Formal statistical analysis of the data in disparate impact cases provides sounder inferences than the U. S. government’s ‘four-fifths’ rule: an examination of the statistical evidence in Ricci v. DeStefano

JOSEPH L. GASTWIRTH*

Department of Statistics, George Washington University, Washington, DC 20052, USA

AND

WEIWEN MIAO

Department of Mathematics, Haverford College, Haverford, PA 19041, USA

[Received on 24 April 2009; revised on 22 May 2009; accepted on 4 June 2009]

Many countries have fair employment laws to protect racial, gender, religious or ethnic minorities from discrimination and courts in the USA can order remedies such as one out of every three new hires should be a member of a protected group after finding an employer discriminated. What steps can an employer undertake to ensure its employment practices do not disadvantage minorities when it does not need to comply with a court order? This issue arose in Ricci v. DeStefano, a ‘reverse discrimination’ case under review by the U.S. Supreme Court. Seventeen Whites and 1 Hispanic who achieved sufficiently high scores qualifying them for promotion to lieutenant or captain of the New Haven Fire Department sued the city because it cancelled the examinations after seeing that no African American could be appointed to an existing vacancy. The City of New Haven justified its action on the basis that both examinations had a disparate impact on African Americans and Hispanics because the ratios of their pass rates to that of Whites were less than 80%, contrary to a ‘rule of thumb’ in the government’s Uniform Guidelines. The city did not conduct statistical tests, which are referred to in the guidelines.

The lower courts accepted New Haven’s explanation and granted summary judgement to it. A statistical study of the various criteria considered by the city and lower courts in their review of the data demonstrates that nearly 70% of the time a fair non-discriminatory test for either position will fail the government’s ‘80% rule’ and at least 60% of the time both fair tests would fail this ‘four-fifths rule’. Since the city created a new criterion after seeing the results, it is difficult to formulate precisely the other ‘rare’ or ‘unusual’ outcomes that would lead to cancellation of the examination. Would New Haven reject a list with no Hispanics or no Whites eligible for an immediate promotion? Would it require that all three groups be represented in the pool eligible for advancement to each position?

From the viewpoint of statistical theory, the hypothesis being tested and the definition of pass or selection rates that will be compared should be decided before examining the data. Formal statistical tests on several relevant pass rates show that the lieutenant examination had a disparate impact on minority applicants, but the differences in the pass rates on the captain examination were not close to statistical significance. Furthermore, when the city cancelled both examinations, it only focused on the demographic mix of the high scorers who could receive an immediate promotion and ignored the 2-year life cycle of the list. Neither likely retirements nor job turnover during the 2-year life cycle.

* Email: jlgast@gwu.edu

© The Author [2009]. Published by Oxford University Press. All rights reserved.
of the results were considered. If this had been done, the city might have realized that three African Americans were likely to be appointed lieutenants along with two Hispanic captains.

**Keywords**: disparate impact; equal employment; ‘four-fifths’ rule; numerical disparity; reverse discrimination; tests of statistical significance.

### 1. Introduction

In a sharply divided *en banc* opinion, the U.S. Court of Appeals for the Second Circuit affirmed a district court’s summary judgement dismissing a reverse discrimination suit by 18 plaintiffs who achieved high scores on promotion examinations for lieutenant and captain in New Haven’s Fire Department. The city cancelled the examination, justifying its decision on the basis that the test had a disparate impact and an insufficient number of minorities would be promoted to an *existing* vacancy, although the test results would be used for 2 years. New Haven’s Charter specifies that when ‘g’ promotions are made, the department must select them from the top $r = g + 2$ scorers. The U.S. District Court for the District of Connecticut noted that the city’s reasons for not certifying the examination were related to the racial distribution of the results. The opinion relied on the government’s four-fifths rule to compare the pass rates, even though the government guidelines specifically state that when the sample sizes are small, differences in selection rates that fail the ‘rule’ may not constitute an adverse impact when they are not statistically significant. The parties did not submit the results of statistical tests of the data, so inferences obtainable from such tests were not discussed by the lower courts. Because city officials looked at both the basic pass rates (having a score of at least 70) and the proportions of individuals from each race-ethnic group with a score putting them in the top $r = g + 2$ candidates who could be promoted to an *existing* vacancy, they applied two ‘statistical tests’ to the data. The Supreme Court granted _certiorari_ to clarify the circumstances that would justify the rejection of the results of an otherwise valid selection

---

1. 530 F. 3d F. 3d 88 (2nd Cir. 2008).
2. 554 F. Supp. 2d 142 (D. CT 2006).
3. Ibid. at 143, 149. At the time of the examination, there were eight lieutenant and seven captain openings and no African American would be promoted immediately to either position.
4. An employment test passes the four-fifths rule when the ratio of the lowest pass rate to the highest pass rate is at least 4/5 or 80%. These Uniform Guidelines, 29 C.F.R. 1607.6(D) (2000), were adopted by the Equal Employment Opportunity Commission, the Department of Labor’s Office of Federal Contract Compliance, the Department of Justice and the Civil Service Commission in order to have a common set of principles for examining the fairness of employment selection procedures. A convenient compilation of the questions and answers concerning the implementation of the guidelines that the agencies have issued can be found at [www.management-advantage.com/media/UniformGuidelinesQ_A.pdf](http://www.management-advantage.com/media/UniformGuidelinesQ_A.pdf).
5. Ibid.
6. See the statement of the City’s counsel, Mr. Ude (Jt. App. 146) indicating that both the pass rates and the distribution of scores were considered. In her deposition, the City’s Chief Examiner, Ms. N. Marcano, noted that while the previous 1999 examination results pass–fail data did not satisfy the four-fifths rule, the proportions of applicants from the three groups who were those selected for advancement were considered (Jt. App. 172–173). More promotions were made that year and members of all three groups were promoted (the data are analysed in Section 4.1).
7. This is important because had the test been certified the list would be used for 2 years. Data on likely retirements and past turnover in the department from which an estimate of the number of future openings could be made were not considered. In fact, eight more lieutenants could have been appointed during this period and three of them could have been given to African Americans (Petitioners brief at 26). One additional captaincy opened up before the end of 2005 (Deposition of Mr. Frank Ricci, Pet. App. 384a).
8. 129 S. Ct. 893 (2009). The questions presented are reproduced in the Appendix. They were obtained from the court’s website ([www.supremecourt.gov/qs/07-1428](http://www.supremecourt.gov/qs/07-1428)).
process that yields an unintended racially disproportionate outcome. This paper demonstrates how statistical methods can clarify the issues in the case. New Haven officials, the federal trial court and the Second Circuit evaluated the examinations for both positions together and determined that both had a disparate impact. Statistical tests confirm that the differences among the pass rates of the three groups on the lieutenant examination were statistically significant but not on the captain examination.

Section 2 of the paper provides more information about the case and arguments of the parties. In Section 3, it will be seen that for the sample sizes involved in either examination, more than 60% of the time a fair test would be classified as having disparate impact if the four-fifths rule is applied to either of the two pass rates of the three race-ethnic groups.

After reading the briefs and a recent case concerning the fairness of a promotion examination in another fire department, we think that several additional pass rates deserve to be considered, including the race-ethnic composition of those likely to be selected from the top group of applicants for the immediately available positions by the city under its diversity program and those eligible for or likely to be selected when the expected new openings during the 2-year life cycle of the examination results are included. Formal statistical analysis shows that the differences between all relevant types of pass rates of the three groups on the examination for captain were not close to statistical significance at the commonly used 0.05 level. In contrast, the differences in the various pass rates of the three groups on the lieutenant examination were statistically significant. Because these inferences are in conflict, combining the results of both tests to obtain an overall effect is questionable. Moreover, the results of the lieutenant examination do not seem very relevant to resolving the claim of a Hispanic plaintiff, who achieved a high score on the captain examinations.

9 Ms. N. Marcano stated that she believed there was adverse impact that would have resulted in disparate treatment of minorities who completed the examinations (Jt. App. 224) and Ms. T. Burgett said ‘yet the end result on these two examinations is always the same we are still left with a significant adverse impact that cannot be accounted for’ (see Pet. App. 579a).


11 The Supreme Court accepted statistical significance testing in Castenada v. Partida, 430 U.S. 482, 496 n. 17 (1977) where it describes a null hypothesis (no difference) as ‘suspect to a social scientist’ when a statistic falls between 2 and 3 SDs from what is expected. Baldus and Cole (1980) at 297 note that these standard deviations are approximately equivalent to probability levels of 0.05 and 0.01, respectively, and that the 0.05 or 5% level is used in many applications, Id. at 291. It should be noted that the results of the statistical tests are correlated as they are being applied to the same set of data. In a statistical study, one would correct for this (see infra n. 19). Since it is for the court, rather than statisticians to select the legally appropriate rate, all analyses are given in Section 3.2. As the ‘expected’ number of vacancies was not considered in the proceedings, the data on the actual openings that became available are used in the analysis for the fourth rate.

12 Because the process by which the tasks to be tested on both examinations was the same (a survey of current incumbents), the lower courts might believe that they were sufficiently similar. It is unclear that the material covered on the examinations for the two positions was the same. Before combining the results of examinations given in different years for the same position, courts review them to ensure that the topics covered and the weightings they received were similar. See Fudge v. City of Providence 766 F. 2d 650, 652-654 (1st Cir. 1985).

13 Most disparate impact cases concern the effect of a test or employment criterion on applicants for a particular position. In Isabel v. Memphis, 404 F. 3d 404 (6th Cir. 2005), the disparate impact of a promotion examination on African American candidates for promotion to lieutenants in the Memphis Police Department was established after the difference between the pass rates of African American and White applicants was shown to be statistically significant. Courts have considered the issue of when it is appropriate to pool statistical data on different minority groups into a larger data set. In Lewis v. Aerospace Community Credit Union, 114 F. 3d 745, 750 (8th Cir. 1997), the court stated that the plaintiff needs to demonstrate a test had a disparate impact on his or her particular group. In contrast, minority groups were combined in Kohn v. City of Minneapolis Fire Department, 583 N.W. 2d 7 (Minn. Ct. App. 1998) after plaintiffs showed that their pass rates were similar. See Mems et al. v. City of St. Paul Fire and Safety Department 224 F. 3d 735 (8th Cir. 2000) for further discussion and cases concerning smaller samples than those in the present case.
Potential insights from other statistical analyses are discussed in Section 4. The differences in pass rates on previous tests are compared with those in the Ricci case. Furthermore, data on turnover and the ages of current employees could have been used to estimate the expected number of future openings during the following 2 years. Then, the city might have realized that three African Americans could probably be promoted to lieutenant. Section 5 discusses the potential implications of formal statistical analyses for Ricci and makes suggestions for improving the assessment of statistical evidence and testimony in disparate impact cases.

2. Background

2.1 The federal trial court opinion

In November and December 2003, the New Haven Fire Department administered examinations, with both a written and an oral component, for promotion to lieutenant and captain. Under the contract between the city and the firefighters union, the written portion received a weight of 60% and the oral 40%. Applicants who achieved a total score of at least 70 passed the examination and became eligible for promotion. The eligibility list was to remain valid for 2 years, and the city charter mandated that the person selected should be among the three highest scorers remaining on the list.

The tests were designed by a 7-year-old firm that specializes in developing appropriate valid examinations for public safety positions. In March 2004, after the results of the examinations were available, the city refused to certify those examinations because an insufficient number of minorities would receive a promotion to an existing position. Eighteen individuals, 17 Whites and 1 Hispanic, who had high scores on the examination and consequently a high probability of being promoted, filed a ’reverse’ discrimination suit.

Forty-one applicants took the captain examination (25 Whites, 8 African Americans and 8 Hispanics). Twenty-two passed (16 Whites, 3 African Americans and 3 Hispanics). There were seven vacancies at the time of the examination and those positions would be filled from applicants with the nine highest scores. Of the nine top scorers, there were two Hispanics but no African Americans. Specifically, the ranks of the minority passers were: Hispanics ranked 7, 8, and 13 and African Americans ranked 16, 19, and 22. Notice that the pass rate of each minority group (37.5%) is less than four-fifths that of the Whites (64%), although the number of applicants is small.

Seventy-seven applicants took the lieutenant examination (43 Whites, 19 African Americans and 15 Hispanics). Thirty-four passed (25 Whites, 6 African Americans and 3 Hispanics). The African Americans ranked 14, 15, 16, 20, 22 and 24 while Hispanics ranked 27, 28 and 31. There were eight vacancies (including the seven jobs of the lieutenants who would be promoted to captain) at the time of the examination and the top 10 scorers were all Whites. Thus, no minority member would be promoted to lieutenant. The pass rates of the three groups were 58.1% (White), 31.6% (African American) and 20.0% (Hispanic). The opinion noted that the minority rates were much less than four-fifths that of the Whites.

The opinion describes the process the developer used to ensure the examination would be valid and focused on the information and tasks needed to successfully perform the positions. The

---


15 554 Supp. at 147. The developer interviewed a sample of current officers in the department to determine the tasks, knowledge and skills needed for each position. The frequency and importance of each task or skill was assessed and then ranked relative to how essential or critical they were. The examinations were reviewed by two experts from other states. While...
panelists for the oral examination were chosen to ensure diversity. The vice president of the testing firm stated that the questions were firmly rooted in the study materials on the syllabus, which were distributed with the promotion applications.

New Haven’s Civil Service Board (CSB) conducted several hearings and an industrial psychologist from a competitive test development firm testified that while he had not studied the test in detail it had a relatively high adverse impact. In contrast, a fire program specialist from the U.S. Department of Homeland Security testified that the test was appropriate. Another expert witness told the CSB that minorities generally perform relatively poorly on written tests, the questions may have been skewed because 67% of the individuals responding to the survey about the knowledge and skills needed for each job were White and firefighters of different backgrounds perform firefighting jobs differently. In addition, she noted that a bilingual speaker or one with an accent may elicit more negative reactions from evaluators in oral examinations.16

At the last hearing on 18 March 2004, counsel for New Haven strongly argued against certifying the examinations because they had a disparate impact, which could violate federal law. Moreover, he claimed that there was a greater disparity on this test than on previous ones created by a different firm. While the selection ratios of the 1999 tests and the 2003 tests in question were similar, there were more African Americans among the top scorers in the previous examination.17 From a statistical viewpoint, the city considered the disparate impact of its tests by looking at whether the four-fifths rule was satisfied in two ways:

1. By the proportions of applicants from the three race–ethnic groups scoring above 70.
2. By the proportions of applicants from the three race–ethnic groups scoring above 70 and being among the top scorers from whom the immediately available promotions would be made.

A representative of the Mayor also advocated discarding the results because the rankings on the eligibility list combined with the number of existing vacancies would lead to no African American being promoted to either position and no Hispanic promoted to lieutenant. It is difficult to formulate the Mayor’s criteria precisely. Did the Mayor mean that when promotions to both jobs are made, at least one member of each of the three groups should receive one of the immediately available promotions or the appointments to each position should include at least one member from each race–ethnic group?18

16 Ibid. at 149.
17 According to the opinion of the district court, 554 F. Supp. 2d at 154, the ratio of the African American pass rate to the White one for the lieutenant examination in 1999 was 0.58 while that of the 2003 examination was 0.59. Similarly, the corresponding ratio for the 1999 captain examination was 0.45 while that of the 2003 examination was 0.59. Our calculation of this adverse impact ratio for the 2003 lieutenant examination is 0.543, rather than 0.59. This is a minor difference. The results of the captain examination, which appears to have been given in 1998, rather than 1999, are discussed in Section 4.1. The decision notes that while witnesses testified that in these two previous examinations, more minority candidates had scores near the top of the list and had more promotional opportunities; the actual scores or ranks were not submitted to the court. The available data are analyzed later in Section 4.1 and the results are compared to those of the examinations at issue.
18 The second formulation translates the requirement for at least one Hispanic to be on the list of candidates eligible for promotion to lieutenant, in addition to the two eligible to be promoted to captain, to mean that both the lieutenant and the captain lists include at least one individual from each of the three race–ethnic groups. Because Whites were the largest group of applicants, under random selection they would be expected to have a correspondingly higher representation in the
No consideration appears to have been given to whether or not the differences in either of the pass rates would meet or come close to the commonly used 0.05 level of statistical significance, which has been accepted in many legal decisions. Furthermore, neither party nor the lower courts appreciated that when multiple tests are carried out at any pre-specified level of significance, even when the null hypothesis of no difference between the pass rates is correct, the probability that at least one of the tests yields a significant result is greater than the pre-specified level.\(^\text{19}\)

The federal trial court opinion notes that both the plaintiffs’ evidence and the defendant’s arguments show that the city’s reason for not certifying the results was related to the racial distribution of the scores. The court observed that the case presents a scenario opposite to the usual one where minority applicants complain of disparate impact and the defendant then submits evidence that the test is a valid job-related predictor of performance. Indeed, plaintiffs argued that the city’s refusal to conduct a validation study of the examination is evidence of reverse discrimination. The court felt that while the government guidelines were written under the assumption that employers would defend and validate their pre-employment tests, the guidelines do not require an employer to conduct a validity study of a selection procedure having a disparate impact.

The plaintiffs also presented evidence indicating that they, and presumably other White applicants, studied harder and took more advantage of the available study groups than most minority candidates.\(^\text{20}\) On the other hand, an African American firefighter said that the study materials were difficult to obtain and not available in all firehouses.\(^\text{21}\)

The court decided that the plaintiffs did not have a viable disparate impact claim because the city’s cancelling the examination affected all applicants equally. Finally, the court accepted the city’s argument that it made its decision on the basis of the disparate impact of the examination on minorities, which would subject the city to Title VII lawsuits from minority applicants that for political reasons it did not want to defend.

### 2.2 The petitioners’ brief to the Supreme Court

While the plaintiffs’ brief primarily focused on legal issues, it provided some additional background information. First, it emphasized that when the decision to cancel the examination was made, the candidates were race-coded.\(^\text{22}\) It also noted that the city knew that previous promotion tests had top scorers. For example, under the first interpretation (at least one member of each group receives a promotion to either position), no African Americans or no Hispanics corresponds to a shortfall of 4.2 or 3.7 promotions. In terms of standard deviation units, used by the court in \textit{Castenada}, 430 U.S. 482 (1977), the city would cancel the examination if African Americans or Hispanics have a deficit of \(-2.53\) or \(-2.34\) SDs, respectively. In contrast, zero Whites or a shortfall of 11.07 places on the two lists of top scorers corresponds to a deficit of \(-5.62\) SDs. Thus, these interpretations of the Mayor’s criteria imply that the city requires an examination to have a greater disparate impact on Whites than on a minority group before questioning its fairness. This appears inconsistent with the Uniform Guidelines, which treat the pass rates for all groups similarly in the calculation of the four-fifths rule or the appropriate statistical test of significance.

\(^{19}\) See Kaye and Freedman (2000) at p. 83 for a general discussion and Gastwirth (2008) for an illustration of the effect of ignoring multiple tests in a discrimination case and references to the statistical literature.

\(^{20}\) 554 F. Supp. 2d at 151 and 156.

\(^{21}\) 554 F. Supp. 2d at 146.

\(^{22}\) See Petitioner’s brief (February 2009) at p. 10 (section IV) which is available from the American Bar Association’s website (www.supremecourtpreview.org). The race coding appears to have been done after the examination was graded but before the results were reviewed by the city. Thus, it may have less importance than in early jury discrimination cases; e.g. in \textit{Avery} v. Georgia 345 U.S. 559 (1953) where no African Americans were on the petitioner’s jury. The fact that the names of African American potential jurors were written on yellow cards, while the names of Whites were on white cards indicated that
similar numerical disparities, but the city had not taken steps to reduce them.\[23\] Although the city decided, in part, to cancel the examination because no African Americans would receive a promotion to an existing vacancy during the entire time period when the examination scores would have been used, eight additional lieutenant positions became available. Thus, a total of 16 lieutenants would have been appointed from the list, 3 of whom could have been African American.\[24\]

The plaintiffs argued that the examinations were content valid, i.e. they tested for knowledge and skills needed for the job, and the composite scores were calibrated to equate to minimal competence. The developer had conducted an appropriate post-examination analysis and was prepared to deliver the technical report until city officials stopped the process.\[25\] While statistical methods are used in the process of validating pre-employment or promotion examinations, they do not appear to be central to the issues raised by the parties and will not be discussed here.

2.3 Respondent’s brief to the Supreme Court

The city’s brief emphasized that its response was prudent because it was faced with examination results indicating an extreme statistical disparity.\[26\] After conducting a deliberative open process, it declined to use the results. The city argued that since there were well-grounded questions about the validity and legality of the test, its response was in accord with the constitution and equal employment law.\[27\]

Candidates for the lieutenant post needed to have 30 months of experience in the New Haven Fire Department, possess a high school diploma and completed training courses in relevant subjects. In addition to having a high school diploma and passing courses in relevant fields, applicants for captain had to serve 1 year as a ‘certified fire lieutenant’ in the city’s fire department.

The city asserted that the test developer did not establish a content-valid legally defensible cut-off score for the examination that corresponds to a minimum level of competency for the position. The brief also questioned the thoroughness of the review of the examination by New Haven’s fire experts as well as by outsiders.\[28\] The city disputed the petitioners’ claim that it had prevented the developer from preparing a technical report that would substantiate the content validity and lawfulness of the test. The City of New Haven also questioned whether that report would have met the legal standards for a validation study that would protect it from future litigation.\[29\]

The city’s brief emphasized the magnitude of the disparity between the pass rates of African Americans and Whites. Specifically, it noted that the pass rates for African Americans on both lieutenant and captain examinations were about one-half of those of Whites and that ‘more jarringly’ out of the 19 possible candidates for the 15 available positions none would be African American.\[30\] It argues that the disparate impact of the 2003 examinations exceeded the disparities on previous

\[23\] Ibid. at 41 (suggesting a special tutoring programme).
\[24\] Ibid. at 61 n. 27. This point is repeated in the Petitioner’s reply brief at 22.
\[25\] Ibid. at 7, 11 and 51–52.
\[26\] Respondent’s Brief on the Merits in Ricci v. DeStefano (March 2009) at 54, which is available from the American Bar Association’s website (www.supremecourtpreview.org).
\[27\] Ibid. at 2.
\[28\] Ibid. at 3 and 4.
\[29\] Ibid. at 11 and 31–32.
\[30\] Ibid. at 5–6.
examinations because the highest ranked African American candidates for lieutenant ranked 3rd in 1996, 5th in 1999, but 13th in the 2003 test. Similarly, the highest ranked African American applicant for captain ranked fifth in 1998, but was 15th in 2003.\footnote{31

Over 10 *amicus* briefs, primarily concerned with legal issues and standards, were also filed. None presented a statistical analysis of the data, so they will not be summarized here.

3. Statistical issues in applying the government guidelines and the ‘four-fifths’ rule to the data and a reanalysis of the data

3.1 The four-fifths rule and the pass–fail data

The trial court used the ‘four-fifths’ rule to assess the disparate impact of the examination, i.e. it compared the ratios of the pass rates of each minority group to that of the Whites to 0.80. The Uniform Guidelines, however, state that statistical tests should be calculated, especially in small samples:

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 . . .

As emphasized by Judge Sippel in *Stewart v. City of St. Louis*\footnote{32

2007 U.S. Dist. Lexis 75342 (E.D. Mo. October 2007) aff’d. 532 F. 3d 939 (8th Cir. 2008).} the Guidelines, at Question and Answer 20, state that the four-fifths rule should not be used in small samples. In such situations, federal courts often focus more attention on whether the differences between the pass rates are statistically significant than on the four-fifths rule.\footnote{33

See *NAACP v. City of Mansfield*, 866 F. 2d 162, 168 (6th Cir. 1989), *Frazier v. Consolidated Rail Corp.* 851 F. 2d 1447, 1451 (D.C. Cir. 1988) (approving the district court’s rejection of the 80% rule when the sample size was small and the plaintiffs did not present a significance test) and *Fudge v. City of Providence Fire Dep’t.* 766 F. 2d 650, 658-9 (1st Cir. 1985) (noting that in small samples, the four-fifths rule is not an accurate test of adverse impact). In the examination at issue in *Fudge*, 1 of 24 African Americans (4%) passed, while 29 of 244 Whites (13%) did. The court, *ibid.* at n. 10, was mindful of the fact that the four-fifths rule was violated but referred to the need for statistical testing of small sample sizes noted in the Uniform Guidelines, *supra* n. 4. From STATXACT, the two-tailed *p* values of Fisher’s exact test and Barnard’s unconditional test are 0.49 and 0.296, respectively. Thus, the difference between the two pass rates is not close to statistical significance at the commonly accepted level of 0.05.}

For each race–ethnic group, the number of applicants and the number who passed or were among the top \( r = g + 2 \) scorers on the examinations at issue in the *Ricci* case are given in Table 1. The next-to-last column reports the data for the likely promotions to the *existing* vacancies, assuming the city considered diversity, i.e. a minority will be promoted if he/she is among the top \( r = g + 2 \)

\footnote{Ibid. at 6. The complete results giving the scores and rank orders of the individuals who took the examinations in 1996 and 1999 were not submitted into evidence and may not have been kept by the city. There are minor discrepancies between the highest ranks for African Americans between the city’s brief, e.g. 13th for lieutenant and the district court opinion, 554 F. Supp. at 145 reporting 14th. The rankings in the opinion will be used in this paper.

\footnote{2007 U.S. Dist. Lexis 75342 (E.D. Mo. October 2007) aff’d. 532 F. 3d 939 (8th Cir. 2008).}

\footnote{See *NAACP v. City of Mansfield*, 866 F. 2d 162, 168 (6th Cir. 1989), *Frazier v. Consolidated Rail Corp.* 851 F. 2d 1447, 1451 (D.C. Cir. 1988) (approving the district court’s rejection of the 80% rule when the sample size was small and the plaintiffs did not present a significance test) and *Fudge v. City of Providence Fire Dep’t.* 766 F. 2d 650, 658-9 (1st Cir. 1985) (noting that in small samples, the four-fifths rule is not an accurate test of adverse impact). In the examination at issue in *Fudge*, 1 of 24 African Americans (4%) passed, while 29 of 244 Whites (13%) did. The court, *ibid.* at n. 10, was mindful of the fact that the four-fifths rule was violated but referred to the need for statistical testing of small sample sizes noted in the Uniform Guidelines, *supra* n. 4. From STATXACT, the two-tailed *p* values of Fisher’s exact test and Barnard’s unconditional test are 0.49 and 0.296, respectively. Thus, the difference between the two pass rates is not close to statistical significance at the commonly accepted level of 0.05.}
TABLE 1  Number of applicants from each race–ethnic group and the number of whom pass each of the five relevant pass rates on the 2003 examinations in the Ricci case

<table>
<thead>
<tr>
<th>Lieutenant</th>
<th>Pass</th>
<th>Fail</th>
<th>Total</th>
<th>Top 10</th>
<th>Top 18</th>
<th>Likely promotions</th>
<th>Likely promotions</th>
</tr>
</thead>
<tbody>
<tr>
<td>African American</td>
<td>6</td>
<td>13</td>
<td>19</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Hispanic</td>
<td>3</td>
<td>12</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>White</td>
<td>25</td>
<td>18</td>
<td>43</td>
<td>10</td>
<td>5</td>
<td>8</td>
<td>13</td>
</tr>
<tr>
<td>Total</td>
<td>34</td>
<td>43</td>
<td>77</td>
<td>10</td>
<td>15</td>
<td>8</td>
<td>13</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Captain</th>
<th>Pass</th>
<th>Fail</th>
<th>Total</th>
<th>Top 9</th>
<th>Top 10</th>
<th>Likely promotions</th>
<th>Likely promotions</th>
</tr>
</thead>
<tbody>
<tr>
<td>African American</td>
<td>3</td>
<td>5</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Hispanic</td>
<td>3</td>
<td>5</td>
<td>8</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>White</td>
<td>16</td>
<td>9</td>
<td>25</td>
<td>7</td>
<td>5</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>Total</td>
<td>22</td>
<td>19</td>
<td>41</td>
<td>9</td>
<td>10</td>
<td>7</td>
<td>8</td>
</tr>
</tbody>
</table>

The next-to-last column reports the data for the likely promotions to the existing vacancies, assuming the city considered diversity. The last column gives the corresponding data for the positions that became available during the 2-year life cycle of the examinations.

The next-to-last column reports the data for the likely promotions to the existing vacancies, assuming the city considered diversity. The last column gives the corresponding data for the positions that became available during the 2-year life cycle of the examinations.

Five possible definitions of ‘pass’ are relevant:

1. First, ‘pass’ is defined as scoring at least 70.
2. Second, a ‘pass’ is having a score of 70 or more and is in the top \( r \) from whom the vacant positions at the time of the examination will be filled.
3. Third, a ‘pass’ is a score of at least 70 and is in the group of high scorers from whom the vacant positions would be filled during the 2-year period when the examination results would be used.
4. The fourth ‘pass’ is being one of the likely selections when the city considers diversity. In other words, once a minority member passed the examination and is among the top \( r = g + 2 \) group, he/she will receive a promotion.
5. The fifth criterion is the analog of fourth for the 2-year life cycle of the examination.

The Uniform Guidelines indicate that when the same test is used for several jobs, the assessment of disparate impact should be done on a job-by-job basis. The examinations in Ricci were designed for different jobs, although they were developed in similar fashion. Hence, it is appropriate to examine the pass rates for each position.\(^{34}\)

Before analysing the data, a simulation study was conducted to determine the probability that a fair test, one where the scores from all three race–ethnic groups have the same distribution, fails the ‘four-fifths’ rule. This calculation assumes that the applicants are random samples from a large population of individuals having the education and experience required to take the examination.

For each test, the specific overall pass rate was calculated \((34/77 = 0.4416 \text{ for lieutenant and } 22/41 = 0.5367 \text{ for captain})\) and the corresponding cut-off value of a standard normal distribution

\(^{34}\) See the Uniform Guidelines, supra n. 4 at Sec. 15A (2) (a) and Question and Answer 27 in the 44 Fed. Reg. No. 43 (2 March 1979).
was obtained.\textsuperscript{35} For the lieutenant examination, samples of 43, 19 and 15 were selected from the same standard normal distribution and the pass rates for each group calculated. Similarly, the captain examination scores for applicants of each group were simulated from samples of 25, 8 and 8. Ten thousand simulations were carried out for each examination. Table 2 reports the fraction of times the ‘four-fifths’ rule was violated by a fair test. Column A refers to the first criterion for passing listed above (having a score of at least 70). Notice that fair tests would fail to satisfy the ‘four-fifths’ rule 70\% (87\%) of the time for the lieutenant (captain) position; so, it is almost inevitable that one of the tests would be deemed to have a disparate impact.\textsuperscript{36} Column B is the proportion of times that the ‘four-fifths’ rule is violated when applied to the second criterion for ‘passing’. The results in Column B indicate that over 85\% of the time, a fair test for each position would violate the four-fifths rule. Column C reports the results when the ‘fourth-fifths’ rule is applied to the third ‘pass’: a score of at least 70 and is in the top group from whom the vacant positions would be filled during the 2-year life cycle of the examination. Again, over 85\% of the time, a fair test for each position would violate the four-fifths rule. Similar results were obtained for the fourth and fifth criteria for ‘passing’. These findings illustrate the need to determine whether the differences among the pass rates of the three groups are statistically significant and are consistent with the Uniform Guidelines and scholarly literature.\textsuperscript{37}

3.2 Formal statistical analysis of the data

In this section, statistical tests are used to assess the significance of the differences in the success rates of the three groups. For the first definition of passing, because the passing score of 70 was set before the examination, the numbers of applicants from each ethnic group who pass the examination are not known in advance. In statistical terms, they are random variables. In this situation, the appropriate statistical test of the equality of the pass rates is the chi-squared test (Mosteller & Rourke, 1979). In contrast, when membership in a fixed number of high scorers is of interest, e.g. belonging to the top \( r = g + 2 \), the appropriate test is the Fisher–Freeman–Halton or FFH test (Freeman and Halton, 1979). For both examinations and all three race–ethnic groups, each applicant’s score is simulated from the standard normal distribution. The cut-off passing score for lieutenant (captain) is the 44.16th (53.67th) percentile of the standard normal. Individuals whose simulated test score is higher than the cut-off value are classified as passing the examination, i.e. having a score 70 or above on the actual examination.

\textsuperscript{35} Since different individuals took each test, the two applicant pools can be considered independent random samples from distinct populations of potential candidates for the two positions. Then, the probability of at least one of the two pass–fail rates of fair tests not satisfying the ‘four-fifths’ rule is 0.70 + 0.87 – 0.70 × 0.87 = 0.96.

\textsuperscript{37} See Shoben (1978) at 805–06 (four-fifths rule can produce anomalous results because it fails to account for differences in sample size and neglects the differences among pass rates), Greenberg (1979) (concluding that both types of errors, classifying a fair test as having a disparate impact or failing to detect a test having a disparate impact, can have a high probability of occurring when the four-fifth’s rule is used) and York (2002) (showing that the usual chi-squared test for the difference between two pass rates and the four-fifth’s rule often yield contradictory conclusions).
1951; Hirji, 2006). This test is used to examine the second through fifth ‘Pass’ rate data. Table 3 gives the p values for chi-squared and FFH tests for all the five definitions of ‘Pass’.

First, notice that the statistical tests for equality of all five relevant ‘pass’ rates on the captain examination have p values exceeding 0.25; clearly greater than the usual threshold of 0.05. The smallest p value, 0.2513, is for data on the first definition of ‘Pass’ (scoring at least 70). This means that one-fourth of the time, differences between the three proportions at least as large as observed in the data would occur if the scores of the applicants were randomly selected from a common distribution. The value 0.2513 is also described as the probability that chance or a lottery would yield an outcome as far from expected as the observed data. The fourth p value, 0.515, referring to the likely immediate promotions, shows that about one-half of the time, a random selection or chance process would produce a disparity at least as large as the observed one. This p value indicates that the most likely pattern of promotions is in the central region of the possible outcomes of a random selection process. Thus, formal statistical analysis supports the plaintiffs who scored sufficiently high on the captain examination to be eligible for a promotion. One reason these results contrast so sharply with the ‘four-fifths’ rule is that the numbers of test takers from the three groups are quite different. There were 8 African Americans, 8 Hispanics and 25 Whites. If one randomly selects a member from this applicant pool, there is a much greater chance that a White applicant would be chosen than a minority member.  

The inadequacy of the ‘four-fifths’ rule in the context of evaluating the captain examination data for disparate impact can be demonstrated by supposing that one fewer White was in the ‘likely to be promoted to an existing position group’ while, one African American was in the same group. Now, the ‘selection rates’ would be White (4/25 = 0.16), African American (1/8 = 0.125) and Hispanic (2/8 = 0.25). The ratios of the selection rates of Whites and African Americans to that of Hispanics would be 64% and 50%, respectively, both less than 80%. The p value of the FFH test applied to the data is 0.849, implying that the data set is very near the ‘centre’ of possible outcomes of a fair examination. Yet, these data violate the ‘four-fifths’ rule. This modification is similar to the one used in the Uniform Guidelines to illustrate problems that can arise when the ‘four-fifths’ rule is applied to selection rates based on small samples; see Question and Answer 21 in the Guidelines.

It is helpful to compare this analysis with a similar study of the data on the captain examination in the St. Louis case.  Two possible scenarios were considered, one for existing vacancies and the other included future appointments. If the 62 existing vacancies were filled, then 58 would be awarded to the 348 White candidates and only 4 would go to the 56 African Americans. Including the expected future openings, there would be 120 promotions. Of these, Whites would receive 112 and African Americans only 8. The opinion noted that the selection ratios were similar, 42.9% for the existing positions and 44.4% for the future promotions, both violated the ‘four fifths’ rule.

---

**Table 3**

<table>
<thead>
<tr>
<th></th>
<th>1st Pass</th>
<th>2nd Pass</th>
<th>3rd Pass</th>
<th>4th Pass</th>
<th>5th Pass</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lieutenant</td>
<td>0.018</td>
<td>0.0077</td>
<td>0.0102</td>
<td>0.0312</td>
<td>0.0285</td>
</tr>
<tr>
<td>Captain</td>
<td>0.2513</td>
<td>0.318</td>
<td>0.297</td>
<td>0.515</td>
<td>0.41</td>
</tr>
</tbody>
</table>

---

38 This statistical model is appropriate for the situation where the score of each applicant is considered as a random sample from one common distribution of scores. This implies that the probability the top scorer would be White is 25/41 = 0.61.

39 Stewart, supra n. 10.


TABLE 4 Results of the 1999 lieutenant examination (after adjustment)

<table>
<thead>
<tr>
<th>Group</th>
<th>Pass (selected)</th>
<th>Fail</th>
<th>Total</th>
<th>% Pass (selected)</th>
</tr>
</thead>
<tbody>
<tr>
<td>African American</td>
<td>8 (8)</td>
<td>18</td>
<td>26</td>
<td>30.77</td>
</tr>
<tr>
<td>Hispanic</td>
<td>9 (9)</td>
<td>7</td>
<td>16</td>
<td>56.25</td>
</tr>
<tr>
<td>White</td>
<td>26 (26)</td>
<td>29</td>
<td>55</td>
<td>47.27</td>
</tr>
<tr>
<td>Total</td>
<td>43 (43)</td>
<td>54</td>
<td>97</td>
<td>44.3</td>
</tr>
</tbody>
</table>

Applying Fisher’s exact test, which is appropriate for comparing equality of two proportions, the two-sided \( p \) values are 0.073 and 0.0069, respectively. Because these \( p \) values are much smaller than 0.2513, the \( p \) value of the test on the captain examination data in Ricci, the disparities between the pass rates of African American and White candidates for captain in the St. Louis case for either data set are of greater statistically significance than the disparities in Ricci.

Now, the same statistical methods will applied to the data for the lieutenant position in Ricci. For the data using the first criterion, passing is a score of at least 70; the chi-squared test yields a \( p \) value of 0.018. Using the second criterion, having a score above 70 and having a score in the top 10, the \( p \) value of the FFH test equals 0.0077, i.e. less than 1%. Thus, under both criteria mentioned in the opinion, the differences are statistically significant at the 0.05 level, so the examination would need to be validated as job related. Similar results were obtained when the third, fourth and fifth ‘pass rates’ were analysed. Thus, under any of the five possible criteria for ‘passing’, the success rates of the three groups on the lieutenant examination were statistically significantly different, implying that the examination had a disparate impact.

It may seem puzzling that the statistical analyses of the impact of the two examinations on minority applicants for the two positions yield different conclusions. Since candidates for the captain examination had to serve for at least 1 year as a lieutenant in the department, it is quite possible that there was less variation in firefighting knowledge, training and experience among them than among the applicants for the lieutenant position.\(^{40}\)

4. Further potential uses of statistical reasoning in the case

This section demonstrates how a more thorough study of the information available might have clarified some of the issues and led to a less contentious outcome. The second subsection describes the potential use of other data not in the record to estimate the expected number of future openings and it is somewhat more speculative as the data are not available for analysis.

4.1 Statistical analysis of the results of previous examinations

Although the complete rankings or scores of the applicants on the previous lieutenant and captain examinations that were accepted and used by the city are not in the record, the available pass rate data are reported in Tables 4 and 5.

\(^{40}\) See Respondent’s brief, n. 25 at 2-3 for the background requirements for each position. All candidates for captain passed a previous examination for the lieutenant position. A similar situation occurred in Stewart, supra n. 10 at pp. 12–13, where the tests in question had a disparate impact in the lower level position but not for the higher level one. Also, see Judge Posner’s opinion in Allen v. Seidman, 881 F.2d 375, 380 (7th Cir. 1989) (noting that the effect of differences in the educational backgrounds of the candidates would wash out with experience on the job).
Notice that there were more applicants (97) and appointments (43) made from the 1999 lieutenant examination than in Ricci. All scores on the 1999 lieutenant examination were adjusted in order to increase the fraction of applicants whose score reached 70 or more. This procedure did not change the relative rankings of the applicants. Hispanics had the highest success (either pass or selected) rate (56.25%) and the ratio of the African Americans pass rate to that of the Hispanics is 54.70%, which fails the ‘four-fifths’ rule. A formal test of significance, however, shows that the three success proportions are not statistically significant (p value = 0.218). Because the p value of the FHH test on the 2003 lieutenant data is significant (p value = 0.018), formal statistical testing supports the city’s claim that the disparity in the pass rates on the 2003 lieutenant examination was greater than the disparity in the results seen of the 1999 test. It does not confirm that assertion with regard to the 2003 captain examination. Indeed, the p value, 0.2513, of the FFH test of the 2003 captain examination is very close to the p value (0.218) on the 1999 lieutenant examination.

For the 1998 captain examination, the rates of selection of both African Americans (37.5%) and Whites (71.8%) were less than four-fifths that of the Hispanics (100%). The difference between the rates, however, is not statistically significant (p value = 0.096). Notice that the disparities in these data are similar to those of the 2003 examination with the exception that none of the African Americans could be promoted to an existing vacancy in 2003. This is due, in part, to the fact that 34 of 50 or 68% of the applicants were promoted in 1998, while only 7 promotions from a pool of 41 were available in 2003. Because the City of New Haven accepted the results of the previous captain and lieutenant examinations, with disparities similar to or greater than those on the 2003 captain examination, it appears that the city added a new criterion that requires that each set of promotions contain at least one person from each ethnic group after seeing the results.

The petitioners’ reply brief describes the disparate impact as ‘razor thin’ and argues that the numerical disparity was not severe because the adverse impact ratios in Ricci, although less than 0.80, were similar to those on the previous examinations. The results of formal statistical tests of the data here indicate that the magnitude of the disparate impact of the 2003 lieutenant examination was greater than that of the previous examinations, but this was not the case for the 2003 captain’s examination.

4.2 Were there reasonable alternative methods to lessen the adverse impact of the examination on minorities?

Since all promotions for the next 2 years would have been made using the examination results, why did New Haven focus only on the immediately available positions? While the plaintiffs’ brief to the court says that by ‘fortuity’, eight more lieutenant positions became available during the next

---

**Table 5 Results of the 1998 captain examination**

<table>
<thead>
<tr>
<th>Group</th>
<th>Pass (selected)</th>
<th>Fail (not selected)</th>
<th>Total</th>
<th>% Pass (selected)</th>
</tr>
</thead>
<tbody>
<tr>
<td>African American</td>
<td>7 (3)</td>
<td>1 (5)</td>
<td>8</td>
<td>87.50 (37.50)</td>
</tr>
<tr>
<td>Hispanic</td>
<td>3 (3)</td>
<td>0</td>
<td>3</td>
<td>100.0 (100.0)</td>
</tr>
<tr>
<td>White</td>
<td>35 (28)</td>
<td>4 (11)</td>
<td>39</td>
<td>89.74 (71.79)</td>
</tr>
<tr>
<td>Total</td>
<td>45 (34)</td>
<td>5 (16)</td>
<td>50</td>
<td>90 (68)</td>
</tr>
</tbody>
</table>

---

41 Filed on 7 April 2009 at 2.
42 Ibid. at 21.
2 years, the city could have estimated the number of likely vacancies from prior data on turnover and retirements.\textsuperscript{43} Surely, the ages of the current lieutenants and captains and the dates they would be eligible for retirement were available.\textsuperscript{44} As eight lieutenant openings arose, a reasonable estimate might well have been in the range of 5–10. Then, the city would have realized that only six vacancies during the next 2 years were needed for it to appoint three African American lieutenants.\textsuperscript{45} Perhaps New Haven could have decided to use the lieutenant list until the sixth additional opening occurred, even if it needed to extend the 2-year period. Then, the difference between the promotion rates of the three groups would no longer be statistically significant at the 0.05 level.\textsuperscript{46}

Had the existing seven captain vacancies been filled with five Whites and two Hispanics, four more positions would be needed before the next minority could be appointed and seven more before the second. Because there were seven vacancies and only 21 incumbents\textsuperscript{47} perhaps an estimate based on the ages of incumbents and past turnover would have indicated that no minority was likely to be reached. To maximize future opportunities, New Haven could fill the existing openings and let the list expire. Whether this would be legally proper, especially if the lieutenant list was used until the sixth additional vacancy occurred, is beyond our expertise. The main point is that data on past turnover and prospective retirements provide a basis for estimating the likely number of additional openings. These estimates would have enabled the city and courts to conduct a more thorough assessment of the overall impact of the examination results during their 2-year life cycle and assist the city to develop a strategy to increase diversity and minimize the disparate impact of the lieutenant examination.

4.3 \textit{Why it is statistically unsound for the city to add criteria after looking at the data}

Superficially, the observations made by New Haven officials that \textit{no} African Americans would be among those promoted to an \textit{existing} vacancy and \textit{no} Hispanic would advance to lieutenant might appear to be quite unlikely to occur. Indeed, \textit{before} seeing the data, the probability that no African American would be among those eligible for an immediate promotion is only 0.0052. The city presumably would have deemed results with \textit{no} Hispanics on either list sufficient to cancel the examinations too. The probability of this outcome is 0.0108. Of course, the city should also cancel the examination if there were too few Whites among those eligible for an immediate promotion.\textsuperscript{48}

\textsuperscript{43} See \textit{supra} n. 22 at 26 n. 11. The limited data in the record suggest that there is a fair degree of turnover, especially for lieutenants. As of the time of the case, there were 51 lieutenants and 21 captains (Jt. App. 217). In 1999, 43 lieutenants passed the examination and were promoted (Pet. App. 426a) and in 1996, 69 individuals passed the examination (Jt. App. 218), but the number promoted is not reported. Even if only 10 were promoted, the number of recently advanced lieutenants would exceed the current total. In 1998, 34 lieutenants were promoted to captain (Pet. App. 427a), a number exceeding the 28, the sum of the current number (21) and immediately available (7) captains in the department.

\textsuperscript{44} It is interesting that the court in \textit{Stewart}, \textit{supra} n. 10 at 356, assessed the impact of the examination in question by analysing data on both the existing and the expected promotions.

\textsuperscript{45} The district court opinion, 554 F. Supp. at 145 n. 3, reports that African Americans ranked 14, 15 and 16th.

\textsuperscript{46} Now 11 Whites, 3 African Americans and no Hispanics would be promoted to lieutenant and the \textit{p} value of the FFH test is 0.0898 or just under 10%.

\textsuperscript{47} See Jt. App. 217 reporting the composition of the fire department at the time. Since there were 7 vacancies out of 28 captains, it appears that several recently retired or left the department. It is now known that only one new vacancy occurred in the 2-year period, \textit{supra} n. 7 (Pet. App. 384a).

\textsuperscript{48} Since there were far more Whites among the applicants observing none among those who might receive a promotion to an existing vacancy is much rarer than finding no member of one of the minority groups on the list. As noted \textit{supra} n. 18, observing zero Whites on the two eligibility lists corresponds to a shortfall of 5.6 SD units, while observing zero African Americans or Hispanics corresponds to a shortfall of about 2.5 SDs. Depending on whether one measures the difference between the
Thus, the total probability of finding one of the three groups ‘under-represented’ for an immediate promotion before looking at the data is larger than the probability of finding any particular group ‘under-represented’. Looking at the data first and calculating the probability that data as extreme as that of the most under-represented group would occur leads us to believe that the observed data are more unusual than they really are. This happens because outcomes where the other groups are similarly ‘under-represented’ are not included in the calculation. If one looks at the data before defining the outcomes that are inconsistent with the hypothesis being tested, quite often one can find something unusual. This is why scientists define the hypothesis under investigation before looking the data.

It is difficult to discern from the statements of the city officials, the precise amount of diversity in the eligibility lists, they would accept. Since the city cancelled both examinations, even though two Hispanics had sufficiently high examination scores to be eligible for promotion to an existing captain vacancy, would it have been satisfied if only one African American could receive a promotion? Nonetheless, suppose that New Haven had a minimum requirement that its list of appointments to the existing vacancies contain at least one person from each of the three race-ethnic groups. The first two rows of Table 6 present the probability that a fair test for each of the positions would fail this criterion. The third row gives the probability that at least one of the two examinations would fail the criterion and the fourth row gives the probability that one of the groups would be excluded from both lists.

The results for each examination separately show that about 15% (20%) of the time, one of the three race-ethnic groups would be substantially under-represented in the list of individuals eligible for an immediate promotion for the lieutenant (captain) position. Because the calculation assumes that the applicants from each group were random samples from the same distribution of scores, applying the city’s criterion to each examination separately implies that about one-third of the time, at least one of two fair tests for the two positions would yield results that would be cancelled by observed and the expected number of Whites on the lists arithmetically or in standard deviation units, the minimum number of Whites that corresponds to finding no Hispanics on the combined eligibility list is six or seven. The probability of observing seven or fewer Whites on the two lists of individuals scoring sufficiently high to be appointed to a current vacancy is 0.035. Thus, the overall probability of at least one of the three groups being ‘under-represented’ on the lists of candidates eligible for an immediate promotion is closer to 0.05 than the 0.005 figure for African Americans. If one considers the actual selections for the current vacancies rather than the eligible list, then zero African Americans (Hispanics) corresponds to a shortfall of 2.23 (2.05) SD units. A similar shortfall of 2.10 SDs for Whites corresponds to five or fewer promotions for them. Now, the probability that one of the three groups would be under-represented, at the 2 SD level, among those selected is 0.083. Thus, 8% of the time, two perfectly fair tests were administered at least one of the three groups would not receive a promotion to an existing lieutenant or captain vacancy. Whether one should focus on the actual selections, the eligible pool or the applicants who passed the test is a legal, rather than statistical question. Since the ultimate inferences one can draw may depend on the choices of pass rate and threshold of statistical significance, it is important that whatever criteria will be used to determine whether an employment practice has a sufficiently large disparate impact to justify cancelling the entire process, should be set before the results are obtained. Otherwise, an employer can choose among the various pass or selection rates and use the one that supports whatever action or policy is desired.

49 See Kaye and Freedman (2000) at 127-28 (observing that after looking at a data set and observing some unusual pattern, if one ignores the previous search, statistical significance is bound to follow).

50 In many situations, e.g. clinical trials that will be used to demonstrate the efficacy of a new drug, the hypotheses to be studied and the statistical tests to be used are specified even before the data are collected. In that context, the null hypothesis is that survival rates of the new and old drugs are equal and the alternative is that the new drug improves survival. In Ricci, the null hypothesis is that the three pass rates are equal and the alternative is that they are different.
TABLE 6 Probabilities of two fair tests failing New Haven’s additional requirement that at least one member of each race–ethnic group be eligible for an immediate promotion

<table>
<thead>
<tr>
<th></th>
<th>At the time of examination</th>
<th>During 2-year period</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lieutenant</td>
<td>0.1440</td>
<td>0.0141</td>
</tr>
<tr>
<td>Captain</td>
<td>0.2144</td>
<td>0.1622</td>
</tr>
<tr>
<td>On both lists</td>
<td>0.3275</td>
<td>0.1740</td>
</tr>
<tr>
<td>On at least one list</td>
<td>0.0160</td>
<td>0.0012</td>
</tr>
</tbody>
</table>

the city. The situation is more interesting if the city would only reject both tests if one of the race–ethnic groups was not represented on the combined list of those eligible for an immediate promotion to either position. Now, a fair non-discriminatory test would only be rejected by New Haven only one and a half percent of the time. This is the first criterion satisfying the standard 0.05 level cut-off used in most statistical applications.\(^{51}\) The above analysis, however, ignores the fact that the city’s additional requirement was created after it saw the results. Suppose the data showed that one Hispanic rather than two were eligible to be promoted to captain, but one African American was eligible to be appointed to lieutenant. The city could have decided to ‘require’ that all groups receive one immediate promotion and the two eligibility lists contain a total of three or more minority members.

Another issue with New Haven’s apparent requirement that each group be represented on the combined list is that the probability that two fair tests would yield no Whites in the top-r, eligible for an immediate promotion, is much smaller than the corresponding probability for either minority group. A detailed calculation shows that no Whites corresponds to a shortfall of 5.6 SDs, while no African Americans or no Hispanics corresponds to a shortfall of about 2.5 SDs.\(^{52}\) Whatever shortfall the court decides is sufficient for an employer to cancel a pre-employment examination, should it not apply equally to all groups?

Suppose New Haven had taken the likely future vacancies into account and its estimates were near the actual numbers (eight lieutenants, one captain). The last column of Table 6 shows that the probability that a fair test would yield results with one of the three groups not being among the top-18 (top-10) on the lieutenant (captain) list is 0.014 (0.16). Considering this criterion jointly for the two examinations, the probability that one group would be excluded from both lists is just over one in a thousand. If this criterion had been established prior to looking at the results and one of the three groups been excluded from the likely promotions over the 2-year life cycle of the examinations, then the occurrence of such a low probability event would lend strong support for the city’s questioning the examinations. The actual data, however, indicates that three African American applicants for lieutenant and two Hispanic applicants for captain could have been appointed during the 2-year life cycle of the examinations.

\(^{51}\) See Sprent and Smeeton (2007) at 12, noting that the \(p\) value reflects the ‘strength’ of the evidence and interpreting \(p\) values between 0.05 and 0.10 as weak evidence against the null hypothesis. Their subsequent discussion also points out the need to consider the power of a test, also see infra n. 60 and 61. Recently, Zou et al. (2009) noted that courts sometime give weight to one-tailed \(p\) values under 0.05, citing In re Phenylpropanomolamine Prods. Liab. Litig. 289 F. Supp. 2d 1230, 1236 (D. Wash. 2003), while others, In re Ephedra Prod. Liab. Litig. 393 F. Supp. 2d 181, 191–192 (D. N.Y. 2005) consider \(p\) values in the range of 0.05–0.10 as weak evidence when the sample size is small.

\(^{52}\) See fn. 18 for details of the calculation.
4.4  *Did New Haven’s cancellation of the examination have a disparate impact?*  

The lower courts accepted the city’s argument that not using the results of the examination treated every applicant equally. Therefore, the plaintiffs did not have a viable claim of disparate impact or disparate treatment. Because applicants for Civil Service promotions often devote a substantial amount of time to prepare for these examinations, other courts might classify those candidates who scored sufficiently high to be eligible for promotions as disadvantaged by the city’s cancellation of the list. Not only did they spend time preparing for the 2003 examination, they will need to prepare for the next examination. In the context of evaluating disparities arising from a test or decision, being *disadvantaged* by New Haven’s action becomes passing the pre-employment examination. The possible disparate impact of the city’s cancellation of the examinations can be assessed by the same statistical techniques used in Section 3.2.

Once one realizes that data in Table 1 just need to be reinterpreted, the statistical analysis is routine. Let us consider the lieutenant data in Table 1 for the first criterion, achieving a score of at least 70 on the examination. Now, those who passed are disadvantaged and those who failed are advantaged by the cancellation. Thus, the ‘pass’ or ‘advantage’ rate becomes the fraction of applicants who failed the examination. The *p* value of the chi-squared test is identical to its previous value, 0.018. The most favourable data set for the plaintiffs concerns the likely promotions, 13 Whites and 3 African Americans to the 16 lieutenant appointments that opened up in the 2 years following the examination (last column of Table 1). The corresponding ‘success’ or ‘advantaged’ rates would be: White 30/43 = 69.77%, African American 16/19 = 84.2% and Hispanic 15/15 = 100%. The four-fifths rule would now classify the Whites as disadvantaged or burdened relative to Hispanics. The FFH test of the equality of ‘advantage rates’ again is statistically significant at the 0.05 level (*p* value = 0.0285) in agreement with the result in Section 3.2. Thus, if the court decides that canceling the examination is a burden on those who ‘passed’, under any of the definitions given in Section 3, the statistical significance of the disparate effect of the cancellation of an examination on Whites is the same as that of the statistical significance of the disparate impact among corresponding examination ‘pass’ rates on minorities.

5. Discussion

This article demonstrates that formal statistical analysis of pass rate data arising in disparate impact yields sounder inferences than simply relying on the ‘four-fifths’ rule. The results in Section 3.2

---

53 The dissenting opinion of Judge Cabranes, 530 F. 3d 88, 98 n. 5 (2008) questions Judge Parker’s conclusion, Ibid. at 90, that discarding the examination results in order to give members of one race a ‘second chance’ constitutes a conferral or denial of a benefit on the basis of race. If the court decides that the city’s action created a burden on the high scoring, predominantly White applicants, the statistical significance of the comparative impact of the cancellation of the examination on the three groups is provided by the analysis in this section. The word burden is used as the *amicus* brief of the U.S. Government, at 27 notes that New Haven’s decision burdened not only White firefighters but also the two Hispanics who were eligible for an immediate promotion and other minority firefighters who would have become eligible during the 2-year life cycle of the test.

54 Essentially, each of the top 10 had an 80% chance of being promoted so their expected loss is 80% of the increased pay they would receive had they been advanced to lieutenant plus the value of the time they may spend preparing for the next examination.

55 The chi-squared test in Section 3.2 is based on the differences between the pass rates of each group, while here it is based on the differences between the failure rates. These differences, however, are identical. For example, if 80% (60%) of Whites (minority group) pass, then 20% (40%) of each group fail. The difference in both cases is 20%.
showed that the pass rates, of the three race–ethnic groups, on the captain’s examination were not statistically significantly different. In Section 4.1, the $p$ value or significance level of the test of equality of pass rates on the 2003 captain’s examination were compared with its value calculated on data from earlier examinations. It was seen that the $p$ value of the statistical tests on the various pass rates in the 2003 captain examination data were very similar to, or even less ‘significant’ than, their values calculated on data from previous examinations that were accepted by the city.

It is important to emphasize that the hypotheses that will be subject to statistical testing should be specified prior to examining the data. When one looks at the data first, essentially one is conducting many comparisons. Selecting one or two ‘unusual’ outcomes, which may agree with a preconceived viewpoint, for formal testing often leads to finding a statistically significant result because the effect of the ‘multiple tests’ when the data were scrutinized is ignored. Just looking at the demographic mix of those eligible for an immediate promotion or of those who would be selected means that two statistical comparisons, one for each type of pass rate, were made. Furthermore, from their comments at the hearing, it is apparent that New Haven reviewed the results of each examination separately as well as jointly. Informally, the Mayor and other city officials made several comparisons, although we cannot determine the exact number. The multiple comparisons or testing issue arose previously in the context of a drug approval case in which a firm conducted about 240 statistical tests on various subgroups of patients. It found six significant (at the 0.05 level) and argued that the drug should be approved for those subgroups. The Food and Drug Administration was correct in not approving the drug because when there is no effect, under chance or random sampling, a statistical test at the 5% level will classify 1 in 20 results as significant. Hence, around 12 of 240 comparisons are expected to reach statistical significance. The six found by the firm were less than would be expected due to sampling variability.

In *Ricci*, the city justified cancelling both examinations on the basis that the results of each one violated the ‘four-fifths’ rule, no African Americans would receive a promotion to an existing vacancy and no Hispanic would be promoted to an open lieutenant position. Before combining the results of several statistical tests on stratified data, i.e. data organized into several subsets, one should check that it is appropriate to combine them. After looking at the results, the city added an extra

---

56 See Kaye and Freedman (2000) at 127-28 (noting that after looking at a data set and observing some unusual pattern, if one ignores the previous search, statistical significance is bound to follow).

57 Warner-Lambert v. Hechler 787 F. 2d (3rd Cir. 1986). The multiple comparison issue is discussed in Question and Answer 93 clarifying the section on validating employment tests in the Uniform Guidelines, 45 Fed. Reg. 29530 (2 May 1980). When statistical tests of the relationship of a selection procedure to each of several different job components are conducted, the answer cautions that one or two statistically significant results may occur by chance when correlations between many job components and job-test score are examined and subject to statistical testing.

58 Not only do courts carefully review the similarity of the examinations, as in *Fudge*, 766 F. 2d 650 (1st Cir. 1985), but one needs to use an appropriate method when combining the results of statistical tests on stratified data. Several procedures use a function, e.g. sum or product, of the $p$ values of the statistical test calculated in each stratum. See Gastwirth (1984) for an introduction to combination methods and Gastwirth and Pan (2009) at n. 57 for a more refined calculation of the type suggested by Justice (then Judge) Breyer in his concurrence in *Fudge* that confirms his intuition. Problems that can arise when simply pooling separate tables of hiring rates for different jobs are discussed in *Caston* v. Duke University, 34 FEP Cases 102, 107 (N.D. N.C. 1983) (plaintiffs applicant flow statistics rejected as it pooled data for several jobs rather than considering qualifications). An example from the *Caston* opinion is reanalysed in Gastwirth (1988) at 235. The amicus brief filed by the Equal Employment Advisory Council, at 4, pools the results of both examinations to determine the pass rate of each race–ethnic group. Then, it calculates the ratios of those pass rates to show the results of the examinations violated the four-fifths rule. Later, at 13, the brief describes the disparity as statistically significant but does not refer to a statistical test.
requirement that depended on the number of immediately available positions. As seen in Section 4.3, only when one interpretation of this new criterion is added, i.e. one of the three groups is not represented on both lists, would the test results reach statistical significance at the 0.05 level. The city also seemed to require that the lists of promotable candidates for each position should contain at least one member from the three race–ethnic groups. Two fair tests for both positions would yield examination results that failed this ‘new criteria’ over 30% of the time. When New Haven assessed the results of both examinations together, without considering the likely demographic mix of future appointments likely to be made from them, it did not realize that it probably would be able to reduce the disparate impact of the lieutenant examination. Unfortunately, the briefs submitted to the Supreme Court do not discuss these statistical issues or the statistical effect of a governmental agency’s adding or modifying a previously set standard or guideline without announcing it beforehand.

Our reanalysis of the data in the case and a related one indicate that it may well be appropriate to use significance testing at the more liberal 0.10 significance level, rather than the 0.05 one, in cases involving small sample sizes for a single position. It is also reasonable to allow an employer to take remedial steps before an internal examination of its hiring or promotion data reaches the 0.05 level. An advantage of statistical testing over allowing an employer to determine whether there is a ‘substantial disproportion’ by just looking at the numbers, as apparently advocated in one amicus brief, is that it replaces a subjective standard by an objective one. This should assist in ensuring a more uniform application of the law. Along with the p value or level of statistical significance of a test applied to the data, the power of the statistical test to detect a meaningful alternative should also be considered. Given the importance of the issues raised in Ricci and the need to conduct a rigorous statistical analysis, the government should review the Uniform Guidelines and provide more detailed guidance to both employers and courts. Recommendations about the significance levels appropriate for the analysis of data of different sample sizes and numbers of hires or promotions would systematize the examination of success rate data.

Finally, it should be noted that often information on characteristics that are related to an applicant’s performance on the examination, e.g. educational background, amount of prior training or experience, is available. These factors can be incorporated into the statistical analysis by stratifying the data into comparable subgroups. A measure of the difference in pass rates between the

---

59 In Motor Vehicles Manufacturer’s Assoc. v. Environmental Protection Agency (EPA) 768 F. 2d 385 (DC. Cir. 1985), the court did not allow the EPA to weaken the statistical standard for assessing whether two fuels led to cars producing equivalent levels of pollutants to allow a new fuel to ‘pass’. The data from the case are discussed by Gastwirth (1988, pp. 611–619) and Finkelstein and Levin (2001, pp. 187–188) (noting the complications arising from the EPA’s use of several different statistical tests and the lack of power of the tests in the small sample size (16) in the case).

60 The reason for this is that the power of a statistical test, which is the probability it would detect a meaningful disparity depends on the sample size (the number of test passers or the number of promotions made) and the proportion of minorities among the applicants. In small samples, especially when the minority fraction of the candidates is substantially less than one-half, the power of a test can be quite low and the only way to increase the power of a test is to increase the type I error or cut-off for statistical significance. See Gastwirth (1988) at 180–184 for an illustrative example and Goldstein (1985) for a discussion of the use of power considerations in the analysis of equal employment data.

61 While the number of hires or promotions as well as the demographic mix of the eligible pool affect the determination of statistical significance, it is reasonable for employers to check for fair treatment if the p value of an appropriate statistical test reaches 0.10, say. Then, they can check to see whether the protected group members were less qualified or were impeded by their employment practices.

62 See amicus brief of the ACLU at 25 n. 10.
race–ethnic groups in each stratum can be calculated and then summarized in an overall statistical test to assess its statistical significance.\textsuperscript{63}

Acknowledgements

The authors wish to thank the staff of the Office of the Clerk of the Court for their help in providing access to the materials on file and cited in the article and Prof. E. Bura for her thoughtful comments. The comments of three referees and the editor improved the paper considerably. Neither author served as an expert or consultant to the parties in the case or to the individuals or organizations who filed an \textit{amicus} brief.

REFERENCES


\textsuperscript{63} A commonly used approach when the pass or promotion rates of one protected group is compared to that of the unprotected group is the Cochran–Mantel–Haenszel test described in Gastwirth (1984) and Finkelstein and Levin (2001). Generalizations of it to accommodate more than two groups can be used. One approach uses a logistic regression incorporating these factors as covariates.


Appendix. The questions presented to the court

This case presents recurring issues regarding proper application of Title VII and the Equal Protection Clause to the civil service. Petitioners, New Haven firefighters and lieutenants, qualified for promotion to command positions pursuant to job-related examinations and merit selection rules mandated by local law. Citing the race of the successful candidates and Title VII’s ‘disparate impact’ provision, city officials refused to promote the petitioners.

1. When an otherwise valid civil service selection process yields unintended racially disproportionate results, may municipalities reject the results and the successful candidates for reasons of race absent the demonstration required by 42 U.S.C. § 2000e-2(k)?

2. Does 42 U.S.C. §2000e-2(l) which makes it unlawful for employers ‘to adjust the scores of, use different cutoff scores for, or otherwise alter the results of, employment-related tests on the basis of race ... ’, permit employers to refuse to act on the results of such tests for reasons of race?

3. If citing the public interest in eradicating political patronage, racism and corruption in civil service, a state’s highest court mandates strict compliance with local laws requiring race-blind competitive merit selection procedures, does 42 U.S.C. §2000e-7 permit federal courts to relieve municipalities from compliance with such laws?

Note added in proof: On 29 June 2009, the U.S. Supreme Court decided the Ricci case in favour of the plaintiffs. The majority opinion stated that, by itself, the city’s fear of being sued for the disparate impact of the examination on minority applicants cannot justify New Haven’s reliance on race to the detriment of individuals who passed the examination and qualified for promotion. While the difference in pass rates would help minority plaintiffs establish a prima facie case, the majority said that before an employer can engage in intentional discrimination to remedy an unintentional disparate impact, it must have a strong evidentiary basis that it will be subject to disparate impact liability if it fails to take the race-conscious discriminatory action. Thus, the plaintiffs prevailed on their disparate treatment claim.

Four justices dissented. They felt that an employer can jettison a selection procedure when its disproportionate racial impact becomes apparent subject to an important condition: The employer must have good cause to believe that the procedure would not meet the criteria for business necessity or job relatedness.

The opinion is available on the Court’s website (last visited on 30 June 2009): www.supremecourtus.gov/opinions/08pdf/07-1428.pdf