NERA
Economic Consulting

6 July 2009

Statistical Lessons of *Ricci v. De Stefano*

By **Jonathan Falk**

The Supreme Court's decision in *Ricci v. De Stefano*[1] has already garnered a great deal of attention from lawyers, political pundits, and Supreme Court watchers. Although the statistical issues received very little attention in the decision from either side, there are important statistical currents in *Ricci*—as there are in any disparate impact case—that are worthy of further attention. This brief will focus on three issues: (1) what do statisticians really have to say about disparate impact? (2) how might statistical analysis have played out in *Ricci*? and (3) going forward, what role do statisticians have to play in the new standard (i.e., strong basis in evidence)?

To frame the discussion, it will help to lay out a few facts about the case. The City of New Haven, CT hired a company to develop a promotional test for firefighters that would accomplish two objectives: test only skills relevant for promotion and, subject to that , minimize disparate impact of the results.[2] It should be noted that under the precedent provided in *Griggs v. Duke Power*, any procedure that accomplishes the first task has a safe harbor against allegations of disparate impact, albeit a safe harbor that might well have been challenged in court on the facts.[3]

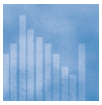
The lieutenant's test[4] was given to 43 white firefighters and 19 black firefighters, producing the following results: 25 whites passed (58 percent) and six blacks passed (32 percent).[5] The New Haven procedure for promotion involved a second phase and the results of that phase meant that 10 whites (40 percent of those passing the test) and no blacks (0 percent) would actually receive promotions.

The Statistics of Disparate Impact

The Civil Rights Act of 1964 and its subsequent amendments prohibit discrimination in promotion. Facially neutral promotion policies are still suspect under Title VII of the Act when their application produces a "disparate impact." Surprisingly, the actual judicial interpretation of what "disparate impact" means is quite thin and the questions of when statistical analysis is appropriate, and to what purpose it should be put, are particularly vague. This is not surprising, since decisions are written by

judges, who are well versed in the law but are unlikely to have had substantial training in statistics. As a consequence, the overwhelming tendency in disparate impact cases is to rely on one of two precedents. First are the Supreme Court's statistical pronouncements, which appear in footnotes in the *Casteneda v. Partida*[6] and *Hazelwood School District v. US*[7] cases.[8] Those footnotes have engendered extensive commentary, but it suffices to note that the Court in these cases allowed, but did not mandate, the use of statistical analysis to help assess disparate impact.

The second reference that is uniformly employed is the so-called "four-fifths rule" promulgated by the Equal Employment Opportunity Commission, in which a rate of promotion for a disfavored group which is less than four-fifths of the rate for the most favored group is "evidence of adverse impact."[9] It is this rule that the majority decision in *Ricci* cited for the proposition that "[t]he racial adverse impact here was significant, and petitioners do not dispute that the City was faced with a *prima facie* case of disparate impact liability." Looking to the lieutenant's exam, the pass rate of blacks was about half that of whites, and the fact that no blacks would have been promoted following the second phase is clearly substantially short of the promotion rate for whites.



The "four-fifths rule" is not a legal standard on its own, nor has it been characterized by the courts as anything more than a "rule of thumb."[10] Nor should it be—the rule itself is qualified in EEOC regulations and has received substantial scorn from statisticians who have looked at it. And as we shall see, the actual disparities observed in the test fall short of the standards set forth in *Castaneda* and *Hazelwood*.

The basic calculation that is normally employed begins with the assumption that, *ex ante*, all candidates ought to be equally capable of securing promotion in an unbiased test. If that is true, then we can calculate the probability that one group will have a particular success rate as a function of the total number of successful candidates, and the number of candidates in one of the groups.[11] This is the basis for the so-called Fisher Exact Test, and it is the most commonly used test in these circumstances, though it is not the only one that might be employed. If the combined probability of all events less than or equally likely than the one which actually occurred is sufficiently low, then we can conclude that one of three things has happened:

- Something unusual has occurred by chance
- The test has taken equally qualified people and is somehow biased in its result
- People weren't equally qualified to begin with

It is the logical disjunction of these three possibilities that explains why a statistician's evidence in a disparate impact case can never be dispositive. The statistician's calculations can shed some light on these three possibilities, but it requires a finder of fact to separate these three causes for the observed result.

The standard procedure for a statistician is to choose a probability below which the first possibility, i.e., "Something unusual has occurred by chance," ought to be discounted. A rule of thumb in social science is 5 percent, but the court should recognize that this is no more than a rule of thumb and needs to be considered in light of both the sample size involved and the other facts in the case.

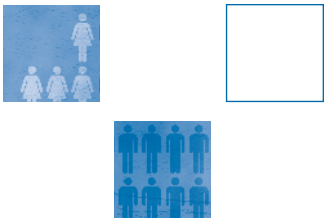
After the statistician has disposed of this possibility by declaring the result "statistically significant," the next step is to assert (sometimes directly, sometimes indirectly) that the test was biased. This is because the assumption that everyone was equally well qualified *ex ante* is an assumption by the statistician that cannot be altered without invalidating the procedure. Thus, of the three possibilities, one is rejected by the use of a rule of thumb and one is rejected by assumption, which leaves only the third possibility.

But the court is not so constrained. Suppose, for example, that the test results make it clear that five black candidates and five white candidates were manifestly unqualified for some reason. Then these candidates should not have been included in the original statistical analysis. If this is the case (using the New Haven numbers), the white pass rate (from among the *ex ante* equally qualified candidates) will rise, but the black rate will rise more. Indeed, this possibility is explicitly mentioned in the EEOC guidelines.[12]

What the statistician can usefully do to aid a court is to explain the assumptions behind the analysis, to explain the calculations made under those assumptions, and to explain what would be the normal conclusion in professional practice subject to those assumptions being true. The finder of fact, weighing the totality of the evidence, is then in a position to make a determination of disparate impact.

A Quick Look at the *Ricci* Results

Suppose we make the common assumption that all lieutenant candidates, black and white, were equally well qualified. When we calculate the probabilities under the Fisher Exact Test, the results may be somewhat surprising. First, the probability that, given entirely equal *ex ante* qualifications, results equally unusual or more unusual than 25 of the 31 passing scores coming from white candidates would be expected 10 percent of the time. This is substantially in excess of the standard statistical rule of thumb and thus is not "statistically significant" as that term is normally used. It is important to note that that does not mean the disparity is not legally meaningful, but a result such as this would not normally pass muster in a refereed journal as a reliable index of statistical inference.



Among those passing, the probability of an all-white group of promotees (or even more unlikely events) is even higher: 14 percent. Again, this is not dispositive on its own but it certainly suggests that, had New Haven been sued over the utilization of these results, they would have had substantial weapons to use in their defense.

Considered as an integrated process, however, the pass rates for blacks are significantly low at standard levels of the significance. The relevant Fisher Exact Test probability is 2.4 percent. Thus, if we regard the compound issue of promotions as the significant issue, then there is at least something mildly suspicious about the entire procedure.

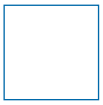
Even here, however, there is a problem. The compound procedure consists of a test and a promotion rule based on the test scores. If the test does what it is supposed to do, then the test itself, under *Griggs*, is a safe harbor. If that is true, then we are left only with the promotion rule from among successful candidates which, as we have seen, is not even close to statistically suspicious. But if the test doesn't do what it is supposed to do, there's no reason it should have been used in the first place. Thus, any inquiry into this process ought to focus on the reliability and job-relatedness of the test, not on the results that come out of it.

The tendency of non-statisticians (and occasionally, even statisticians) to use the unexpected results of a procedure to criticize the procedure is what statisticians call the p-value fallacy.[13] The time to look for anomalies in the test is not after you've observed results, but before. The City of New Haven held hearings *after* the results came out to decide whether to certify them and gave great weight to experts who opined (without having looked at the test) that the results were indicative of a problem. With all due respect, they cannot have known that, because the proper calculation of that probability would have to take into account that there never would have been a hearing in the first place if there had not been a disparity in the test results.

We can draw an analogy with the causes of winning baseball. A team wins the World Series and it is argued that their superior cohesion caused them to win. And indeed the team might have superior cohesion. But we cannot draw conclusions that their cohesion was a reason for their winning without examining the cohesion of teams that had poorer records. This is particularly true when we reason backwards, observing the winning team and then inventorying their characteristics. The chance of spurious attribution rises dramatically when we proceed in this fashion.[14]

Lessons from the New Rule

The Supreme Court's main holding in *Ricci* is that one can't simply be afraid one will be sued before throwing out the results of a test: one must have a so-called "strong basis in evidence" that one will lose such a suit. What useful advice can a statistician give about such a standard?



First, it is clearly insufficient to simply look at the results of the test without seriously analyzing the results. On their own, without further fact-finding, it appears that the results of the *Ricci* test would have been defendable in court. While additional inquiry and analysis might have undermined that tentative conclusion, the City did not engage in any relevant fact-finding about the meaning of the results. Neither the majority opinion nor the dissent make any comment on this, apparently taking the discussion of the four-fifths rule as being dispositive, at least for the purposes of this case. However, the statistical insignificance of this result (if, upon full discovery, it turns out to be insignificant) would certainly have been an important issue at trial. Indeed, a similar situation arose in 1996, when the New Haven Police Department was sued by a group of minority policemen who argued that the results of a promotional test to sergeant had a disparate impact. Statistical analysis showed that while the disparities exceeded the four-fifths rule, they were not statistically significant.[15] The court agreed that, as a result, no disparate impact had occurred.[16]

Second, the test results should be the last place to look for test bias, not the first. In an ideal world, the test could be certified as fair in advance, which would obviate any need to look at the results. It is certainly possible that the results could still be surprising, but once we have calibrated the test for fairness, it would take extremely discrepant results to raise a reasonable suspicion. At that point, the strong basis in evidence standard makes more sense, because only very strong evidence would ever cause a reopening. If courts are going to require strong basis in evidence, it is imperative that statisticians be used to help assess the strength of that evidence. If a suit is eventually filed, statistical evidence will undoubtedly be used. So it is impossible to assess the basis in evidence without the help of statisticians.

Third, finders of fact should spend less time searching the literature for rules of thumb and more time trying to understand what the statisticians are telling them. This lesson is, of course, a two-way street, and many statisticians are guilty, consciously or unconsciously, of trying to make their results either more mystifying than they need to be or more certain than they could possibly be. The role of the expert witness in court testimony is too far afield to be discussed at length here, but the *Daubert* line of cases springs from the proposition that finders of fact seem to be overawed by scientific testimony into ceding their responsibility to the expert. In addition to serving as a gatekeeper to bar "unscientific" theories, judges should disallow testimony that does not provide sufficient understanding of the underlying assumptions to allow the finder of fact to assess those assumptions in light of the facts of the case.

Fourth, the case law needs to recognize the p value fallacy that arises many times in litigation and is now an endemic problem with the legal system. *Ricci* presents an excellent example of the p value fallacy in action. The City called several witnesses who said they had seen results similar to those that New Haven had in exams that were biased. That is undoubtedly true, but it is not the question to be answered. Since the results showed a disparity, and since biased tests show disparities, the

testimony of these witnesses is fine as far as it goes—but unbiased tests sometimes show disparities as well. That witnesses who had not even reviewed the tests should feel qualified to comment on the *cause* of a disparity is certainly statistically invalid. And a qualified statistician who could have commented usefully on the results would have to preface any findings with a list of assumptions, most critically that everyone was equally qualified *ex ante*. Surely that assumption could not possibly be properly the subject of disposition on summary judgment, and must form an important part of any assessment of the strong basis in evidence standard.

End notes

1 www.supremecourtus.gov/opinions/08pdf/07-1428.pdf.

2 "Minimize" is used here in a nonmathematical sense.

3 http://www.law.cornell.edu/supct/html/historics/USSC_CR_0401_0424_ZO.html.

4 There was a test for Captain as well, which I am ignoring for expository purposes.

5 For expository purposes, I will also leave out the Hispanic firefighters who took the test. There are numerous issues involved in doing so, but they need not concern us here because these other factors complicate the analysis but do not essentially change it.

6 http://supreme.justia.com/us/430/482/case.html.

7 http://supreme.justia.com/us/433/299/case.html.

8 For an excellent reference to the statistical issues involved, see Meier, Sacks, and Zabell, "What Happened in Hazelwood," in DeGroot, Fienberg and Kadane, eds., *Statistics and the Law*, John Wiley, 1986.

9 http://law.justia.com/us/cfr/title29/29-4.1.4.1.8.0.21.4.html at D. This rule is also discussed in Meier, Sacks, and Zabell, ibid.

10 *Watson v. Fort Worth Bank and Trust*, http://supreme.justia.com/us/487/977/case.html.

11 The number of successful candidates in the other group is automatically determined by subtraction of successful candidates in one group from successful candidates in total.

12 "Greater differences in selection rate may not constitute adverse impact where … special recruiting or other programs cause the pool of minority … candidates to be atypical of the normal pool of applicants from that group."

13 See, for example, Sellke, Bayarri, and Berger, "Calibration of p values for Testing Precise Null Hypotheses," *American Statistician*, 55(1) (2001), pp. 62-71.

14 Ibid. In the article cited, Sellke, Bayarri, and Berger demonstrate that, on the assumption that the test is as likely to be biased as unbiased, finding a p value of 5 percent (the benchmark level) will actually mean that the test is unbiased at least 29 percent of the time and will often be unbiased close to half the time. In other words, far from being "statistically significant," the 5 percent benchmark level provides almost no evidence at all.

15 Testimony of Jonathan Falk in *New Haven County Silver Shields, Inc., et. al. vs .City of New Haven, et.al.*, District of Connecticut Case 94CV01771 (PCD).

16 Decision of Chief US District Judge Peter C. Dorsey in *New Haven County Silver Shields, Inc., et. al. vs. City of New Haven, et.al.*, District of Connecticut Case 94CV01771 (PCD).

NERA
Economic Consulting

About NERA

NERA Economic Consulting (**www.nera.com**) is a global firm of experts dedicated to applying economic, finance, and quantitative principles to complex business and legal challenges. For nearly half a century, NERA's economists have been creating strategies, studies, reports, expert testimony, and policy recommendations for government authorities and the world's leading law firms and corporations. We bring academic rigor, objectivity, and real world industry experience to bear on issues arising from competition, regulation, public policy, strategy, finance, and litigation.

NERA's clients value our ability to apply and communicate state-of-the-art approaches clearly and convincingly, our commitment to deliver unbiased findings, and our reputation for quality and independence. Our clients rely on the integrity and skills of our unparalleled team of economists and other experts backed by the resources and reliability of one of the world's largest economic consultancies. With its main office in New York City, NERA serves clients from over 20 offices across North America, Europe, and Asia Pacific.

Contacts

For further information and questions, please contact:

Jonathan Falk
Vice President
NERA Economic Consulting
+1 212 345 5315
jonathan.falk@nera.com