Did the Results of Promotion Exams Have a Disparate Impact on Minorities? Using Statistical Evidence in *Ricci v. DeStefano*

Weiwen Miao
Haverford College

Journal of Statistics Education Volume 19, Number 1 (2011)
www.amstat.org/publications/jse/v19n1/wilson.pdf

Key Words: Discrimination; Disparate impact; Legal case data; *Ricci v. DeStefano*; Guided senior project

Abstract

This paper shows how to use data from the *Ricci v. DeStefano* case in statistics courses. The *Ricci v. DeStefano* case was about disparate impact of firefighters' promotion exams in New Haven, Connecticut. A statistical analysis of the test scores of both Lieutenant and Captain exams indicates that there is significant difference between the average test scores of minority and majority applicants. Analysis of the passing rates and the rates of potential promotion to the Captain position, however, does not show significant difference. This apparent contradictory result shows students that in real situations, different ways of analyzing data can lead to completely different conclusions. During the trial, the court used the government's "four-fifths rule" or guideline to reach its decision. The paper also presents a guided senior thesis project to assess the statistical soundness of this "four-fifths rule". The analysis reinforces a previous study that showed that the "four-fifths rule" guideline was not appropriate for the data in the *Ricci* case.

1. Introduction

Statistical educators have long advocated the use of real-life data in classrooms (e.g., Cobb 1992). Real data from biology, psychology, sports, etc. are now used in both examples and exercises in recent textbooks (e.g., Chance and Rossman 2005; Devore and Peck 2007; Moore et al. 2009; Utts and Heckard 2002). However, examples from actual legal cases are relatively rare. As pointed out by Gastwirth (2000), the last thirty years has seen an increasing use of scientific evidence, and in particular, statistical evidence, by the legal system. In *Castaneda v. Partida* (1977), the U.S. Supreme Court decided that formal statistical comparison of the proportion of minorities in the population eligible for jury service and the proportion actually called should be used in equal protection cases. After 1977, the use of statistical evidence in discrimination cases became commonplace. Both statisticians and law professors have developed scholarly literature in this area (e.g., Aitken and Taroni 2004; Fienberg 1988; Finkelstein 1980; Freidlin and Gastwirth 2000; Kadane 1990, 2005; Kaye 1982; Kaye and Aickin 1986; Gastwirth 1988; Gastwirth 1997; and Gastwirth and Miao 2002).

The *Ricci v. DeStefano* case was a "reverse discrimination" suit brought against the City of New Haven by eighteen firefighters who achieved high scores on promotion examinations in New Haven's fire department. The City of New Haven invalidated the test results, because an insufficient number of minorities would be promoted to an *existing* position, although the test results would be used for 2 years. Gastwirth and Miao (2009) provides detailed description of the case. This paper demonstrates how to use the data from the *Ricci v. DeStefano* case in an introductory statistics class and as a guided senior project. In an introductory statistics class, the data can be used in several ways to illustrate different statistical methods, it can also be used in a review session at the end of a semester or as a mini-project of data analysis with *guided* questions. Analysis of the test scores of both the Lieutenant and Captain exams shows that the average test scores for the three races are significantly different, while analysis of pass rates as well as rates for potential promotion to the Captain position does not come close to statistical significance. This contradictory result demonstrates that different ways to analyze data may lead to different conclusions.

During the trial, the court used the government's "four-fifths (80%) rule" to compare the pass rates of blacks and Hispanics to that of whites (see Section 3.1.2 for detailed description). That is, the court calculated the p_B/p_W and p_H/p_W, and then compared those ratios to the 80%, where p_B, p_H, p_W are the pass rates for blacks, Hispanics and whites, respectively. Both ratios are less than 80%, indicating disparate impact on minorities according to the government's "four-fifths (80%) rule" from the Uniform Guideline (29 C.F.R. 1607(D), 2000). As a guided senior project, students can use the *Ricci* case data to assess statistical soundness of this "four-fifths rule". Simulation results show that more than 80% of

the time, a *fair* test for either Lieutenant or Captain exam would fail the government's "four-fifths rule". Furthermore, for the *Ricci* case situation, the "four-fifths rule" would fail for *all possible selections* for the Lieutenant position. In other words, no matter how the City did its promotion, the "four-fifths rule" would fail. For the Captain position, the "four-fifths rule" would be satisfied for only 1 possible selection (out of 53). This analysis strongly suggests that the government should use formal statistical tests in situations similar to *Ricci*.

The paper is organized as follows: Section 2 provides a brief introduction to the *Ricci v. DeStefano* case and the related data set. The use of the *Ricci v. DeStefano* data in an introductory statistics course and as a guided senior project are presented in Section 3. Section 4 contains some discussion.

2. The Story of *Ricci v. DeStefano*

2.1 Brief History of the Case

In November and December of 2003, the New Haven Fire Department administered oral and written exams for promotion to Lieutenant and Captain. Under the contract between the City of New Haven and the firefighter's union, the written exam received a weight of 60% and oral exam received a weight of 40%. Applicants with a total score of 70% or above pass the exam and become eligible for promotion. A total of 118 firefighters took the exam. Among them, 77 took the Lieutenant exam, and 41 took the Captain exam. For the Lieutenant exam, 6 out of 19 (31.6%) blacks, 3 out of 15 (20%) Hispanics, and 25 out of 43 (58.1%) whites passed the exam. For the Captain exam, 3 out of 8 (37.5%) blacks, 3 out of 8 (37.5%) Hispanics, and 16 out of 25 (64%) whites passed the exam (see Table 1). Obviously, the whites had the highest pass rates in both the Lieutenant and Captain exams. At the time of the exam, there were 8 Lieutenant and 7 Captain positions available. The City Charter of New Haven specifies that when "g" promotions are made, the Department must select them from the top $g+2$ scorers. For the Lieutenant exam, all top 10 scorers were whites, and for the Captain exam, the top 9 scorers included 7 whites and 2 Hispanics. It appeared that no blacks would be promoted to either Captain or Lieutenant position, and at most 2 Hispanics would be promoted to Captain position. Furthermore, the eligibility list was to remain valid for 2 years. During the two-year period, a total of 16 Lieutenant and 8 Captain positions became available. Consequently, the top 18 scorers for Lieutenant position and top 10 scorers for Captain position would be considered for potential promotions. The top 18 Lieutenant scorers included 3 blacks and 15 whites and the top 10 Captain scorers included 2 Hispanics and 8 whites. Table 1 lists the number of test-takers that passed the exam as well as the number of each ethnic race group that were

among the top $g+2$ positions for potential promotions. The passing percentage for each race is also given in the table.

Table 1. Data from *Ricci v. DeStefano*

Lieutenant	Pass (% pass)	Fail	Total	Top 10	Top 18
Black	6 (31.6%)	13	19	0	3
Hispanic	3 (20%)	12	15	0	0
White	25 (58.1%)	18	43	10	15
Total	34	43	77	10	18

Captain	Pass (% pass)	Fail	Total	Top 9	Top10
Black	3 (37.5%)	5	8	0	0
Hispanic	3 (37.5%)	5	8	2	2
White	16 (64%)	9	25	7	8
Total	22	19	41	9	10

The City of New Haven decided not to certify the exam and promoted no one, because an insufficient number of minorities would receive a promotion to an *existing* position. Ricci and other test-takers who would be considered for promotion had the city certified the exam sued the city for reverse discrimination. The District Court decided that the plaintiffs did not have a viable disparate impact claim because the city's canceling the exam affected all applicants equally. The court accepted the city's argument that the City made its decision on the basis of the disparate impact of the exam on minorities. The opinion from the trial court (United States District Court for the District of Connecticut, 554 F. Supp. 2d 142) used the government's "four-fifths rule" (29 C.F.R. 1607(D), 2000) to assess the disparate impact of the exam. The trial court noted that for the Lieutenant exam, the pass rates for whites, Hispanics and blacks were 58.1%, 20% and 31.6%, respectively. (The opinion reported 60.5% pass rate for whites, but the correct pass rate is 58.1%.) The adverse impact ratios are the ratios of those pass rates, i.e. $31.6\%/58.1\% = 54\%$ and $20\%/58.1\% = 34.4\%$ for blacks and Hispanics, respectively. Both adverse impact ratios are below the 80% from the Guideline, 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. Neither party submitted a formal report with a full description of the results of statistical tests.

On appeal, a three-judge panel heard arguments in this case of discrimination, and confirmed the district court's ruling in Feb, 2008 (530 F. 3d). In a 5-4 decision issued on June 29, 2009, the Supreme Court decided that the City's failure to certify the tests was a violation of Title VII of the Civil Rights Act of 1964 (129 S. Ct. 2658).

2.2 Data Set Arising from the Case

The data set presented in the Appendix contains the oral, written and combined test scores, together with the race and position for each test taker. There are 5 variables and 118 observations in the data. The variables are:

- Race: Race of each test-taker. W = white, H=Hispanic and B = black;
- Position: Captain or Lieutenant;
- Oral: Oral exam scores;
- Written: Written exam scores;
- Combine: Weighted total scores, with 60% written and 40% oral.

This paper concentrates on the combined test score, because the decision of passing the exam as well as potential promotion are based on this combined score.

3. Pedagogical Uses

In this section, I present how to use the *Ricci* case data in an introductory statistics course as well as a guided senior thesis project. In a one semester introductory statistics course, students perform two sets of analysis of the data: one concentrates on the actual test scores, and the other on different kinds of pass rates the courts considered. The analysis on test scores demonstrates that for both Lieutenant and Captain exams, the average test scores for the three races are significantly different. But analysis of the pass rates and rates of being among the top "g+2" positions for potential promotion for the Captain exam shows no significant difference among the three races. As a guided senior project, students assess the statistical soundness of the government's "four-fifths rule". Simulation results show that the "four-fifths rule" is not consistent with formal hypothesis testing for the *Ricci* case data. All the analysis was done by the software R, which can be freely downloaded from: http://www.r-project.org.

3.1 The Use of *Ricci* Data in an Introductory Statistics Course

The *Ricci* case data can be used several times during the semester for different topics in an introductory statistics course or can be used in a review section at the end of the semester. I used this data as a review project at the end of the semester in my introductory course. My course is designed for students with a *strong* math background, and has a two-semester

calculus prerequisite. Besides three-hour lectures per week, students also have 1.5-hour computer lab (I used software R) each week. At the beginning of the project, students were asked to analyze the data to see whether there is significant disparate impact on minorities, *without* any specific instructions (about 30 minutes). Then we discussed how to approach the problem and I posted the following *guided* questions for students to answer.

1. For the Lieutenant exam, what are the means, medians and standard deviations of the test scores for the three races? According to those summary statistics, do the three races have approximately the same average scores? How about for the Captain exam?
2. What can you say about the distributions of those test scores? (Open-ended)
3. Draw a graph to compare the Lieutenant test scores for the three races, what can you conclude? Do the same for the Captain exam.
4. Do you think the average test scores for the Lieutenant exam are the same for whites and minorities? Carry out a formal hypothesis test to answer this question. Do the same for the Captain exam scores.
5. Do you think the average test scores for the Lieutenant exam are the same for all the three races? Carry out a formal hypothesis test to answer this question. Do the same for the Captain exam scores.
6. The trial court considered the pass rates as well as the rates of being among the top "g+2" scorers. Do you think the pass rates are the same for all the three races for the Lieutenant exam? The Captain exam? How about the rates of being among the top "g+2" scorers?

Those questions provided directions for students to follow. Finally the whole class discussed what they discovered. More importantly, we also discussed what students should do when they analyze their own data. I used 2 lab periods (1.5 hours each) on this data set. Students loved this project and believed that this case study combined together all the material they learned in class. The fact that the Supreme Court was reviewing the case when they did the analysis made the class very exciting. I think this review project helped students to develop a sense of statistical thinking, which is one of the recommendations in the Guidelines for Assessment and Instruction in Statistics Education (GAISE 2005) report.

3.1.1 Analysis on Test Scores

Before conducting a formal hypothesis test on the test scores, students would have learned that the first step for data analysis is to explore the data through drawing some appropriate graphs and calculating various summary statistics (e.g., Pardoe 2008). We start with numerical summaries of the combined test scores for the three different races, and then use the side-by-side boxplot to compare those test scores for different races.

Numerical summary of the test scores

Table 2 lists summary descriptive statistics for both Lieutenant and Captain test scores, separated by race.

Table 2. Summary statistics for the *Ricci. v. DeStefano* data

Position	Race	Size	Mean	Median	Standard Deviation
Lieutenant	Black	19	63.72	61.07	9.08
Lieutenant	Hispanic	15	63.62	63.27	5.77
Lieutenant	White	43	71.84	70.73	9.15
Captain	Black	8	63.78	63.9	8.49
Captain	Hispanic	8	68.55	67.52	8.70
Captain	White	25	74.11	73.73	8.25

The table clearly indicates that for the Lieutenant exam, the whites have a higher mean score than the other two races; and the mean scores for blacks and Hispanics are about the same. For the Captain exam, the whites have the highest mean score, followed by the Hispanics and then the blacks.

> **Helpful Hint:** *Students may also notice that for both the Lieutenant and Captain exams and all the three races, the mean and median scores are roughly the same, indicating that the distributions of the scores may be symmetric. In terms of variability, for the Lieutenant exam, the standard deviations for blacks and whites are about the same, both are larger than the standard deviation of the Hispanics scores. For the Captain exam, all the three races have about the same standard deviations.*

Stem-and-leaf

The sample sizes in each position*race category are not large, and stem-and-leaf is a perfect graphic method to use for the shape of the distributions. Figure 1 shows the stem-and-leaf

graphs for the Lieutenant and Captain exams for black, Hispanic and white test takers. Clearly all the six stem-and-leaf graphs are approximately symmetric with no obvious outliers.

Lieutenant Exam

The decimal point is 1 digit(s) to the right of the |

4 | 6		
5 | 14	5 | 4	
5 | 799	5 | 579	5 | 77788
6 | 00114	6 | 1133	6 | 0023
6 | 67	6 | 5689	6 | 56778999
7 | 234	7 | 111	7 | 0001112223
7 | 667		7 | 556789
		8 | 2333444
		8 | 67
		9 | 0
Race=Black	**Race = Hispanic**	**Race = White**

Captain Exam

The decimal point is 1 digit(s) to the right of the |

5 | 1		
5 | 68	5 | 79	5 | 6
6 | 1	6 | 2	6 | 13
6 | 7	6 | 69	6 | 8999
7 | 02	7 |	7 | 001144
7 | 5	7 | 68	7 | 566678
	8 | 0	8 | 011
		8 | 59
		9 | 3
Race=Black	**Race = Hispanic**	**Race = White**

Figure 1. Stem-and-leaf graphs for Lieutenant and Captain exams for black, Hispanic and white test takers.

Side-by-side boxplots

After looking at those exam scores separately for each race, instructors can ask students to do a graphic comparison for the three different race groups. A side-by-side boxplot is a nice tool to use. Figure 2 shows the side-by-side boxplots for Lieutenant and Captain exams. Those side-by-side boxplots indicate that for both Lieutenant and Captain exams, whites have generally higher median scores than the other two race groups, and blacks and Hispanics have approximately the same median scores. (Some students may say that Hispanics have slightly higher median scores than the blacks.) The boxplot for blacks,

Lieutenant exam shows that the median is closer to its first quartile than to the third quartile; but the whisker of the lower end is longer than that of the upper end. The other five boxplots are roughly symmetric. Furthermore, it's clear that for the Lieutenant exam, the Hispanic test scores are less spread out compared to the scores for blacks and whites, which is consistent with its smaller standard deviation listed in Table 2. Those boxplots provide preliminary comparisons on test scores. In order to find out whether the differences in test scores are due to chance or to different race groups, one needs to use formal statistical hypothesis testing.

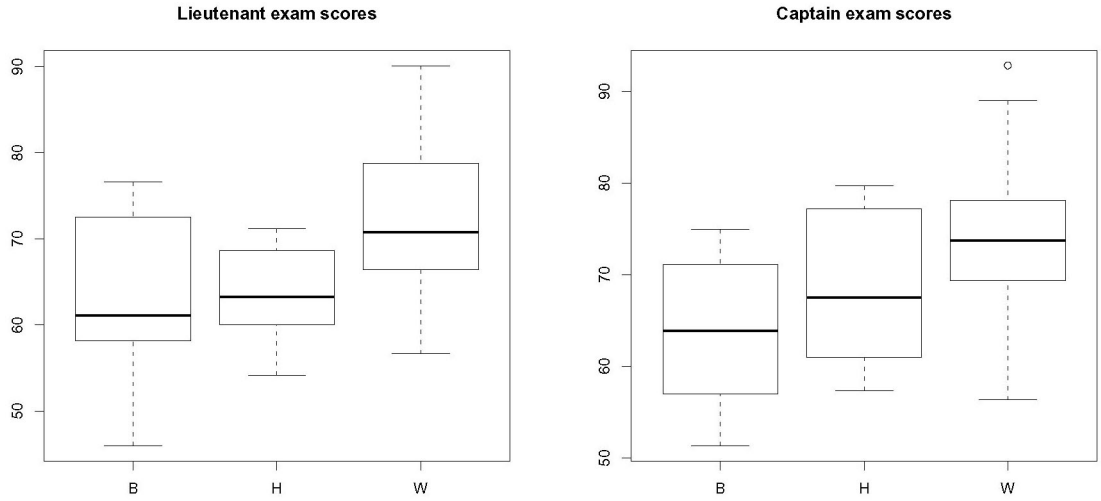

Figure 2. Side-by-side boxplots for the combined test scores

Two-sample *t*-test

Even through the actual law case concerned three race groups, instructors can combine the black and Hispanic into one minority group, and ask students to compare the average test scores for white and minority test takers. Table 3 provides the results of the Welch two sample *t*-test (two-sided) for combined test scores for Lieutenant and Captain. The *p*-values for both Lieutenant and Captain exams are less than 0.01, indicating that minorities had significantly different test scores than the whites.

Table 3. Welch two sample *t*-test on combined test scores

Position	Group	Mean	S.D.	t-value	df	p-value
Lieutenant	Majority	71.84	9.15	-4.2559	74.695	5.973e-05
	Minority	63.68	7.69			
Captain	Majority	74.11	8.25	-2.9191	30.955	0.006489
	Minority	66.16	8.66			

One-way ANOVA

For the purpose of the applications, a better way to analyze test scores is to treat blacks and Hispanics separately. Tables 4 and 5 are the ANOVA outputs for Lieutenant and Captain exams, respectively. Clearly, the average test scores for the three race groups are significantly different for both exams, as both *p*-values are less than the traditional cut-off of 0.05.

Table 4. One-way ANOVA output for Lieutenant test scores

	DF	Sum Sq	Mean Sq	F value	Pr(>F)
Race	2	1266.5	633.2	8.5789	0.0004458
Residuals	74	5462.2			

Table 5. One-way ANOVA output for Captain test scores

	DF	Sum Sq	Mean Sq	F value	Pr(>F)
Race	2	707.19	353.59	5.034	0.01150
Residuals	38	2669.15			

Potential Pitfall: *Some of my students tried the one-way ANOVA with all the 118 test scores. I.e. analyze the Lieutenant and Captain scores together. The corresponding p-value is 5.014e-6, showing that the test scores are significantly different for different races.*

Alternative Applications: *In my introductory course, I only covered the F-test in one-way ANOVA. However, after the H_0 is rejected in one-way ANOVA, students can perform multiple comparisons to find out which pairs of means differ, for both Lieutenant and Captain exams. Those comparisons can also lead to the discussion of possible interaction between Race and Position and the two-way ANOVA.*

1. **Multiple Comparison:** *The following table gives the Tukey's 95% confidence intervals and the p-values adjusted for multiple comparisons.*
   It's clear that for the Lieutenant exam, the differences between whites and blacks, and whites and Hispanics are significant. But the difference between blacks and Hispanics is not significant. For the Captain exam, the difference between whites and blacks is significant, but the differences between Hispanics and blacks, and whites and Hispanics, are not significant.

Table 6. *Tukey's 95% confidence interval and adjusted p-values*

Position	*Group*	*Diff*	*Lwr*	*Upr*	*p-adj*
Lieutenant	H-B	-0.0927	-7.19	7.00	0.9995
	W-B	8.126	2.47	13.79	0.002786
	W-H	8.219	2.06	14.38	0.005849
Captain	H-B	4.765	-5.46	14.98	0.4977
	W-B	10.331	2.03	18.63	0.01178
	W-H	5.566	-2.74	13.87	0.2436

2. **Interaction Graphs:** *The interaction graph on the left-hand side in Figure 3 clearly shows that for the Captain exam, whites have the highest average scores, followed by Hispanics, and blacks have the lowest scores. But for the Lieutenant exam, the pattern is different: blacks and Hispanics have about the same average scores, their scores are lower than that of the whites. This graph was made in R, using the default settings. Some students might think that there is some interaction between the Race and Position. However, the p-value for Race*Position is 0.6447 (Table 7), indicating that there is no interaction between Race and Position. Students may wonder how to interpret those results. This is a great opportunity to reinforce the concept of making sensible graphs. Instructors can point out that the range of the average test scores is from 0-100. But the y-axis range for graph on the left-hand side in Figure 3 is approximately from 63-74. In this narrow range, a small difference appears large and hence makes the possible interaction look stronger. The interaction graph on the right-hand side in Figure 3, where test score range is set to (0,100), does not show a significant interaction between Race and Position. This graph is more consistent with the p-value of 0.6447 reported in Table 7. On the other hand, the sample sizes for Captain exam are very small. Consequently, the power of the interaction test is low. We do not know whether or not there is an interaction. However, a p-value of 0.6447 is not evidence for interaction. Figure 4 presents another set of interaction graphs.*

3. **Two-Way ANOVA**: *Table 7 is the two-way ANOVA output. The table clearly indicates that the average test scores for three ethnic races are significantly different. However, the difference between the two positions is not significant, and there is no interaction.*

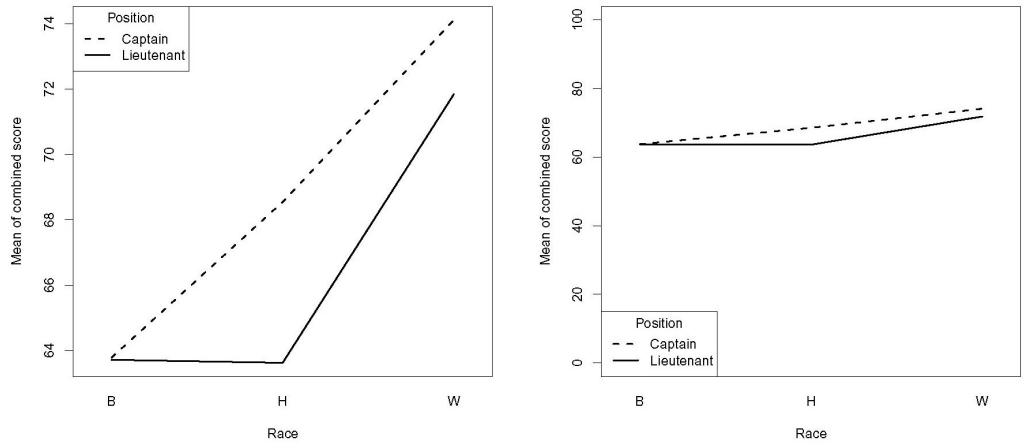

Figure 3. Interaction Graphs for Race*Position. The Left Graph has Approximately Y-axis Range from (63, 75). The Right Graph has Y-axis Range from (0,100).

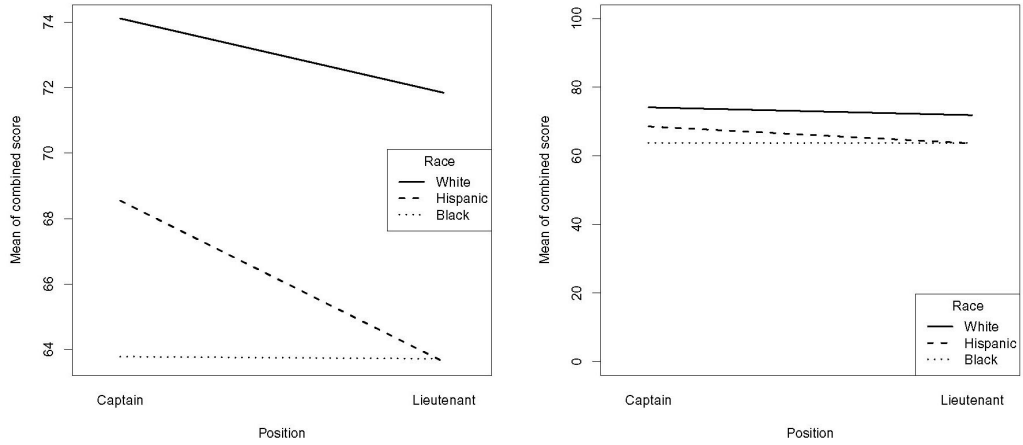

Figure 4. Interaction Graphs for Race*Position. The Left Graph has Approximately Y-axis Range from (63, 75). The Right Graph has Y-axis Range from (0,100).

Table 7. *Two-way ANOVA output with interaction*

	DF	*Sum Sq*	*Mean Sq*	*F value*	*Pr(>F)*
Race	2	1749.4	874.7	12.048	1.82e-05
Position	1	121.7	121.7	1.6766	0.1980
Race:Position	2	64.0	32.0	0.4407	0.6447
Residuals	112	8131.3	72.6		

The above analysis on the actual test scores shows that for *both* Lieutenant and Captain exams, the average test scores for the three races are significantly different. Of course, there are other variables that also influence test scores, such as educational background,

experience and how devoted applicants are to preparing for the exam. Ideally, information on those variables should also be included in the formal analysis.

3.1.2 Analysis on Passing Rates and Rates of Top $g+2$ Scorers

In the actual legal case, the City of New Haven and the trial court treated blacks and Hispanics separately and compared rates for different races instead of the actual test scores. Furthermore, in the past, the City of New Haven adjusted the test scores in order to increase the fraction of applicants whose scores reached 70 or more (Gastwirth and Miao 2009). In other words, what the City and the court were really concerned with was the rates of top $g+2$ scorers for the three races. From a statistical viewpoint, they considered the disparate impact of the tests by looking at the first two rates listed below:

1. **Pass rate**: the proportions of applicants from the three race-ethnic groups scoring 70 or above;
2. **Top $g+2$ rates, existing positions**: the proportions of applicants from the three race-ethnic groups being among the top $g+2$ scorers from whom the *immediately* available promotions would be made;
3. **Top $g+2$ rates, 2-year period**: the proportions of applicants from the three race-ethnic groups being among the top $g+2$ scorers from whom the vacant positions would be filled during *the two year period* when the exam results would be used.

Chi-square test

When the different rates are considered, the data constitute a 2-way contingency table with Race*Pass. The chi-square test can be used to test whether the rates are the same for all the three groups. In this subsection, I consider three race groups: black, Hispanic and white. Table 8 shows the p-values of the chi-square test on pass rates and the top $g+2$ rates for both the existing positions and the available positions during the two-year period. For the Lieutenant exam, all the p-values are smaller than 5%, indicating that the pass rates and the top g+2 rates are significantly different for different races. But for the Captain exam, all the p-values are higher than 0.15, meaning that for the Captain exam, there is no statistical evidence of discrimination for all the three rates considered. This result is different than the one using actual test scores.

> **Helpful Hint:** *In Table 8, five out of six chi-square approximations may be incorrect due to small expected values. Software R provides simulated p-values*

Table 8. *P-values for chi-square test*

	Pass rates	Top g+2 rates (existing positions)	Top g+2 rates (two-year period)
Lieutenant	0.01675	0.01064*	0.01526*
Captain	0.2522*	0.2433*	0.1857*

*Some expected numbers are small, chi-square approximation may be incorrect.

with the command simulate.p.value=TRUE. The simulation is done by random sampling from the set of all contingency tables with given marginals, and works only if the marginals are strictly positive. Table 9 reports one set of simulated p-values for the chi-square test.

Table 9. *P-values for chi-square test*

	Passing rates	*Top g+2 rates* (*existing positions*)	*Top g+2 rates* (*two-year period*)
Lieutenant	0.01675	0.01149*	0.01399*
Captain	0.2594*	0.3123*	0.3038*

**Simulated p-values*

Alternative Application: *When the top g+2 rates are considered, the appropriate statistical test is the Fisher-Freeman-Halton test or FFH test (Freeman and Halton 1951). This is because the total number of promotions, $g+2$, is fixed. As this test is not covered in the usual introductory statistics course, I used the chi-square test in my introductory statistics course. However, the two-sided p-values of the Fisher-Freeman-Halton test are given in Table 10. The R commands and a brief description of the process used to calculate the p-values of the FFH test are given in Appendix 5.3.*

Table 10. *P-values for the Fisher-Freeman-Halton test (two-sided)*

	Top g+2 rates (existing positions)	*Top g+2 rates (two-year period)*
Lieutenant	0.00769	0.01012
Captain	0.3177	0.2968

The "four-fifths rule"

The trial court used the so-called "four-fifths rule" from the Uniform Guideline (29 C.F.R. 1607(D), 2000). The following is an excerpt from the Guideline:

> "A selection rate for *any* race, sex, or ethnic group which is less than four-fifths (4/5) (or 80%) 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…".

Table 11. The "four-fifths rule" ratios for the different rates considered

	Passing rates	Top g+2 rates (existing positions)	Top g+2 rates (two-year period)
Lieutenant	34.4	0	0
Captain	58.6	0	0

Table 11 lists the ratio of the lowest rate to the highest rate for all the three rates considered. For example, consider the Captain position, top $g+2$ rates for existing positions. According to Table 1, seven out of 25 whites, 2 out of 8 Hispanics and 0 out of 8 blacks were among the top 9 scorers. Consequently, the top $g+2$ rates for whites, Hispanics and blacks are: $7/25 = 0.28$, $2/8 = 0.25$, and $0/8 = 0$, respectively. The ratio of the lowest rate to the highest rate is $0/0.28 = 0$. The other ratios are calculated similarly. Note that none of the ratios is higher than the 80%. Hence the "four-fifths rule" from the Government Guideline shows

disparate impact on *both* the Lieutenant and Captain exams. However, formal statistical analysis on the Captain exam shows no evidence of disparate impact on minorities. Those results indicate that using the "four-fifths rule" here is not appropriate.

Alternative Application: *Although the previous analyses on the different rates were done with three races (black, Hispanic and white), the same analysis can be carried out to compare majority v. minority. Table 12 reports p-values of the chi-square test, the Fisher's exact test (Agresti 2002) and the "four-fifths rule" ratios for the three rates considered. For completeness of the table, the p-values for both the chi-square test and the Fisher's exact test are reported. Even though the pass rate should be analyzed by the chi-square test, and analyzing the top $g+2$ rate should use the Fisher's exact test. The conclusions here are the same as the analysis on three races.*

Table 12. *The p-values for the chi-square test, the Fisher's exact test and the 4/5 ratio*

		Pass rates	*Top g+2 rates (existing positions)*	*Top g+2 rates (two-year period)*
chi-square test	*Lieutenant*	0.0055	0.002574	0.007295
	Captain	0.09694	0.2421	0.1561
Fisher's exact test	*Lieutenant*	0.00633	0.001867	0.01321
	Captain	0.1197	0.4409	0.2654
Four-fths ratio	*Lieutenant*	0.4553	0	0.2529
	Captain	0.5859	0.4464	0.3906

Probability calculation

As the number of minorities belonging to the top $g+2$ positions is small, instructors can also use this data to ask students to calculate the probability of obtaining data "as extreme or more extreme" when the selection is actually done randomly. This corresponds to the following questions:

1. Randomly choose 10 applicants from 19 blacks, 15 hispanics and 43 whites, what is the probability that all 10 chosen are whites? (Answer: 0.0017)
2. Randomly choose 18 applicants from 19 blacks, 15 hispanics and 43 whites, what is the probability that 3 or fewer minorities are selected? (Answer: 0.0065)

3. Randomly choose 9 test takers from 8 blacks, 8 Hispanics and 25 whites. What is the probability that 2 or fewer minorities are chosen? (Answer: 0.2199)

4. Randomly choose 10 test takers from 8 blacks, 8 Hispanics and 25 whites. What is the probability that 2 or fewer minorities are chosen? (Answer: 0.1478)

Those probabilities confirm the conclusions from formal hypothesis testing that the Lieutenant exam shows significantly different rates for three ethnic races; but the rates for the Captain exam are not significantly different.

3.2 Guided Senior Thesis

The *Ricci* data can also be used as a senior project. For example, it can be used to learn the Fisher-Freeman-Halton test (FFH test). Another direction is to study when the "four-fifths rule" from the Government Guideline is consistent with formal statistical testing.

Section 3.1 shows that (Table 11) for all the three rates considered, the "four-fifths rule" from the Government Guideline are violated for both Lieutenant and Captain exams. But the formal statistical tests and the probability calculation demonstrate that for the Captain exam, the three rates considered are not statistically significant, which contradicts the results from the "four-fifths rule". As also pointed out in the Guideline, the rule depends on the sample sizes involved. Students can assess how this "four-fifths rule" works in the *Ricci* case situation. There are several different ways to approach the problem.

1. For the given sample sizes and given the overall pass rate (or overall top g+2 rate), how likely is it that a *fair* test would fail the "four-fifths rule"?

   This question can be answered by the following simulation: Assume that all three race groups have the *same* test score distributions. Consider the pass rates for Lieutenant, for example. The overall pass rate was $\frac{34}{77} = 0.4416$. (The top $g+2$ rate was $\frac{10}{77} = 0.1299$.) Samples of sizes 43, 19 and 15 were selected from the *same* standard normal distribution and scores that are above the $1 - 0.4416 = 0.5584$th percentile of the standard normal passed the exam. Then the pass rate (or the top g+2 rate) for each ethnic race group is calculated and the ratio of the minimum pass rate (or the top g+2 rate) to the maximum pass rate (or the top g+2 rate) is obtained. If the ratio is less than 0.8, the "four-fifths rule" is violated. Table 13 shows the percentage of the time a *fair* test violates the "four-fifths rule" based on 10^4 simulations. Amazingly, for the Lieutenant exam, 81% of the time a *fair* test would fail the "four-fifths rule" in the pass rate and 100% of the time a *fair* test would fail the "four-fifths rule" when considering the top $g+2$ rates for both the existing positions and the possible

positions during the two-year cycle of the exam. For the Captain exam, more than 80% of the time the "four-fifths rule" is violated for all the three rates considered. This simulation shows that the "four-fifths rule" from the Government's Guideline is not appropriate to use in the *Ricci* case data. Formal hypothesis testing should be used.

Table 13. Percentage of the simulations in which a *fair* test fails the four-fifths rule

	Passing rates	Top g+2 rates (existing positions)	Top g+2 rates (two-year period)
Lieutenant	0.8147	1	1
Captain	0.8233	0.8803	0.8751

2. The second way to assess the consistency between the formal hypothesis testing and the government's "four-fifths rule" for the situation in the *Ricci* case is the following: given that there were 19 blacks, 15 Hispanics and 43 whites who took the Lieutenant exam, how many Lieutenant positions are needed to guarantee that a *fair* test will fail the "four-fifths rule" at most 5% of time?

   Students can tackle this problem by simulating the percentage of the time a *fair* test will fail the "four-fifths rule" for a given number of positions. There are at least two equivalent ways to approach this problem. (a) Samples of sizes 19, 15 and 43 are selected from the *same* normal distribution. For the given g available positions, find the race of the top $g+2$ scorers. Let b, h, w be the number of black, Hispanic, and white top $g+2$ scorers, respectively ($b+h+w=g+2$). Let $r_{min} = min(b/19, h/15, w/43)$ and $r_{max} = max(b/19, h/15, w/43)$ be the minimum and maximum top g+2 rate. Calculate the ratio, $R = r_{min}/r_{max}$, of the minimum to the maximum rate. Repeat the simulation 10^4 times to obtain the percentage of the time that the ratio $R < 80\%$, i.e. the percentage of the time that a *fair* test fails the "four-fifths rule". (b) Just randomly select $g+2$ test-takers from 19 blacks, 15 Hispanics and 43 whites. For each selection, find out whether the ratio, $R = r_{min}/r_{max}$, of the minimum to the maximum top g+2 rate is less than 80%. Table 14 reports the results.

   Table 14 shows that only when the available positions are over 68 for Lieutenant and 37 for Captain, will a *fair* test fail the "four-fifths rule" less than 5% of time. Recall that 77 and 41 applicants took the Lieutenant and Captain exams, respectively. This means that in the *Ricci* situation, only when the numbers of potential promotions are close to the number of test-takers, will a *fair* test satisfy the "four-fifths rule" more than 95% of the time, which is certainly not realistic. This also shows that for the sample sizes given in the *Ricci* case, the "four-fifths rule" does not work.

Table 14. Percentage of the simulations that a **fair** test fails the 4/5 rule for a given # of available positions

Lieutenant Exam							
# of available positions	10	50	65	66	67	68	69
test score selection	0.9136	0.5719	0.1687	0.1279	0.0809	0.0539	0.0306
random selection	0.9123	0.5748	0.1763	0.1288	0.0852	0.0536	0.0337
Captain Exam							
# of available positions	8	10	30	35	36	37	38
test score selection	0.8764	1	0.571	0.3271	0.1859	0.0683	0
random selection	0.8759	1	0.5672	0.3313	0.19	0.0681	0

Potential Pitfall: *Students may wonder why for the Captain exam, the percentage under 10 is 1, higher than the percentage under 8. This is due to the sample sizes involved. Let b, h, and w be the number of black, Hispanic and white top* $g+2$ *scorers, respectively. Then the top* $g+2$ *rates for blacks, Hispanics and whites are:* $b/8, h/8$ *and* $w/25$*. The condition that* $min(b/8,h/8,w/25)/max(b/8,h/8,w/25) \geq 0.8$ *is equivalent to the following three inequalities:* $1.25 \geq b/h \geq 0.8$*;* $0.4 \geq b/w \geq 0.256$ *and* $0.4 \geq h/w \geq 0.256$*. When the number of available positions is 10, the City chooses from the top* $12 = 10+2$ *scorers. It can be shown that there does* ***not*** *exist positive integers b, h, and w, such that b+h+w=12, and the three inequalities are satisfied. But when the available positions is 8, there exists positive integers b, h, and w, such that* $b+h+w = 10 = 8+2$ *and the 3 inequalities are satisfied. For example,* $(b,h,w) = (2,2,6)$ *is a possibility.*

3. A third way to assess the consistency between the formal hypothesis testing and the government's "four-fifths rule" for the situation in the *Ricci* case is the following: given that 19 blacks, 15 Hispanics and 43 whites took the Lieutenant exam, and that there were 8 positions available, how would the top 10 (8+2) test-takers be chosen (in terms of race), so that both the *p*-value of the Fisher-Freeman-Halton test (FFH test) is larger than 0.05 and the "four-fifths rule" is satisfied? When will the two criteria give different conclusions?

   Table 15 shows the possible ways to fill the top 10 Lieutenant positions. Note that in the table, $0 \leq b \leq 10$, $0 \leq h \leq 10$ and $0 \leq b+h \leq 10$. Let $k = b+h, k = 0,1,\ldots,10$. For any given k, there are $k+1$ possible ways to arrange the (b,h) values $((0,k),(1,k-1),\ldots,(k,0))$. Hence for Lieutenant, there are $1+2+\ldots+11 = 66$ possible tables. For each possible table, students can check whether the government's "four-fifths rule" is satisfied, whether the *p*-value for the FFH test is less than the traditional

0.05. The results are reported in Table 16.

Table 15. Possible ways to fill the top 10 Lieutenant positions

	Black	Hispanic	White	Total
Top 10	b	h	10-b-h	10
Not in top 10	19-b	15-h	33+b+h	67
Total	19	15	43	77

Table 16 shows that 40 out of 66 tables fail the "four-fifths rule" and also have the FFH test p-values less than 0.05. The rest of the 26 tables fail the "four-fifths rule" and the p-values of the FFH test are bigger than 0.05. Furthermore, of the 66 possible ways to choose the top 10 scorers, none of them would pass the "four-fifths rule". In other words, no matter how those top 10 scorers are chosen, the "four-fifths rule" will be violated. This calculation also confirms the simulation results given in Table 13. The result suggests that for the given sample sizes and the given number of available positions, the "four-fifths rule" should not be used.

Table 16. Consistency between the "4/5" rule and the FFH test for Lieutenant exam

	FFH test $p < 0.05$	FFH test $p \geq 0.05$	Total
Fail 4-5 Rule	40	26	66
Pass 4-5 Rule	0	0	0
Total	40	26	66

For the Captain exam, there are $(1+2+\ldots+10)-2 = 53$ possible ways to select the top 9 scorers, as there are only 8 blacks and 8 Hispanics who took the Captain test. Table 17 indicates that 35 out of 53 tables the FFH test and the "four-fifths rule" give consistent results. Note that of those 53 possible ways, only 1 situation satisfies both the government's "four-fifths rule" and has FFH test p-value higher than 0.05. That's the case when the top 9 scorers include 2 blacks, 2 Hispanics and 5 whites. (See Table 18.) In this case, the top g+2 rates for black, Hispanic and white are 25%, 25% and 20%, respectively, with the $R = 0.2/0.25 = 0.8$. This barely passes the "four-fifths rule". The p-value for the FFH test is equal to 1.

Potential Pitfall: *Students may wonder whether Table 17 is consistent with Table 13. As Table 17 indicates that there is only 1 out of 53 possible ways to choose the top 9 applicants for Captain position that satisfies*

the "four-fifths rule", that's about $1/53 \approx 2\%$. *But Table 13 shows that the probability that a* ***fair*** *test fails the "four-fifths rule" is about 88.03%. It's not easy for students to realize that probabilities of obtaining different selections of top 9 scorers are different. It would be interesting to ask students to calculate the probability of getting 2 blacks, 2 Hispanics and 5 whites from a* ***random*** *selection (out of 8 blacks, 8 Hispanics and 25 whites). This probability is 0.119. In other words, the probability of obtaining Table 18 via random selection is 0.119, which is consistent with the 88% given in Table 13.*

Table 17. Consistency between the "4/5" rule and the FFH test for Captain exam

	FFH test $p < 0.05$	FFH test $p \geq 0.05$	Total
Fail 4-5 Rule	34	18	52
Pass 4-5 Rule	0	1	1
Total	34	19	53

Table 18. Situation satisfies both the "4/5" rule and has FFH test p-value higher than 0.05

	Top 9	Not in Top 9	Total
Black	2	6	8
Hispanic	2	6	8
White	5	20	25
Total	9	32	41

Tables 16 and 17 strongly suggest that for the situation like *Ricci*, the government's "four-fifths rule" should not be used.

Alternative Application: *The guided senior thesis presented here treats blacks and Hispanics separately. The same analysis can be carried out by comparing majority v. minority. The following three tables are the corresponding results.*

Table 19. *Percentage of the time that a **fair** test fails the 4/5 rule (Majority v. Minority)*

	Pass rates	*Top g+2 rates (existing positions)*	*Top g+2 rates (two-year period)*
Lieutenant	0.4169	0.7463	0.5974
Captain	0.4603	0.7218	0.7101

Table 20. *Percentage of the time that a **fair** test fails the 4/5 rule for a given # of available positions (Majority v. Minority)*

Lieutenant Exam						
# of available positions	10	30	50	55	60	61
test score selection	0.7554	0.4873	0.2197	1192	0.0745	0.0351
random selection	0.7535	0.4795	0.2161	0.1072	0.0815	0.0329
Captain Exam						
# of available positions	8	10	30	32	34	35
test score selection	0.706	0.7336	0.1226	0.0888	0.0689	0.0175
random selection	0.7135	0.7246	0.117	0.0867	0.0649	0.0175

Table 21. *Consistency between the "4/5" rule and the Fisher's exact test (Majority v. Minority)*

		Fisher's exact test $p < 0.05$	*Fisher's exact test* $p \geq 0.05$	Total
Lieutenant	*Fail 4-5 Rule*	5	5	10
	Pass 4-5 Rule	0	1	1
Captain	*Fail 4-5 Rule*	4	5	9
	Pass 4-5 Rule	0	1	1

4. Conclusion

The paper demonstrates how to use the data set from the *Ricci v. DeStefano* case in statistics courses. In an introductory statistics course, analysis can be done on both the test scores and the different rates. Analysis on actual test scores shows that the average test scores for different races are significantly different, for both Lieutenant and Captain exams. But the courts considered the pass rates as well as the rates of top $g+2$ positions. Formal analysis on those three rates shows that for the Lieutenant exam, the rates for three ethnic race groups are significantly different. But for the Captain exam, none of the rates is significantly different. In other words, for the Captain exam, formal analysis on test scores and different rates give different results. This naturally leads to the question: in terms of the disparate impact, should one consider the actual test scores or the pass rates (or top $g+2$

rates)? This can generate a lively discussion on statistics and ethics, which is usually not covered in an introductory statistics course.

As a guided senior project, students can use the *Ricci* data to assess the consistency between the government's "four-fifths rule" and formal hypothesis testing. The results strongly suggest that this rule should not be used, at least for samples of sizes comparable to those in the *Ricci* case. Formal statistical testing should be used to assist the courts to make its decisions.

Acknowledgments

The author would like to thank the editor of Datasets and Stories, the editor of the Journal of Statistics Education and the referees for helpful comments. Their suggestions helped to improve the article greatly.

5. Appendix

5.1 The Ricci Data

Please see attached files RicciData.csv and Ricci.txt.

5.2 Simulation Programs Used in the Paper

The attached file Appendix5(2) contains the programs that generate Tables 13, 14, 16, and 17. The programs that generate Tables 19, 20, and 21 are very similar and hence are not attached here.

5.3 R Commands to Obtain the *P*-value of the FFH Test

The following are the R commands to obtain the *p*-values of the FFH test given in Table 10.

```
x2=matrix(c(0,0,10,19,15,33),nrow=2, byrow=T); fisher.test(x2)

x3=matrix(c(3,0,15,16,15,28),nrow=2,byrow=T); fisher.test(x3)

y2=matrix(c(0,2,7,8,6,18),nrow=2,byrow=T); fisher.test(y2)

y3=matrix(c(0,2,8,8,6,17),nrow=2,byrow=T); fisher.test(y3)
```

Those *p*-values are calculated as follows: (1) Calculate the probability of obtaining the

observed table by random selection, call this probability L. For example, consider the "top g+2 rate" for Lieutenant, existing positions. In this case, $g+2=10$, and the L is the probability that for a random selection of 10 applicants from 19 blacks, 15 Hispanics and 43 whites, all the 10 chosen are whites. This L is: $L = \binom{19}{0}\binom{15}{0}\binom{43}{10}/\binom{77}{10} = 0.001747809$. (2) List all the possible 2*3 tables with the same margins as the observed table. (3) For each table, calculate the probability of obtaining the table by random selection. (4) The *p*-value is the sum of those probabilities that are less than or equal to the L. The software R calculates the p-values automatically. The command is `fisher.test(x)`, where x is the observed 2*3 table. The *p*-values in Table 12 are obtained in the same way.

References

Agresti, A. (2002). "*Categorical Data Analysis*", 2nd ed, Wiley.

Aitken, C.G.C. and Taroni, F. (2004). "*Statistics and the Evolution of Evidence for Forensic Scientists*", 2nd Ed, Wiley.

Castaneda v. Partida, 430 U.S. 482, 496 n. 17 (1977).

Cecil, J. (ed.) (2000), *"Manual on Scientic Evidence"* , 2nd Ed, Federal Judicial Center.

Cobb, G. (1992), "Teaching Statistics", in *Heeding the Call for Change: Suggestions for Curricular Action*, ed. L. A. Steen, MAA Notes, No. 22, Washington DC: Mathematical Association of American, 2-23.

Chance B. and Rossman A. (2005), "*Investigating Statistical Concepts, Applications and Methods*", Duxbury Press.

Devore, J. L. and Peck, R. (2007), *Statistics - The Exploration and Analysis of Data* (5th ed.), Belmont, CA: Thomson Learning.

Fienberg, S.E. (1988), *The Evolving Role of Statistical Assessments as Evidence in the Courts*, Springer-Verlag.

Finkelstein, M. O. (1980). "The judicial reception of multiple regression studies in race and sex discrimination cases", *Columbia Law Review*, 80, 737-754.

Freeman, G. H. and Halton, J. H. (1951). "Note on an exact treatment of contingency, goodness of fit and other problems of significance", *Biometrika*, 38, 141-149.

Freidlin, B. and Gastwirth, J. L. (2000) "Changepoint tests designed for the analysis of hiring data arising in employment discrimination cases", *Journal of Business & Economic Statistics*, 18, 315-322.

Gastwirth, J. L. (1988) *"Statistical Reasoning in Law and Public Policy"*, San Diego, CA: Academic Press.

Gastwirth, J. L. (1997). "Statistical evidence in discrimination cases". *Journal of the Royal Statistical Society*, Series A: Statistics in Society, 160, 289-303.

Gastwirth, J. L. (2000). *Statistical Science in the court room*, Springer.

Gastwirth J. L. and Miao, W. (2002). "The Potential Effect of Statistical Dependence in the Analysis of Data in Jury Discrimination Cases: *Moultrie v. Martin* Reconsidered", *Jurimetrics*, 43, 115-128.

Gastwirth, J. L. and Miao, W. (2009). "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. DeStephano*", *Law, Probability and Risk*, 8, 171-191.

Guidelines for Assessment and Instruction in Statistics Education (GAISE) (2005). American Statistical Association. *http://www.amstat.org/education/gaise/*

Kadane, J. B. (1990). "A statistical analysis of adverse impact of employer decisions", *Journal of the American Statistical Association*, 85, 925-933.

Kadane, J. B. (2005). " Ethical Issues in Being an Expert Witness", *Law, Probability and Risk*, 4, 21-23.

Kaye, D. (1982). "Statistical evidence of discrimination", *Journal of the American Statistical Association*, 77, 773-783.

Kaye, D. H. (ed.) and Aickin, Mikel (ed.) (1986). *Statistical methods in discrimination*

litigation, Marcel Dekker Inc (New York).

Moore, D., McCabe, G. P. and Craig, B. A. (2009) *Introduction to the Practice of Statistics*, W. H. Freeman and Company, New York.

Pardoe, Iain (2008), "Modeling Home Prices Using Realtor Data," *Journal of Statistics Education*, Vol 16, No. 2.

Ricci v. DeStefano, 530 F. 3d F. 3d 88 (2nd Cir. 2008).

Ricci v. DeStefano, 554 F. Supp.2d 142 (D. Conn. 2006).

Ricci v. DeStefano, 129 S. Ct. 2659 (Super Court, 2009).

Uniform Guideline, 29 C.F.R. 1607 (D) (2000). These guidelines 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.

Utts, J. M. and Heckard, R. F. (2002), "*Mind on Statistics*", Duxbury/Thomson Learning.

Weiwen Miao
Department of Mathematics
Haverford College
Haverford, PA 19041 E-mail: wmiao@haverford.edu
