured Tobacco Coop. Stabilization Corp. v. EPA, 4 F. Supp. 2d 435, 461 (M.D.N.C. 1998) (criticizing the EPA for changing the significance level from ninety-five percent to ninety percent because that "looks like a[n] attempt to achieve statistical significance for a result which otherwise would not achieve significance." (alteration in original) (quotation omitted)), *rev'd*, 313 F.3d 852 (4th Cir. 2002).

[112] *See* Rachel E. Barkow & Kathleen M. O'Neill, *Delegating Punitive Power: The Political Economy of Sentencing Commission and Guideline Formation*, 84 TEX. L. REV. 1973, 2006–2007 & tbl.2 (2006); John C. Coates IV & Guhan Subramanian, *A Buy-Side Model of M&A Lockups: Theory and Evidence*, 53 STAN. L. REV. 307, 369 (2000); Mary Eschelbach Hansen & Daniel Pollack, *Unintended Consequences of Bargaining for Adoption Assistance Payments*, 43 FAM. CT. REV. 494, 505 tbl. 5 (2005); Beth A. Simmons, *Money and the Law: Why Comply with the Public International Law of Money?*, 25 YALE J. INT'L L. 323, 348 n.102 (2000).

[113] *See* 2 CHARLES T. MCCORMICK, MCCORMICK ON EVIDENCE § 339, at 568 (Kenneth S.

as the optimal decision point for civil cases generally because "the consequences of an error in one direction are just as serious as the consequences of an error in the other."[114] Although it might be hard for most to accept this level for disparate impact cases, we should ask ourselves why a plaintiff in a torts case can prevail by showing a likelihood of fifty-one percent that the defendant is responsible for his injury but a much higher level is necessary merely to shift the burden in disparate impact cases.

Further, a lower level would decrease problems associated with evaluating statistical significance in small samples. Because most employers have a relatively small number of employees,[115] requiring plaintiffs to establish significance at the ninety-five percent level makes it difficult to enforce the mandates of antidiscrimination statutes. A lower level would increase plaintiffs' likelihood of bringing successful cases.

Although changing the level would require more defendants to defend their practices, this increased burden on defendants might be warranted given that the standard for business necessity is rather deferential to employers.[116] Indeed, many courts seem willing to accept the employer's word that a challenged practice is necessary. For instance, an Illinois court concluded that a background check was a business necessity for police officers even though the police department presented no evidence supporting the need for the requirement.[117]

2. The Case for Requiring a Higher Level

There are countervailing reasons, however, to increase the level required to establish statistical significance or at least to keep it at the current level—ninety-five percent. The higher the level, the greater the certainty that random chance is not producing the observed disparity. According to the widely quoted "reasonable interpretation of the results of significance tests" there are four confidence levels: (1) a confidence level of ninety-nine percent or greater is required to indicate "very strong evidence" of a difference; (2) levels between ninety-five percent and ninety-nine percent provide "moderate evidence;" (3) levels between ninety and ninety-five percent provide "suggestive evidence;" and (4) anything below those levels yields "little or no real evidence" of difference.[118] We need look no further than *Hazelwood* to find support for a 99.75% level, the level corresponding to the Court's "three standard deviations" language.[119]

Broun ed., 6th ed. 2006).

[114] John Kaplan, *Decision Theory and the Factfinding Process*, 20 STAN. L. REV. 1065, 1072 (1968).

[115] According to census data, of the approximately 5.9 million employer firms, about 4.8 million (eighty-two percent) have fewer than 100 employees and about 2.8 million (forty-seven percent) have fewer than twenty employees. *See* U.S. CENSUS BUREAU, STATISTICS ABOUT BUSINESS SIZE (2004), http://www.census.gov/epcd/www/smallbus.html.

[116] *See* N. Y. City Transit Auth. v. Beazer, 440 U.S. 568, 592 (1979) (suggesting that it is sufficient for the employer to show that the policy served "the general objectives of safety and efficiency").

[117] *See* McCraven v. City of Chicago, 109 F. Supp. 2d 935, 945 (N.D. Ill. 2000) (concluding that the plaintiffs had failed to disprove business necessity).

[118] WALTER BURDETTE & EDMUND GEHAN, PLANNING AND ANALYSIS OF CLINICAL STUDIES 9 (1970).

[119] *See* Hazelwood Sch. Dist. v. United States, 433 U.S. 299, 311 n.17 (1977).

A higher level would reduce the doctrine's interference with employers' freedom to control their own businesses. As Professors Theodore Blumoff and Harold Lewis have observed, the disparate impact doctrine reflects an effort to "preserve the defendant's freedom of contract rather than commit the 'false positive' error that favors the victim."[120] As far back as 1978, the Supreme Court observed in a Title VII case that "courts are generally less competent than employers to restructure business practices."[121] Those who believe the current requirements are already too onerous for employers should support a higher level.[122] Proponents of a higher level likely worry that a lower level or the current level: (1) wastes money and resources by requiring employers to gather evidence of their practices' effectiveness and defend those practices in courts; and (2) forces some employers to abandon their preferred practices rather than expose themselves to potential liability.

Further, we should prefer a higher level of confidence if we are concerned that the current level exposes firms to an unacceptable degree of risk. The easier we make it for plaintiffs to establish a prima facie case, the greater the likelihood that enterprising plaintiffs will bring cases with little or no merit and coerce companies into providing lucrative settlements.[123] The risk of distortion effects is arguably lower in disparate impact cases than in other types of litigation because prevailing plaintiffs have limited remedies and cannot receive punitive damages. Nonetheless, the potential of encouraging inefficient litigation should caution us against supporting too low a level.

3. Preferred Approach

In my view, a ninety percent statistical significance level is preferable. At this level, we can be ninety percent confident that an observed disparity is not due to chance. The ninety percent level appropriately reflects the need for a robust test for statistical significance, as ninety percent is used frequently in research literature, while also recognizing the difficulties of establishing statistical significance in small samples.

A ninety percent level might seem too low to some because it is of accepted levels of statistical significance. But choosing such a level on the outer bounds makes sense in this context because the test does not establish liability, but only establishes part of the plaintiff's prima facie case. The doctrine's burden-shifting framework establishes a "built in means of limiting errors."[124] A ninety percent level might seem too high to

[120] Theodore Y. Blumoff & Harold S. Lewis, Jr., *The Reagan Court and Title VII: A Common-Law Outlook on a Statutory Task*, 69 N.C. L. REV. 1, 76–79 (1990).

[121] Furnco Constr. Co. v. Waters, 438 U.S. 567, 578 (1978); *see also* Deborah C. Malamud, *The Last Minuet: Disparate Treatment After* Hicks, 93 MICH. L. REV. 2229, 2265 (arguing that the Court is unwilling to use the judicial system to "require the employer to restructure his employment practices to maximize the number of minorities and women hired" (quoting Texas Dep't of Cmty. Affairs v. Burdine, 450 U.S. 248, 259 (1981))).

[122] If we choose a lower confidence level, we might choose to decrease the burden on employers to establish business necessity, so that the change is less onerous for employers.

[123] *See generally* John C. Coffee, Jr., *The Regulation of Entrepreneurial Litigation: Balancing Fairness and Efficiency in Large Class Actions*, 54 U. CHI. L. REV. 877 (1987) (documenting plaintiffs and the plaintiffs' bar bringing nonmeritorious cases for financial benefit and thereby exposing firms to unnecessary risk).

[124] David A. Strauss, *The Law and Economics of Racial Discrimination in Employment:*

some because at such a level, it will still be difficult for plaintiffs in cases with small samples to establish a statistically significant disparity. This difficulty, however, will exist to some extent at whatever level is chosen. Aggregation—across years, different offices, or industries—presents a partial, albeit somewhat unsatisfying, solution.[125] The ninety percent level reduces the difficulty of establishing causation in small samples while fulfilling the need for sufficient certainty with respect to the conclusion that a practice caused the disparity.

B. Choosing the Practical Significance Level

In evaluating practical significance we determine at what point a (statistically significant) disparity is large enough to impose liability. The EEOC guidelines already direct courts to analyze practical significance,[126] a decision likely aimed at directing judicial and agency enforcement resources toward disparities with practical significance.

Some readers invariably believe that any disparity, however small, should be sufficient to establish disparate impact. To such readers, establishing a minimum level potentially contravenes the clear purpose of Title VII to identify and rectify discrimination.[127] The problem with this view is that it equates the mere showing of a difference with the showing of discrimination. Practical significance is a policy determination of the point at which an observed, nonrandom difference becomes actionable discrimination. Even assuming (erroneously in my view) that all practices that cause disparities are injuries, we regularly decide that there are some injuries for which the law should be concerned and some for which it should not.[128] We must balance the benefits and goals of Title VII and similar statutes with the costs of holding defendants liable for disparate impact.[129]

Further, requiring more than a small disparity is consistent with the two views of the doctrine. Under the smoking-out-discrimination view of the doctrine, we should require a greater showing of animus before labeling an employer as a "discriminator." It is unlikely that an intentional discriminator would impose a practice with minimal disparate impact (unless he or she intended to discriminate to the maximum extent permitted under the law). Under the promoting equality view of the doctrine, practical significance is a means of targeting resources; challenging practices with little

The Case for Numerical Standards, 79 GEO. L.J. 1619, 1650 (1991).

[125] *See* Coates v. Johnson & Johnson, 756 F.2d 524, 541 (7th Cir. 1985) ("Pooling data is sometimes not only appropriate but necessary, since statistical significance becomes harder to attain as the sample size shrinks.").

[126] *See* Questions and Answers, *supra* note 60, at 11,999–12,000.

[127] McDonnell Douglas Corp. v. Green, 411 U.S. 792, 801 (1973) (stating that "Title VII tolerates *no* discrimination, subtle or otherwise") (emphasis added).

[128] For instance, the Supreme Court has held that "The Eighth Amendment's prohibition of 'cruel and unusual' punishments necessarily excludes from constitutional recognition de minimis uses of physical force." Hudson v. McMillian, 503 U.S. 1, 9–10 (1992) (emphasis in original). The plaintiff must show that "the alleged wrongdoing is objectively 'harmful enough' to establish a constitutional violation." *Id.* at 2 (internal citation omitted).

[129] The remedies for a disparate impact case include reinstatement or promotion of plaintiff class members and discontinuation of the challenged practice. 42 U.S.C. § 2000e-5(g)(1) (2000). Defendants may face high costs in creating jobs for the plaintiffs and developing new selection practices.

noticeable impact has a relatively unnoticeable impact on establishing a diverse workforce.

In establishing a uniform level for practical significance, there are three basic options: First, we could establish a selection ratio at which the disparity is actionable, as the four-fifths rule currently does. Second, we could require a difference of a specified number of percentage points in the selection rates. Third, we could require a difference of a specified number of persons. Of course, alternatively, we could just direct courts to require more than a de minimis disparity, but that risks the same arbitrariness that characterizes current case law.

1. Selection Ratio

The first option for establishing practical significance requires plaintiffs to show that the minority group's selection rate is less than a given ratio, or percentage, of the majority group's rate. This is how the four-fifths rule currently functions—it establishes four-fifths, or eighty percent, as the level at which a disparity is actionable. If both tests are combined into one standard, then four-fifths is too low a level because there is no need to leave room for other factors or chance to operate, as a means of addressing causation.

The problem with the ratio option, as discussed above, is that it is sensitive to the magnitude of the selection rates being analyzed.[130] Whatever the selected ratio, the lower the overall selection rates are, the easier it is for plaintiffs to establish liability. For instance, if we choose ninety percent as the selection ratio, practical significance is established where the selection rates are forty-five percent and fifty percent and where the rates are nine percent and ten percent; however, we will likely be more concerned with the first disparity than the second.[131] One way to minimize this problem is to use the fail ratio instead of the pass ratio when the selection rates are very low.[132] In the above example, a pass ratio of nine percent to ten percent corresponds to a fail ratio of ninety-one percent to ninety percent, which is ninety-four percent—above the ninety percent threshold for liability. Defining "very low" might be problematic, thus leading to arbitrary determinations.

2. Difference in Selection Rates

Alternatively, we could require plaintiffs to show that the difference was equal to, or greater than, a specified number of percentage points in the selection rate to establish practical significance. For instance, we could designate a difference of two percentage

[130] *See supra* text accompanying notes 92–94.

[131] One might wonder whether such low selection rates actually exist in the workplace. Although such rates are probably unlikely where employers use written tests, low rates are likely where employers use subjective selection systems such as interviews in which few persons are ultimately chosen for hiring or promotion.

[132] Currently, the fail ratio provides another option to courts, increasing the arbitrariness of the tests. "[D]epending on whether the court focuses on pass rates, fail rates, or minority applicants versus hirees, the apparent disparity can be minimized or maximized." Dean Booth & James L. Mackay, *Legal Constraints on Employment Testing and Evolving Trends in the Law*, 29 EMORY L.J. 121, 154 (1980).

points as the required level. Under this standard, both a 48/50 selection rate split and an 8/10 split would be actionable. This rule establishes a uniform threshold for liability. But readers who care about the magnitude will likely think that the second disparity matters more—indeed the first split reflects a four percent disparity, while the second represents a disparity of five times that, or twenty percent. Such readers will prefer the selection ratio method outlined above.

3. Difference in Number of Persons

We could also require that a disparity be equal to, or greater than, a specified number of persons to establish statistical significance. If the addition of a certain number of persons to the pass group of the minority eliminates any statistically significant disparity, then the disparity is *not* practically significant. To illustrate, consider selecting five as the relevant number. If a statistically significant disparity between the performance of blacks and whites on a written test remains statistically significant after adding five more blacks to the pass group, practical significance is established. If the difference disappears after the addition, then the difference is not practically significant.

The problem with this standard, however, should be rather obvious: it is sensitive to the magnitude of sample sizes being analyzed. Certainly, the movement of five persons is going to have a much greater effect in a small workforce than in a large one. In a twenty-person workforce, five persons constitute twenty-five percent of the total workforce, while in a 1000-person workforce, those five persons only make up 0.5% of the total workforce.

4. Different Standards

Once we select the standard for practical significance, we might then consider the possibility of creating different standards for different employers. For instance, we could establish a graduated standard system that rewards "good" behaviors (e.g., recruiting at minority job fairs). Indeed, the EEOC regulations currently suggest that a less stringent rule might apply to an employer who conducted an extensive recruiting campaign and developed a larger pool of minority and female applicants and a more stringent rule might apply where an employer's reputation might have discouraged members of particular groups from applying.[133] We also could use different standards to treat large and small employers differently, or to accomplish other policy goals.

Different standards based on employers' behavior are probably most appealing to adherents of the smoking-out-discrimination view of the disparate impact doctrine. Presumably, the "bad" employers the view seeks to smoke out are less likely to engage in behaviors that promote a more diverse workforce. In contrast, under the promoting equality view, different standards are problematic because they reward effort rather than results and allow "good" employers to have less diverse workforces.

[133] *See* 29 C.F.R. § 1607.4(D) (2008).

5. Preferred Result

A difference in the selection-rates standard for practical significance is preferable, in my view, because disparities with the same absolute magnitude should matter the same amount. A selection-rate disparity of ten percent versus fifteen percent is equally practically significant as a selection-rate disparity of twenty percent versus twenty-five percent or a disparity of thirty percent versus thirty-five percent, even though the underlying percent differences are different. A ratio approach would make sense if we were measuring change and thus the starting point was important. But in disparate impact cases, we care about the difference in rates, not how those rates are related proportionally. Further, using the difference rather than the ratio avoids the difficulties of evaluating ratios where selection rates are low.[134]

Whatever standard is selected, in my view, the standard should be uniform across employers and cases. Although the impulse behind rewarding "good" behaviors through the standard is understandable, any differentiation of standards risks the same potential for arbitrary decision making that characterizes the current choice of two tests. Further, measuring and evaluating "good" behavior is both difficult and highly subjective. How does a judge determine, for instance, how much recruiting constitutes an "extensive" recruiting campaign?[135]

From my research and review of numerous disparate impact cases, my sense is that courts may have already chosen five as the baseline required for practical significance. Under that level, if one group's selection rate was fifty percent, the other group's rate would have to be forty-five percent or lower for the disparity to be actionable. In addition to its frequent use, five makes intuitive sense because it is large enough to put employers on notice of potential liability and to weed out smaller disparities, thus causing employers to target resources at disparities with greater effect. Therefore, five is a good starting point when thinking about the appropriate standard.

In determining the standard, however, more research is necessary. Ultimately this choice is a policy decision, but that decision should be informed by information about the range of observed disparities and the shape of that range. For instance, it would be helpful to have a rough estimate of the incidence of disparities and the average size of a disparity (i.e., in a sample of 100 employers, how many use employment practices resulting in disparities, and how large are these disparities?). We might also seek information on the magnitude of observed disparities in disparate impact claims filed (e.g., is there some observed minimum point at which plaintiffs are bringing claims?). More research about how employers' practices are affecting workplace diversity would help us to determine how to best target judicial and employer resources at practically significant disparities.

[134] *See supra* text accompanying note 86–90.

[135] Evidence of good behavior might be relevant in determining whether applicants represent the relevant labor market. *See* Reynolds v. Sheet Metal Workers Local 102, 498 F. Supp. 952, 967 (D.D.C. 1980) (using labor market instead of applicant data for analysis because employers requirements "dissuaded or 'chilled' potential applicants").

CONCLUSION

This Article is certainly not intended to end the debate about the appropriate test but to start it. The Article recognizes the need for using both tests as complements and the necessary considerations in choosing an appropriate level for those tests. Choosing the proper standard is necessarily a policy-laden decision that must balance the need to make it feasible for plaintiffs to make out prima facie cases of disparate impact discrimination with the desire not to impose unfairly on employers' control of hiring and promotion.

Whatever test is chosen, it is important that it be uniform. Otherwise, individual litigants and judges will continue to choose the appropriate test on an ad hoc basis as a means of dictating their desired result. The way to distinguish statistics from the "lies and damned lies" in disparate impact cases is to select a single coherent test, while recognizing the normative considerations implicit in its selection.