An analysis of the properties of a potential replacement of the four-fifths rule

Elizabeth Jaikes[1]
Department of Economics
A.E. Rodriguez

Abstract

How does a potential plaintiff account for the increasingly common manipulation of a workforce event? The professionalization of Title VII-related advice on workforce interventions enables firms contemplating such a move to manage the results to comply with received case-law and enforcement agency regulations. The objective of the firm's tinkering, reduced to its essence, is to forestall litigation by gerrymandering favorable statistical tests of significance to achieve a seemingly facially neutral employment outcome. Presumably, the favorable gender-ratios or race-ratios resulting from the planned process will pre-empt litigation or, at the very least, dramatically reduce its chances. After all, in practically all forums, plaintiff's rebuttable presumption in disparate impact and disparate treatment cases is seemingly established by a statistical showing of outcomes which can be, and is, artfully altered. We propose a more ample interpretation of the EEOC's rule of thumb – the four-fifth's rule in a manner that will provide relief to aggrieved plaintiffs. Specifically, in this paper we appraise the incremental impact on the Type I and Type II error rates of raising the threshold ratio to establish a rebuttable presumption of discrimination. We examine a threshold of 0.9 for purposes of illustration.

Introduction

Any modern firm entertaining a reduction-in-force, a job-promotion program or any similar workforce selection event is likely to carefully consider the possible abridgment of the various Title VII laws[2]. In this role it is increasingly assisted by professional workforce event managers[3].

Professional input is advisable and solicited because stakes are high for the firm. Monetary losses derived from an unfavorable court decision on an adverse impact matter can be substantial whether a company prevails or chooses to settle. Indirect costs may similarly be considerable. Few firms can stomach the negative publicity that often accompanies popular media coverage of legal proceedings involving discrimination – and most would avoid it altogether, given the choice.

The professionalization of Title VII-related advice on workforce interventions enables firms contemplating such a move to manage the results to comply with received case-law and enforcement agency regulations.

The objective, reduced to its essence, is to forestall litigation by gerrymandering favorable statistical tests of significance to achieve a seemingly facially neutral employment outcome. Presumably, the favorable gender-ratios or race-ratios embodied in the perquisite statistical nomenclature resulting from the planned process will pre-empt litigation or, at the very least, dramatically reduce its chances. After all, in practically all forums, plaintiff's rebuttable presumption in disparate impact and disparate treatment cases is seemingly established by a statistical showing of outcomes[4].

Firm's internal human resources specialists, legal counsel, both in-house and external, as well as specialized consulting firms are fully aware of the potential legal downsides of an "unplanned" employment event (York 2002). In fact, these planning services are now an integral part of the portfolio of services peddled by law firms and professional consulting firms.

For example, K&L Gates' *Labor and Employment Alert*, December 10, 2008 in a piece by
Michael Pavlick and George Barbatsuly titled "To Test or Not to Test: Statistical Analysis for Small Layoffs?" note the following: "As a general guideline, layoffs of fewer than 100 employees are ripe for analysis using an exact test like Fisher's Exact Test. When conducting such a test, the employer should retain legal counsel to facilitate the testing. Experienced counsel can not only identify the right questions to ask, but in the event that the statistical analysis reveals a disparate impact, subsequent reconsideration of the reduction in force will enjoy some level of attorney-client privilege protection[5]."

The well-known economic consulting firm NERA offers the following service on its website: "NERA's workforce reviews help to minimize employment law risk in several ways. With our detailed statistical analysis of workforce composition, NERA experts can: Assess the impact of proposed reductions-in-force (RIFs)." [6]Biddle Consulting Group's (BCG) web page rhetorically asks "Want to know more about Biddle's proactive EEOC/AI consulting services" – as a lead-in to their consulting services web page[7]. BCG offers to assist clients to "take proactive steps to minimize your risks. Biddle Consulting Group, Inc. provides several services that allow employers to "manage" the risk involved in their employment decisions[8]."

[1]Jaikes is corresponding author; Email: ejaik1@unh.newhaven.edu.
[2]For purposes of illustration I will use a reduction in force or RIF as my archetypical workforce event. The argument I propose here and the empirical tests apply with equal validity to all Title-VII workforce events
[3]The argument here may also apply to disparate impact under the Age Discrimination in Employment Act (ADEA), the Equal Credit Opportunity Act (ECOA) and the Fair Housing Act. However, we have not examined the applicability of our argument here to any other but Title VII matters.

The objective of the firm anticipating having to institute a reduction-in-force or any such event is to deploy a suitable strategy within the recognized parameters established by the courts to immunize the firm and thereby avoid the burden and cost of litigation[9]. In successfully forestalling litigation the firm curtails the increased cost and perhaps more importantly, the increased scrutiny that would accompany a lawsuit.

Self-serving classifications and resulting favorable outcome ratios of those considered for a RIF workforce event proffered by defense may contain any number of necessarily "misclassified" individuals. The concern applies to those employees ultimately selected as well as to those who were overlooked.

In other words, some employees originally exempted from the RIF – based on some unobserved (to plaintiff) standard or valid selection measure- were ultimately reclassified as terminated for purposes of meeting statistical thresholds. Similarly, (or, in the alternative) some employees originally selected were subsequently reclassified as retained or not-terminated – again upon appraisal of the original race or gender ratios. In yet another variant, a firm may opt to negotiate ex ante settlements, separation packages, or any such idiosyncratic agreement with one (or many) selected individuals to effectively remove them from the "pool" of employees so as to again engineer the desired statistical outcomes.

The costs associated with the statistically-sanitized outcome described above, the one incurred by impacted individuals who are denied relief, may amount to substantial amounts of monies. Still, perhaps the most egregious failing of the process is the social cost of justice denied.

A showing of no association between an employment event and gender or race (as the case might be) necessarily follows the traditional methodology of postulating the existence of no ex-ante observable difference in the realized rates – the null-hypothesis significance testing (NHST). But this formulation provides the right answer to the wrong question posed; a question advanced as an integral part of the active management of the event. This correct answer to the wrong question is what is known as a type III error (Schwartz and Carpenter 1999).

The concept of a type III error is attributed to Kimball who warned of the possibility of correctly answering the wrong question (Kimball 1957). [10]But where Kimball assumed that a type III pitfall might be a result of a statistical expert's carelessness or haste and therefore unintended, in the modern world statistical consultants are unashamedly advocates and invariably protected by attorney-client privilege. In such a setting it is clear that for a firm interested in minimizing its exposure – shoehorning plaintiffs into the wrong question is the desired precedent to the desired outcome.

The controversy surrounding the applicability, soundness and relevance of "social framework analysis" generally and around Betty Dukes, et al. v. Wal-Mart Stores, Inc11. specifically, attests to the desire to illuminate processes and practices associated with and underlying workplace events that may not be reflected in the statistical tests commonly required by the courts. Betty Dukes and other Wal-Mart employees alleged gender discrimination in pay and promotion policies and practices in Wal-Mart stores. Their ensuing complaint contained 120 sworn affidavits describing what they characterized as anecdotal evidence of discrimination.

Whereas Social Framework Analysis sought to provide a theoretical context to explain the persistent and compelling influence of unconscious or implicit bias against women and minorities in the workplace, our focus, on the other hand, is to highlight the role of mechanisms that preclude the possibility of actionable events emerging. Still, both efforts point to the filed plaintiff affidavits and other similar qualitative information as informative and relevant.

In this paper we argue that given evidence of the active management of a workforce event a higher threshold for establishing a rebuttable presumption of adverse impact may be warranted. Thus, we examine the relative tradeoffs obtained by altering the adverse impact threshold. Specifically, we examine the costs and benefits of replacing the four-fifths rules with a more stringent 9/10ths rule to enable a showing of adverse impact.

In Search of a Higher Threshold: a More Stringent Rule of Thumb or More Accommodating Confidence Limits

Other commentators have argued that enlarging the acceptance interval in statistical tests would achieve the same outcome. This is of course correct. A test of significance of 10 percent is more likely to find a statistical teste demonstrative – allowing plaintiff to meet her burden. However, given the heuristic and practical reliance on rules of thumb by labor consultants and human resource professionals, altering the four-fifths rule would convey a similarly more accommodating threshold.

In effect, by increasing the threshold from 0.8 to 0.9 (for example) we reduce the probability of incurring a Type II error but increase the Type I error. That is to say, we increase the probability of mistakenly rejecting the null of no discrimination. But given that the null has been vitiated by the gerrymandering – the appropriate null is one of discrimination. And if discrimination exists – then an decrease in the possibility of Type II error enhances the chances of a plaintiff being provided relief.

[4]A matter is considered actionable if statistical significance tests of the observed disparity can be distinguished from the outcome that would occurentirely by chance assuming that no disparity exists. Put differently, we have to be confident (at some acceptable level) that the observed outcome is actually occurring and is not simply a chance realization.

[5]Pavlick & Barbatsuly (2008, December 10) K&L Gates' Labor and Employment Alert, December 10, 2008 (at 2).

[6]NERA Economic Consulting at http://www.nera.com/59_2144.htm {viewed June 16, 2011}.

Adverse Impact

The Adverse Impact doctrine was established by the Supreme Court in *Griggs v. Duke Power Company*. The court held that it is not legal for a firm contemplating a reduction-in-force to use a selection metric unrelated to job performance that affects legally protected groups at a disproportionately higher rate. The Equal Employment Opportunity Commission subsequently articulated Griggs and defined Adverse Impact as follows:

A substantially different rate of selection in hiring, promotion, or other employment decision which works to the disadvantage of members of race, sex, or ethnic group.

And laid forth how the EEOC will determine whether adverse impact exists.

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. Smaller differences in selection rate may nevertheless constitute adverse impact, where they are significant in both statistical and practical terms. . Greater differences in selection rate may not constitute adverse impact where the differences are based on small numbers and are not statistically significant. … (Section 1607.3D).

The Four-fifths Rule

In contrast to statistical metrics such as the Chi-Square test and the Fisher-Exact test, the four-fifths rule is intuitive and remarkably simple to deploy. A rebuttable presumption of adverse impact is established if the selection rate of any protected group is less than 4/5ths of the rate for the group with the highest rate. Not surprisingly given its comparative ease of deployment, a recent study found the four-fifths rule to be the most widely used and routinely applied measure of adverse impact (Bobko and Roth 2010).

The 0.90 Rule

The EEOC's decision to adopt the four-fifths rule as the appropriate ratio rather than the five-sixths or nine-tenths rule is entirely arbitrary. The EEOC has noted that the four-fifths rule is a "rule of thumb" and "not a legal definition of discrimination, rather it is a practical device to keep the attention of enforcement agencies on serious discrepancies. " As the name suggests, the 0.90 rule is proffered to establish the threshold for a finding of adverse impact when the selection rate of a protected group is less than 90 percent of the rate for the group with the highest rate.

Testing the Indicators

To apply the rule parties must first calculate the selection rate for each group and then divide the selection rate of the minority group $SR_{minority}$ by the selection rate of the majority group $SR_{majority}$. The Adverse Impact Ratio (AIR) is defined as follows:

$$AIR = \left(\frac{NP_{min}}{N_{min}} \Big/ \frac{NP_{maj}}{N_{maj}} \right)$$

Table 1: Cross-Tabulated Frequency Table of Selected Outcomes

Reference	Fail/Not Selected	Pass/Selected	Total	Proportion
Minority	NF_{min}	NP_{min}	N_{min}	P_{min}
Majority	NF_{maj}	NP_{maj}	N_{maj}	$1 - P_{min}$
Total	NF_T	NP_T	N	
Proportion	$1 - SR_T$	SR_T		

If the AIR is lower than four-fifths or eighty percent constitutes a prima facie case of disparate impact.

To illustrate: suppose that a workforce event has selected 17 women for termination and 11 men. There were originally 30 women and 15 men in the pool originally considered. Thus, the female selection ratio is 17/30 = 0.57 whereas the male selection ratio is 11/15 = 0.73. Dividing the two selection rates to calculate the Adverse Impact Ratio we conclude that the selection ratio of females to males is seventy-eight percent (.57/0.73 = 0.78). Because this ratio is less than eighty percent, the disparity is actionable under the four-fifths rule.

The Size of the Tests

A statistical test T is said to have statistical size α for testing a null hypothesis H0 when rejection of H0 is defined as $T > Z\alpha$ and $Pr(T>Z\alpha|H0 \text{ is true}) = Pr(\text{rejecting } H0|Ha \text{ is false}) = \alpha$.

The AIR test proffers as a null hypothesis that there exists no evidence of a disparity.

Thus, the rejection of a null entails a measure of a Type I error.

[7]Biddle Consulting Group at http://www.biddle.com/eeo-litigation-support.php {viewed July 7, 2011}.

[8]http://www.biddle.com/consulting.php {viewed July 7, 2014}.

[9] The "gerrymandering" point made here is not necessarily limited to the field of employment. In their book on statistical significance, Ziliak & McCloskey (2008) surmise that Merck employees or the scientists who ran the clinical trials testing the effectiveness of Merck's Vioxx, dropped three instances of unhelpful observations in the data "in order to get an amount of statistical significance low enough to claim …[ ] … a zero effect." Stephen T. Ziliak and Deirdre McCloskey, The Cult of Statistical Significance (Ann Arbor, University of Michigan Press), at 29.

Statistical power is denoted as $1-\beta$, with the Pr($T > Z_\alpha | H_0$ is false) = Pr(rejecting $H_0 | H_a$ is true) = $1-\beta$; the probability of rejecting the null when the alternative hypothesis is true. Power and size are reciprocals to one another in the sense that the power of a test is increased when a larger, less stringent statistical size α is selected (or equivalently, a smaller Z_α is used).

Empirical Methodology,

We constructed STATA program to replicate the data generating process underscoring various Adverse Impact Ratio distributions.

The generated distributions are parametrized by The SRs for each of two subgroups, a majority and a minority. The steps of the simulation are as follows:

1. We chose an applicant pool of size, n, where n = {10, 20, 30, 40, 60}; a composition of the minority group within the pool (Pmin); and the pool selection rate (Psel).
2. The number of minorities selected for each particular realization is a result of a random draw from a hypergeometric distribution with integer valued parameters; N is the population size, K is the number of elements in the population that have the attribute of interest, and n is the sample size.
3. We estimate the realized distribution of the Adverse Impact Ratio (AIR).
4. We measure the Type I error rate for both adverse impact ratios, AIR ={0.8, 0.9} – assuming that the data generating process reflects an AIR of 1, i.e. a state of the world of no discrimination.
5. We measure the Type II error rate for both adverse impact ratios, AIR ={0.8, 0.9} – assuming that the data generating process reflects an AIR of 0.5; i.e. a state of the world where discrimination is present.
6. This process is reproduced 10000 times via a Monte Carlo simulation.

We generate via monte carlo simulation the distribution of the test statistic AIR = 0.9. We run 10,000 iterations using Stata. We conduct a binomial test determine whether the realized distribution is equal to AIR = 8. The size of the test is the Type I error rate.

Rejection rates correspond to Type 1 error rates for IR = 1.0 and to power when IR = 0.8 and IR = 0.6 and IR = 0.4.

Results

We examine whether the levels of Type I errors for both the four-fifths and the three-fifths rule by using resampling methods to draw multiple simulated samples to calculate statistical significance (Good 2001).

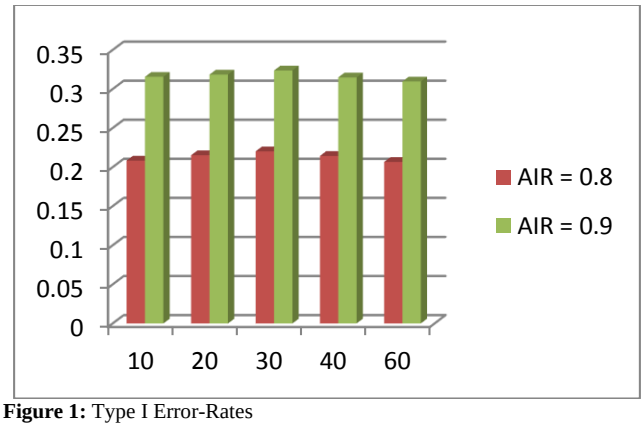

Figure 1: Type I Error-Rates

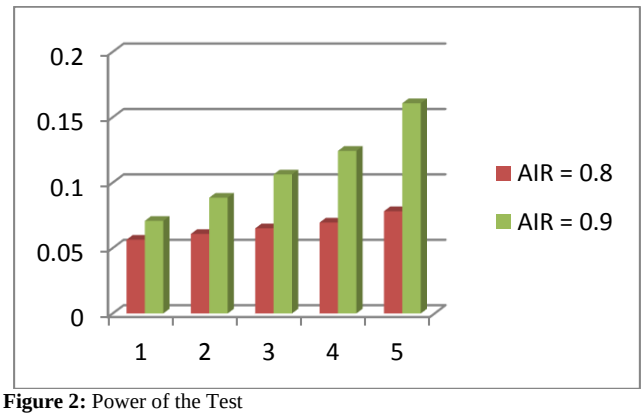

Figure 2: Power of the Test

Interpretation of Results and Concluding Comments

While there is no single uniform test the Courts have relied on either tests of statistical significance or on the four-fifths rule and even on both to adjudicate Title VII matters. The appeal of the four-fifths rules is its simplicity and intuitive appeal.

Unfortunately, legitimately aggrieved plaintiffs can be denied relief in potential Title VII disparate impact discrimination cases if firms actively manage the tests that a establish prima facie case. Succinctly, firms can pick and choose before the reduction-in-force to derive favorable statistical ratios. After all, in practically all forums, plaintiff's rebuttable presumption in disparate impact and disparate treatment cases is duly established by a favorable showing of the four-fifths rule or a statistical showing of outcomes.

[10] There is an alternative definition of a type III error in the statistical literature. A type II error occurs when a false null hypothesis is rejected by the claimed direction of truth is opposite to what it really is (MacDonald 1999). Put differently, the direction of the statistical inference is opposite from the real direction.

[11] (Betty Dukes, et al. v. Wal-Mart, Inc. 2007)

In fact the professionalization of Title VII-related advice on workforce interventions enables and facilitates such this activity. Firms retain consultants to manage workplace event results to comply with received case-law and enforcement agency regulations. The objective of the firm's tinkering, reduced to its essence is to forestall litigation by gerrymandering favorable statistical tests of significance to achieve a seemingly facially neutral employment outcome.

Presumably, the favorable gender-ratios or race-ratios resulting from the planned process will pre-empt litigation or, at the very least, dramatically reduce its chances.

Here we propose a more ample interpretation of the EEOC's rule of thumb – the four-fifths rule in a manner that will provide relief to aggrieved plaintiffs.

We appraised the incremental impact on the Type I and Type II error rates of raising the threshold ratio to establish a rebuttable presumption of discrimination. We examine a threshold of 0.9 for purposes of illustration. We find that the improved opportunities for plaintiffs of the presence of a nine-tenths rule are not offset by the reduced likelihood of reductions in Type II errors.

References

Betty Dukes, et al. v. Wal-Mart, Inc. 474 F.3rd 1214 (9th Circuit, 2007).

Bobko, P., and P. L. Roth. "An Analysis of Two Methods for Assessing and Indexing Adverse Impact: a Disconnect Between the Academic Literature and Some Practice." In *Adverse Impact: Implications for Organizational Staffing and High Stakes Selection*, by J.L. Outtz, 29-49. New York: Routledge, 2010.

Breiman, Leo. "Statistical Modeling: The Two Cultures." *Statistical Science* 16, no. 3 (2001): 199-231.

Collins, M. W., and S. B. Morris. "Testing for Adverse Impact When Sample Size is Small." *Journal of Applied Psychology* 93 (2008): 463-471.

Good, Phillip I. *Resampling Methods.* Boston: Birkhauser, 2001.

Imbens, Guido M., and Jeffrey M. Wooldridge. "Recent Developments in the Econometrics of Program Evaluation." *NBER Working Paper No. 14251*, 2008.

Kimball, A. W. "Errors of the Third Kind in Statistical Consulting." *Journal of the American Statistical Association* 52, no. 2 (1957): 133-142.

MacDonald, Paul. "Power, Type I, and Type II Error Rates of Parametric and Nonparametric Statistical Tests." *The Journal of Experimental Education* 67, no. 4 (1999): 367-379.

Mehrotra, Devan V., S. F. Chan, and Roger L. Berger. "A Cautionary Note of Exact Unconditional Inference for a Difference Between Two Independent Binomial Proportions." *Biometrics* 59 (June 2003): 441-450.

Mehta, Cyrus R., and Pralay Senchaudhuri. "Conditional versus Unconditional Exact Tests for Comparing Two Binomials." 2003.

Nyhart, Nick. "A Consensus for Reform: Connecticut Lawmakers Opt for Public Financing." *National Civic Review*, Summer 2006: 3-10.

Rodriguez, A.E. *Beyond Fisher: On the Evidentiary Foundations of Alteranative Tests of Significance.* San Diego: 86th Annual Western Economics Association Conference, 2011.