Statistical Testing of Peremptory Challenge Data for Possible Discrimination: Application to *Foster v. Chatman*

*Joseph L. Gastwirth**

Ensuring that minority groups receive fair treatment in the legal system is currently an important concern. The Castaneda v. Partida *and* Duren v. Missouri *decisions enable courts to monitor the demographic composition of the pools of potential jurors to ensure that they represent the age-eligible population of the jurisdiction. A variety of statistical measures and techniques have been used to examine data on a large sample of individuals called for jury service to check that minorities form an appropriate proportion of the jury pool. After a venire, chosen from the individuals selected from a larger jury pool to serve as potential jurors for the day, is sent to a courtroom, some members are removed for cause or by the peremptory challenges made by the parties. Although summary statistics concerning the proportion of protected group venire members challenged by the prosecution or defense are considered by courts when they evaluate a* Batson *challenge, the data rarely have been analyzed with a formal statistical hypothesis test. This paper shows that when* Batson *issues are raised, Fisher's exact test, one such statistical hypothesis test, is appropriate for examining the data on peremptory challenges. In addition to being a well-established statistical method, it evaluates the challenges made by each party assuming the other side is fair. Thus, it is consistent with the Supreme Court's statement in* Miller-El *that the defendant's actions during the jury selection process are not relevant for deciding whether the prosecution's challenges were fair. Although information is available regarding the entire population of potential jurors and the number of peremptory challenges, which are regarded as a sample from the venire, both the population and the sample are small in size. This limits the power of the test to detect a system in which the odds a minority member is challenged are two or three times those of a majority*

* Professor of Statistics and Economics, George Washington University. It is a pleasure to thank Ms. Lihui Cai, Prof. Edward Cheng, and the Editors of *Vanderbilt Law Review En Banc* for carefully reviewing a draft of the manuscript and making several helpful suggestions.

member.[1] *When data are available for similar or related trials, one can combine the results of Fisher's exact test for each trial. In every case where adequate data were reported and the Supreme Court found discrimination in peremptory challenges, Fisher's exact test found a statistically significant difference in the proportions of minority members of the venire and majority members of the venire that were removed. It also found that the prosecutor in* Foster v. Chatman *challenged a statistically significantly greater number of African-Americans than non-African-Americans. In a case where the Court did not find bias in peremptory challenges, applying the test to the data does not yield a statistically significant result. The data configuration in this case case was such that the power of the test to detect a substantial disparity was very low, and the Court properly did not give the statistics much weight.*

1. The power of a statistical test is the probability of finding a statistically significant result when the hypothesis being tested (African-American and majority members of the venire have the same chance of being challenged in our context). To calculate its numerical value one needs to consider an alternative hypothesis, e.g., the odds the prosecutor will strike a African-American member of the venire are three times those of a white. For further discussion and references, see Joseph L. Gastwirth & Wenjing Xu, *Statistical Tools for Evaluating the Adequacy of the Size of a Sample on Which Statistical Evidence Is Based*, 13 L. PROBABILITY & RISK 277 (2014).

INTRODUCTION

The Sixth Amendment of the U.S. Constitution guarantees certain parties accused of a crime the right to a public trial by an impartial jury chosen from the state and judicial district where the crime was committed.[2] The Court has interpreted the Sixth Amendment as requiring jury pools to be "fair cross section[s]" of the community, and federal law now incorporates this standard. In *Taylor v. Louisiana* and *Duren v. Missouri*, the Court set out the criteria that a defendant who is challenging the fairness of the jury must demonstrate to establish a prima facie case. Gastwirth and Pan describe the appropriate statistical methods defendants can use to support their claims and provide references to relevant literature.[3] The *Duren* opinion states "In order to establish a prima facie violation of the fair-cross-section requirement, the defendant must show (1) that the group alleged to be excluded is a 'distinctive' group in the community; (2) that the representation of this group in venires from which juries are selected is not fair and reasonable in relation to the number of such persons in the community; and (3) that this

2. While the jury should be impartial and the venires should represent the community, the fair cross section requirement does not apply to petit juries because individuals can be removed from the venire for cause or by peremptory challenges, which may diminish the representativeness of the final jury. *See* Holland v. Illinois 453 U.S. 474, 489 (1990); Lockhart v. McCree, 476 U.S. 162 (1986).

3. Duren v. Missouri, 439 U.S. 357, 364 (1979); Taylor v. Louisiana, 419 U.S. 522, 538 (1975); Joseph L. Gastwirth & Qing Pan, *Statistical Measures and Methods for Assessing the Representativeness of Juries: A Reanalysis of the Data in* Berghuis v. Smith, 10 L. PROBABILITY & RISK 17 (2011).

underrepresentation is due to systematic exclusion of the group in the jury-selection process." The same criteria and statistical methods are also appropriate for examining the representativeness of federal jury pools in civil cases, as the Seventh Amendment guarantees parties' right to a jury trial in federal actions that involve at least twenty dollars.[4]

The Equal Protection Clause of the Fourteenth Amendment also prohibits purposeful or intentional discrimination by the state, so defendants may also challenge their convictions under this Amendment if minorities were systematically underrepresented in jury pools in the jurisdiction.[5] Because defendants claiming the state violated their rights guaranteed by the Equal Protection Clause need to show that the minority underrepresentation was intended by the state, they need to present more convincing evidence than those who bring claims under the Sixth Amendment. However, the same statistical methods are appropriate for examining the data to determine whether there is a pattern of underrepresentation in both types of cases.

Even when the minority proportion of venire members is consistent with their share of the age-eligible population, some prosecutors use their peremptory challenges to eliminate or substantially reduce the representation of minorities on actual juries. During the 1986 and 1994 period, the Supreme Court expanded the scope of the constitutional prohibition against racial discrimination in the creation of juries by disallowing the use of peremptory challenges by prosecutors or defendants in criminal trials or the parties in civil trials to exclude individuals from serving on juries on account of their race or gender.

The first case to address this issue, *Batson v. Kentucky*, restricted prosecutors from using peremptory strikes to remove minority members of the venire solely based on their race.[6] A few years later, the Court also barred race-based peremptory strikes by defendants in *Georgia v. McCollum*.[7] Between those two decisions, the Court held that the Equal Protection Clause protected the right of potential jurors not to be excluded because of their race, even if they are not of the same race as the

4. *See* Hardware Dealers' Mut. Fire Ins. Co. v. Glidden Co., 284 U.S. 151, 158 (1931).

5. *See* Hernandez v. Texas, 347 U.S. 475, 482 (1954) (recognizing "the right to be indicted and tried by juries from which all members of his class are not systematically excluded").

6. 476 U.S. 79, 89 (1986) ("[T]he Equal Protection Clause forbids the prosecutor to challenge potential jurors solely on account of their race . . . .").

7. 505 U.S. 42, 59 (1992).

defendant.[8] In 1994, the Court prohibited gender-based peremptory challenges[9] and extended *Batson* to ensure fairness of peremptory challenges in civil litigation.[10] Under the *Batson* doctrine, courts assessing a party's use of peremptory strikes should consider all relevant circumstances, including the questions and statements made by the prosecutor during voir dire and the pattern of strikes.[11]

Since *Batson* and its progeny, lower courts have used a wide variety of comparisons to determine whether a pattern of strikes supports a prima facie case of purposeful discrimination. If the defendant's evidence establishes a prima facie case, the burden shifts to the prosecution to present a neutral reason for the questioned strikes. There is wide variation between circuits, and courts have not established a statistically sound method for assessing peremptory challenge data.[12] Part I of this Article discusses the most commonly used methods courts have employed to examine statistical data on peremptory challenges; some of their strengths and weaknesses are also noted. Because of the variety of approaches and the lack of common criteria in their use, similar data may strongly support a *Batson* challenge in one court but not in another.

Fisher's exact test is an appropriate statistical method for analyzing the pattern of strikes in the context of a *Batson* challenge and is described in the context of the most common setting: a defendant in a criminal case challenging the fairness of the peremptory challenges made by the prosecutor. Part II presents the statistical model underlying Fisher's exact test, and Part III applies the test to the data from *Miller-El* and *Batson*. In Part IV, Fisher's exact test is used to uncover unfairness even in situations where the courts did not realize the statistical strength of the evidence. Part V will describe some cases where a formal analysis of the data does not support a prima facie case. The data from *Foster v. Chatman*[13] are analyzed in Part VI, and the

8. Powers v. Ohio. 499 U.S. 400, 409 (1991).

9. J.E.B. v. Alabama, 511 U.S. 127, 129 (1994).

10. Edmonson v. Leesville Concrete Co., 500 U.S. 614, 616 (1991).

11. *Batson*, 476 U.S. at 96–97.

12. *See* Mickal C. Watts & Emily C. Jeffcott, *A Primer on* Batson*, Including Discussion of* Johnson v. California*,* Miller-El*,* Rice v. Collins *&* Snyder v. Louisiana, 42 ST. MARY'S L.J. 337, 357–409 (2011) (discussing the *Batson* holdings of each of the federal circuits between 2005 and 2010).

13. The Supreme Court granted certiorari in this case on May 26, 2015. 135 S. Ct. 2349. The prior decisions are *Foster v. State*, 525 S.E.2d 78 (Ga. 2000), and *Foster v. State*, 374 S.E.2d 188 (Ga. 1988).

analysis shows that the number of African-Americans removed by the prosecutor is statistically significantly higher than expected under random selection. Indeed, disparities of the magnitude seen in *Foster* have a probability of occurring, under random selection, about 1 time in 1,000. Because the usual criteria for statistical significance is the occurrence of data with a probability of random occurrence less than 0.05 (one in twenty), the result provides strong support for the defendant's claim.

When data on the peremptory challenges from a few venires in related cases are available, an approach that combines the results of Fisher's exact test can be applied to those data. This approach and its challenges are described in Part VII. A summary of the conclusions and implications of formally analyzing data on peremptory challenges is given in the final Part. Finally, in order to provide a more complete picture of the usefulness of the procedure, the Appendix analyzes the remaining Supreme Court cases concerned with the fairness of peremptory challenges.

I. CURRENTLY USED APPROACHES FOR EVALUATING PEREMPTORY CHALLENGE DATA

Typically, judges make an intuitive assessment of the statistical data on the exercise of peremptory challenges and rarely request that the parties conduct a formal analysis of the data. Melilli lists and discusses eight approaches courts have used when evaluating data on the peremptory challenges made by a prosecutor or the defendant.[14] The methods and Melilli's assessment of them are:

A) Method A simply examines the final jury to see if it contains a sufficient number of minority members. Notice that Method A ignores the pattern of peremptory challenges made by the prosecution[15]

B) Method B compares the minority percentage of the actual jury to the minority percentage of the venire. This method also does not examine the peremptory challenges made by the prosecution.[16]

C) Method C is similar to Method B except that the minority percentage of the jury is compared to the minority

14. Kenneth J. Melilli, Batson *in Practice: What We Have Learned About* Batson *and Peremptory Challenges*, 71 NOTRE DAME L. REV. 447, 471–79 (1996).

15. *Id.* at 471, 474–75,

16. *Id.* Importantly, Methods A and B also fail to consider the demographic composition of the venire and the number of peremptory challenges allotted to the prosecution.

percentage of the population in the jurisdiction. Again, the pattern of strikes is not examined.[17]

D) Method D simply counts the number of peremptory challenges used by the prosecution against members of the protected group. The minimum number needed to support a prima facie case is not specified. Thus, different courts use a different threshold number. Furthermore, the composition of the venire is not considered, which can result in inconsistent determinations. For example, when a prosecutor uses 5 of his or her allowed 6 peremptory challenges to remove 5 out of 6 minorities from a venire of 30, this should be strong evidence of unfairness. On the other hand, if there were 20 minorities on a venire of 30, the occurrence of preemptory challenges to strike 5 of them in a sample of 6 would not be surprising.[18]

E) Method E focuses on the minority percentage of the individuals the prosecutor peremptorily challenged. The problem with this measure is that it is not utilizing the appropriate benchmark, i.e., the minority proportion of the venire. Evidence of unfairness or possible discrimination arises only when the minority proportion of those challenged substantially exceeds the minority proportion of those who could be struck.[19]

F) Method F checks whether all members of the protected class on the venire (or the part of the venire that had a reasonable chance of being selected for the actual jury) have been struck. This method misconstrues *Batson*'s main concern as the actual makeup of the jury and implicitly assumes that defendants are entitled to have someone of their own race or gender group on their jury. In addition, the method does not take into account the number of minorities on the venire nor the total number of challenges allowed each side.[20]

G) Method G focuses on the percentage of protected group members of the venire removed by the prosecutor's peremptory challenges. It differs from Method F, because it does not require that all minority members of the venire are struck. The problem

17. *Id.* at 471, 476. Another problem with this method is that examines the result of the entire process, the selection of potential jurors from the eligible population and the subsequent effects of both challenges for cause and peremptory strikes. If minorities are under-represented in the venire as a consequence of inadequacies in the jury selection procedures of the jurisdiction, a fair prosecutor might be faced with a *Batson* complaint because the venire included too few minorities.

18. *Id.* at 472, 476.

19. *Id.* at 471, 476–77.

20. *Id.* at 471, 477.

with this method is it does not compare this percentage to the appropriate benchmark, i.e., the minority percentage of the venire. Melilli observes that the same percentage of minority venire members can occur in a situation clearly supporting a prima facie case under *Batson* as in a situation where it is doubtful that a prima facie case could be established.[21]

H) This method compares the percentage of the prosecutor's challenges used to challenge protected group members of the venire with the minority percentage of the venire. Melilli recommends this method as it focuses on the venire rather than the final jury. It recognizes that, on average in a minority-neutral system of peremptory challenges, the minority proportion of those challenged should be near the minority proportion of the venire. This is the basis of a statistically sound approach; however, as stated, it does not consider the number of peremptory challenges, nor does it specify by how much the minority percentage of the prosecutor's challenges needs to exceed the minority percentage of the venire in order to provide strong support for a prima facie case under *Batson*.[22] Fisher's test resolves these issues.

Several cases cited by Melilli in which courts did not find that the defendant established a prima facie case had a substantial African-American representation on the venire. Thus, the final jury had a reasonable number or proportion of African-Americans. Common sense suggests that this occurred because the defendants used their challenges to remove non-African-Americans.[23]

---

21. *Id.* at 471, 477–78. Melilli gives the following examples: In the first, there are 20 minority members on a venire of 60. The prosecutor has ten challenges and uses all of them to remove 10 or 50% of the minority members of the venire. It is difficult to imagine a court not deciding that such strong evidence supports a prima facie case. In contrast, suppose 1 of only 2 minorities on the venire of 60 is included among the 10 struck. This evidence is much less convincing than the data in the first case.

22. *Id.* at 471, 478.

23. Consider *Scott v. State*, in which there were 13 African-Americans and 15 non-African Americans on the venire and the jury consisted of 6 members of each race, 599 So. 2d 1222, 1226–27 (Ala. Crim. App. 1992). Even though the prosecutor used 7 of 8 peremptory challenges to remove African-Americans, the court did not find a *prima facie* case of purposeful discrimination. *Id.* at 1227. Since there were only 6 African-Americans remaining after the state removed 7 of them, the 6 must have served on the jury. This implies that the defendant used all 8 of their challenges to remove non-African Americans. Later, it will be seen that it is likely that both sides violated the principles established in *Batson*; however, *Scott* was decided before the Supreme Court's decisions in *Georgia* and *McCollum* barring defendants from excluding jurors based on race. This case illustrates the flaws in methods A and B described, *supra.*

Courts have used two other, related approaches. In *Miller-El v. Dretke*, the Supreme Court compared the proportion of African-Americans on the venire removed by the prosecution with the corresponding proportion of non-African-Americans on the venire.[24] This comparison is quite similar to method H. While the difference between the two proportions in *Miller-El*—79%—is so large that it is evident that a discriminatory system of challenges likely existed, the Court did not specify the magnitude of a difference between the proportions that is needed to establish a prima facie case.[25] Because these percentages depend on the number of peremptory challenges allowed in the type of case and the jurisdiction, in fairness to the Court, we should note that it might *not* be possible to specify one value for this difference that would be applicable in all situations.

In *U.S. v. Battle*, the Eighth Circuit compared the proportion (5/6, or 83%) of the prosecutor's challenges used to remove African-American members on the venire to the proportion (5/7, or 71%) of the minorities on the venire challenged by the prosecutor. The court concluded that the defendant established a prima facie case under *Batson*.[26] This approach can be considered a blend of methods E and G. Although it may seem convincing, there are two potential issues: First, the minority proportion of the venire is not considered. Second, no criterion for how large the difference between these proportions must be to support a prima facie case is specified.

II. UNDERSTANDING FISHER'S EXACT TEST AND APPLYING IT TO *BATSON* CLAIMS

The approaches discussed in Part I contrast with the formal statistical comparison of the minority fraction in a large sample of jury venires to their fraction of the age-eligible population in the jurisdiction, which was approved by the Supreme Court in *Castaneda v. Partida*.[27] As an alternative, DiPrima suggests using a formal statistical test to analyze data

24. 545 U.S. 231, 266 (finding purposeful discrimination when 91% of African-American members of the venire were removed in comparison with only 12% of the non-African-American venire members).

25. *See id*. at 265-66.

26. 836 F.2d 1084, 1085–86 (8th Cir. 1987) (finding a prima facie case of purposeful discrimination under *Batson* when prosecutor used five of its six peremptory challenges (83%) to strike five of seven African-Americans from the venire (71%)).

27. 430 U.S. 482, 495–96 & n.17 (1977).

on peremptory challenges and provides examples demonstrating that this yields more sensible results in several cases than the "intuitive" assessments of the data made by trial courts.[28] Both DiPrima and I recommend using Fisher's exact test in this context.[29]

It is useful to describe and illustrate the application of Fisher's exact test in the context of the data from *Johnson v. California*. This case considered the fairness of the prosecutor removing all three African-Americans from the 43 people who were eligible for jury service after excusals for cause. Since the prosecutor made 12 peremptory challenges, one might ask how likely it is that all 3 African-Americans would be included in a selection of 12 from this pool. If the race did not affect the prosecutor's decisions, the 12 challenged individuals should be similar to the racial composition of a random sample of the 43, in which the probability or chance of each member of the venire being challenged is the same. Intuitively, the chance that any selection would be African-American is 3/43=0.07, so we expect 12x0.07=0.84 or about 1 of the 12 selections to be African-American. Instead, three African-Americans were among the 12 selections, so we ask how frequently a random sample of 12 from 40 non-African-Americans and 3 African-Americans would include all 3 African-Americans. This probability is *less* than 0.02 or 1 in 50, which is smaller than 0.05, the most frequently used probability level for determining statistical significance, which is equivalent to the two-standard deviation criterion the Court noted in *Castaneda*.

Before presenting the probabilities of the possible outcomes of a random selection from a jury pool, which are used to determine the statistical significance of Fisher's exact test, it is important to emphasize that the data available in peremptory challenge cases consists of the *entire population* of individuals who could be removed. The individuals struck by the prosecutor form the sample from the jury-eligible venire. Furthermore, our inference will *only* apply to the case at hand. Thus, we are not relying on the data from the specific case to draw a general

28. Stephen P. DiPrima, Note, *Selecting a Jury in Federal Criminal Trials after* Batson *and* McCollum, 95 COLUM. L. REV. 888, 922–28 (1995). (recommending the use of Fisher's exact test).

29. *See* Joseph Gastwirth, *Statistical Tests for the Analysis of Data on Peremptory Challenges: Clarifying the Standard of Proof Needed to Establish a Prima Facie Case of Discrimination in* Johnson v. California, 4 L. PROBABILITY & RISK 179, 181–83 (2005). Fisher's test compares the actual number of minorities struck by the prosecutor to the number expected if those challenged were a random sample of the venire.

inference about the fairness of the peremptory challenges made by the prosecution in a large number of cases, for example over a substantial time period. This situation differs from the usual one, where one takes a random sample from a large population in order to draw inferences about that population. For example, the media sponsors pre-election polls in order to predict the outcome of an election. Similarly, the monthly unemployment rate in the nation that is reported by the Bureau of Labor Statistics is obtained from a survey of a large sample of households. Not unsurprisingly, the sample sizes needed to make reliable inferences applicable to a large population—for example, voters in a presidential election or the fraction of residents of the nation who are available for work but do not have a job—are quite large. Large samples are used in the legal setting to assess whether the system used by a jurisdiction to obtain the venires from which juries are chosen leads to a noticeably smaller minority proportion of venires than the minority fraction of the jury-eligible population.[30]In Contrast, when one has data for the *entire* population, as in the context of *Batson* analyses, one knows the proportions of individuals of the various races or gender groups on the venire and among those, the proportion the prosecutor removed. Fisher's test uses a probability model to assist our understanding of the difference between those proportions. When the probability that a random selection from the population would yield a disparity at least as large as the actual one is not small—say 0.15 or higher—one infers that the outcome could have resulted by chance.[31] On the other hand, if the probability under random selection of an outcome at least as large as the one that occurred is small—for example, less than 0.05—one doubts that the data were the result of a random process or chance.[32]

30. Duren v. Missouri, 439 U.S. 357, 364–66 (1979) (analyzing data on over 10,000 individuals who were called for jury service over an eight-month period).

31. This means that the probability a random selection or chance process would yield an outcome as far from expected as the observed data is not small (at least .15) and is not the probability that the data occurred by chance.

32. It is important to distinguish the probability that chance or random selection would produce an outcome or disparity at least as large as the one in the data from the probability that the data arose by chance. The first calculation is from a well-defined probability model; in our case, every individual remaining on the venire after challenges for cause has the same probability of being removed by the prosecution (or defendant). To calculate the second probability one would need to know the probability that the prosecutor was using a biased or unbiased (chance) selection process to make their challenges. Of course, we do not know the probability the prosecution will use a biased system, so we cannot calculate the probability that the challenges were due to chance. Quite a few legal opinions describe the first probability, known as the p-value of the test, see *infra* note 35 and surrounding text, as though it was the second. For example, the opinion in *Tabor v.*

More generally, consider a pool with n_1, a specific quantity, items from group *A* (minority) and n_2 items from group *B* (non-minority). Suppose *m* items are randomly chosen from the total of n_1+n_2 items. The probability that this random sample will contain exactly *k* of type *A* is given by:

$$P[k] = \frac{\binom{n_1}{k}\binom{n_2}{m-k}}{\binom{n_1+n_2}{m}}$$

(Eq. 2.1)

The expected number of minorities in a sample of *m* is *m* times the minority proportion of the venire—that is, $n_1/(n_1+n_2)$. Fisher's test rejects the hypothesis that the number of minorities challenged (*k*) is consistent with random selection when *k* is sufficiently larger than the expected number, such that the probability that a random sample would contain *k* or more minorities is sufficiently small. From the *Johnson* data, m=12, n_1=3 and n_2=40, so the expected number of minorities in a sample of 12 from the venire is 12×(3/43)=0.837, or just under one.

There are only four possible values—0, 1, 2, and 3—for the number of *A*s in a sample of 12 from the pool in *Johnson,* and, assuming random sampling from the pool, Equation 2.1 gives their probabilities as: P[0]=0.3642, P[1]=0.4522, P[2]=0.1658, and P[3]=0.0178.[33] Notice that the probability that a random sample of 12 from the pool would contain at least 2 minorities is 0.1836, or nearly 1 in 5, which is not low.[34]

Hilti, Inc., 703 F.3d 1206, 1223 (10th Cir. 2013), stated: "Statistical significance measures the likelihood that the disparity between groups is random, i.e., solely the result of chance." The calculation of the statistical significance of a disparity assumes that the selections are random from the venire without regard for protected group status. When the probability of observing as large a disparity as actually occurred is small, the data are not consistent with the assumption that the actual challenges were made in a neutral or random (with respect to protected group status) process. Under such circumstances, a selection process disadvantaging the protected group seems more plausible unless the prosecutor can explain or justify the challenges.

33. For example, to calculate P[3], we set k=3, n_1=3 and n_2=40; details of the calculation are given in Gastwirth, *supra* note 29, at 182.

34. The reason we consider the probability that 2 or more, i.e., P[2]+P[3] were in the sample is that if one considered 2 to be sufficiently small to conclude that the number of minorities was statistically significantly larger than its expected value, we would also conclude that any observed number larger than 2 was also statistically significant. In *Castaneda v. Partida*, the Court noted that statisticians consider disparities between the observed and expected data in the range of two to three standard deviations as statistically significant. 430 U.S. 482, 496 n.17 (1977). The two-standard deviation criterion corresponds to the commonly used .05 level of significance, while the three standard deviation one corresponds to a probability level less than 0.01. The court based its calculations on the normal approximation to the binomial distribution. *Id.* Modern computers and

In general, Fisher's exact test rejects the assumption or hypothesis of race-neutral selections when the p-value of the test, or probability of observing a disparity in the number of minorities at least as large as the prosecutor removed, is less than a pre-specified level (typically, 0.05).[35] For data sets of similar size and minority composition, the smaller the p-value, the more confidence one has in concluding that the assumption of race-neutral selection is not true.

To see that Fisher's exact test is appropriate when both parties are acting fairly, suppose the defendant also has m challenges. If both the defendant and prosecution are using a race-neutral system for deciding whom to challenge, we can consider the m individuals randomly chosen by the prosecutor (or defendant) to be the result of two stages. First, $2m$ individuals are selected from the total pool of n_1+n_2. Then one randomly divides the $2m$ into two groups of m, one group for each party. Thus, Equation 2.1 gives the probabilities of the number of minorities in each group of m. If the parties are allowed different numbers of challenges, say m_1 and m_2, the same logic applies. Then one could obtain the prosecutor's sample of m_1 from the total pool of n_1+n_2 by first taking a sample of m_1+m_2 and then taking a further random sample of m_1 from the first sample.

The assumption that every member of the venire should have the same probability of being peremptorily challenged might be considered simplistic, as the degree of undesirability of the venire members to the prosecutor is probably not the same. However, the same probability model underlying Fisher's exact test is still appropriate as long as the distributions of undesirable

---

the availability of statistical software make it easy to obtain the exact probabilities of the possible outcomes. The two-standard deviation or 0.05 probability level determine an outcome is statistically significant was developed from the analysis of data from scientific experiments. Subsequently, many social sciences have adopted it. The p-value, however, depends on several factors, including the sample size and magnitude of the disparity between the outcome and its expected value.

35. This describes the p-value of a one-sided test, where we are only concerned with observing an excess of minorities challenged by the prosecutor. When the prosecutor uses k peremptory challenges to remove b African-Americans from a venire of n, this p-value is the probability a random sample of k from the venire would include b or more African-Americans. If one thought that a prosecutor might also challenge more Whites when the defendant is African-American, then a two-sided test is appropriate. This test rejects when the observed number of African-American challenges is at least as far from its expected number, in either direction, is small. For the data in *Johnson*, it is impossible for the number of African-Americans to be as far below their expected value as the observed value (3) exceeds their expected number (just below 1) in a random sample, as the difference between zero and the expected number (0.837) is less than 1.0. Thus, the p-value of a one-sided test will equal that of a two-sided test applied to the data from *Johnson*.

characteristics in the two groups (protected and majority) are the same.[36] With the *Johnson* data, for example, one asks what is the probability that all 3 African-Americans would be among the top 12 in undesirability to the prosecutor when the prosecutor has rank-ordered the venire of 40 non-African-Americans and 3 African-Americans? More generally, one asks: what is the probability that exactly k minorities are among the highest-ranking m in a pool of n_1 minority and n_2 majority members? Equation 2.1 gives the answer. Because members of the venire are excused for cause before the peremptory challenges occur, it is reasonable to assume that the remaining individuals of different races or genders have similar distributions of characteristics related to their suitability as jurors. Thus, Fisher's exact test is valid in this more realistic context.

When Fisher's test finds a statistically significant excess of minorities among those challenged by the prosecutor and other evidence supports this conclusion, the trial judge should decide that the defendant established a prima facie case, and allow the prosecution to argue that the minority venire members were challenged for appropriate reasons. Often, courts make a side-by-side comparison of the characteristics of the challenged minority members and majority members who were not challenged to check whether the prosecutor's strikes were based on reasonable criteria applied in a consistent manner to all members of the venire.

Courts have accepted Fisher's exact test for the analysis of data in equal opportunity employment cases concerning promotion and termination, so there should not be any serious questions about its suitability for examining peremptory challenge data.[37]

Another advantage of Fisher's test is that it evaluates the fairness of one party's peremptory challenges under the assumption that the other side is fair—that is, the p-value obtained from Fisher's exact test is unaffected by the challenges of the opposing party. In light of the Supreme Court's statement

36. For an intuitive discussion, see 1 JOSEPH L. GASTWIRTH, STATISTICAL REASONING IN LAW AND PUBLIC POLICY 228 (1988). For a more formal treatment, see Chapter One of ERICH L. LEHMANN, NONPARAMETRICS: STATISTICAL METHODS BASED ON RANKS (1975).

37. For a case where Fisher's exact test was used to examine termination data, see Johnson v. Perini, No. 76-2259 (D.D.C. June 1, 1978). The data from that case are analyzed in Gastwirth, *supra* note 36, at 218–19. For cases where Fisher's exact test was used to examine promotion data, see Chin v. Port Auth. of N.Y. & N.J., 685 F.3d 135, 143–44 (2d Cir. 2012); Jurgens v. Thomas, No. CA-3-76-1183-G, 1982 WL 409, at *13–17 (N.D. Tex. Sep. 9, 1982).

in *Miller-El* that the defendant's conduct is "flatly irrelevant" to the question of whether the prosecutor's conduct revealed a desire to exclude African-Americans, this property of Fisher's test makes it even more appropriate for analyzing peremptory challenge data.[38]

III. APPLICATION OF FISHER'S EXACT TEST TO FOUR IMPORTANT PEREMPTORY CHALLENGE CASES

In order to show that Fisher's exact test is appropriate for the analysis of peremptory challenge data, it is useful to examine the data from three seminal cases considered by the Court as well as an interesting case in which both sides used their peremptory challenges to remove members of different racial groups. In the three cases in which the courts found a *Batson* violation, Fisher's test yields a statistically significant result. The results were not significant in the case in which the Court did not find a violation.

A. Miller-El v. Cockrell *and* Miller-El v. Dretke

The statistical data in the case of *Miller-El* was reviewed by the Court on two occasions.[39] *Miller-El I* concerned the conditions a state prisoner needs to satisfy in order to appeal a lower court's denial or dismissal of his petition for writ of habeas corpus,[40] while *Miller-El II* dealt with the propriety of the habeas petitioner's *Batson* claim.[41] After the original panel was screened and individuals excused for cause or by agreement of the parties, 42 individuals remained on the venire; 11 of the remaining individuals were African-American. The prosecution used 10 of its 14 challenges to remove African-Americans. Notice that 10 is substantially larger than the expected number—$(11/42)\times14=3.67$ or about 3 or 4 African-American members in a random sample of 14 from the 42 potential jurors. Table 1 presents the data.

38. Miller-El v. Dretke (*Miller-El II*),545 U.S. 231, 255 n.14 (2005).

39. Miller-El v. Cockrell (*Miller-El I*), 537 U.S. 322 (2003); *Miller-El II*, 545 U.S. 231.

40. *Miller-El I*, 537 U.S. at 327.

41. *Id.*

Table 1: Prosecutor's Challenges in *Miller-El*

Group	Number Struck	Number Kept	Total
African-American	10	1	1
Non-African-American	4	27	31
Total	14	28	42

From Table 1, one sees that 10 of the 11 African-Americans, or 90.9%, were challenged, while only 4 of the 31 non-African-Americans, or 12.9%, were challenged. The ratio of the probability an African-American was removed by the prosecutor to the corresponding probability of a non-African-American is 0.909/0.129=7.05, so an African-American was 7 times more likely to be removed by the prosecutor than a non-African-American. Statisticians have also used the ratio of the odds members of the two groups have of being challenged.[42] The odds of an African-American being challenged were 0.909/0.091=9.989 or 10. The odds of a non-African-American being challenged were 0.129/0.871=0.148, and the ratio of the odds is 9.989/.147=66.8.[43] Another formula for calculating the odds ratio from data arranged in the format of Table 1 is to calculate the ratio of the products of the two diagonals, i.e., (10x27)/(1x4)=67.5.[44] Whether one considers the ratio of the challenge rates, the difference between them (78%), or the odds ratio, the disparity in the challenge rates is clear. Indeed, Fisher's exact test yields a p-value of 6.63×10^{-6} or less than 1 in 100,000.

To appreciate the strength of the statistics in *Miller-El*, it is useful to consider two calculations. The first is the probability distribution of the number of African-Americans that would occur

42. If an event has probability p of occurring, it has probability $1-p$ of not occurring. The odds of the event happening is the ratio $p/(1-p)$. Bets are described in terms of the odds of their outcome. "Even money" bets correspond to odds of 1:1, or a probability, p, of 0.50. If an event has probability 1/3 of occurring the odds are (1/3)(2/3) =1/2 or 1:2. If one made a fair bet of one dollar on the event happening, you should receive two dollars if the event occurred. The odds ratio, of course, is the ratio of the odds of the two events occurring.

43. Because each odds is a ratio, *see supra* note 42, the ratio of two odds can be quite large or small.

44. For further discussion and illustrative examples, see ALAN AGRESTI, CATEGORICAL DATA ANALYSIS 44–46 (2d ed. 2002); Gastwirth, *supra* note 36, at 207–10; MICHAEL FINKELSTEIN & BRUCE LEVIN, STATISTICS FOR LAWYERS 36–38 (2001). The minor discrepancy between the odds ratios, 66.8 and 67.5, is due to using only three decimal places in the fractions of each group challenged.

in a sample of 14 from a pool of 11 African-Americans and 31 others. These probabilities are:

$$P[0]=5.0167\text{x}10^{-3},\ P[1]=0.0429,\ P[2]=0.1468,\ P[3]=0.2643,$$
$$P[4]=0.2769,\ P[5]=0.1762,\ P[6]=0.0689,\ P[7]=0.0164,$$
$$P[8]=0.0023,\ P[9]=0.00018,\ P[10]=6.548\text{x}10^{-6},$$
$$\text{and } P[11]=8.504\text{x}10^{-8}$$

Notice that the probabilities of observing numbers of minorities near the expected value, 3.67—that is, in the range 2-5, is quite high, while the chance of observing seven or more or none are comparatively small. Because there are only twelve possible values (0–11) for the number of African-Americans, it is difficult to obtain a region with the exact probability 0.05 to use as a threshold for determining statistical significance. Notice that the probability of observing 7 or more is slightly below 0.02, while the probability of observing no more than one African-American member is about 0.043. A strict two-sided test, using the traditional 0.05 cutoff, would find a statistically significant disparity against African-Americans if 7 or more were peremptorily challenged and a statistically significant disparity against non-African-Americans if none of the African-Americans were challenged—that is, if all 14 individuals removed were not African-American. To obtain the region with probability closest to 0.05 one needs to require one less African-American to define the region of "unlikely" selections. The resulting statistical test would be using a significance level of about 0.065, rather than 0.05. One advantage of Fisher's test is that one can determine the possible outcomes that are statistically significant before one examines the data from the case, reducing potential subjectivity. While this was not necessary in *Miller-El*, because the data are so highly statistically significant, it may be useful in other cases for courts to know the actual probability of a random selection falling into the region they will classify as statistically significant, rather than simply assume they are using a 0.05 level test, especially because a 0.05 cutoff will not always be possible. Thus, a court may accept a p-value somewhat higher than 0.05 as indicating a discriminatory pattern of strikes.[45]

45. Courts may also use the p-value as a guide to how unlikely the observed data would occur if challenges were random with respect to protected group status and require more justification for the challenges when the p-value is less than 0.05, say 0.01 than when it is near 0.05.

The discussion so far has focused on the determination of whether the probability of observing at least as many African-American members as the prosecution challenged in a random sample from the venire is low enough to reach the level of statistical significance. Requiring a low probability of this event before deciding that the prosecution might be using a biased system ensures that fair prosecutors are not called on to justify their challenges very often. Statisticians also consider the power of the test of significance,[46] which in our context is the probability that a biased process of strikes will yield a statistically significant result and make the prosecutor justify his or her challenges. When the prosecutor uses a system in which the *ratio* of the odds an African-American member is removed to those of a majority person is θ, a number bigger than 1.0—e.g., 4.0—the probabilities of the possible numbers of African-American members struck depends on θ. A biased prosecutor is essentially using a system where θ is greater than one,perhaps as large as three or more. The larger the value of θ, the larger is the probability that the number of minorities struck is statistically significantly higher than the number of majorities struck.[47]

Earlier, I discussed the probabilities of the possible numbers of minorities among the 14 challenges from the available pool in *Miller-El II*, assuming a race-neutral or random process. The probability that 7 or more would appear was slightly less than 0.02. Notice that the probability of observing 6 exceeds 0.05, so a test at the 0.05 level is actually a 0.02 level test. The probabilities of observing 7 or more minorities among the 14 challenges when the prosecutor adopts a system where the odds an African-American is chosen are θ times those of a non-African-American will increase with the value of θ. When the odds ratio, θ, is in the range 2–5, these probabilities or powers (denoted by *PW*) are

$$PW[\theta=2]=0.1389,\ PW[\theta=3]=0.3066,\ PW[\theta=4]=0.4610,\ \text{and}\ PW[\theta=5]=0.5853.$$

46. For further discussion and examples, see Gastwirth, *supra* note 36, at 132–50; FINKELSTEIN & LEVIN, *supra* note 44, at 181–84; DAVID KAYE & HANS ZEISEL, PROVE IT WITH FIGURES 88 (1997).

47. This probability distribution is the non-central hypergeometric, while the distribution of probabilities given by Eq. 2.1 that specify the probabilities relevant to Fisher's exact test is the standard hypergeometric. The formula for the probabilities for the non-central hypergeometric distribution is given in AGRESTI, *supra* note 44, at 99 (2002). The power calculations given here were obtained using the Package *"Biased Urn"* in R developed by Agner Fog (2015).

A power of 0.307, when θ is 3, means that the probability of detecting a system in which the odds a prosecutor challenged an African-American were three times those of a non-African-American is 30%, that is, there is only a 30% chance of detection given the available data in *Miller-El*. Indeed, one only has about a 60% chance of detecting a prosecutor's system in which the odds an African-American is challenged are five times those of a non-African-American. These values indicate that the power of Fisher's test to detect bias in this data set is low. Obtaining a statistically significant result when the power of the test is low is quite meaningful, as statistical significance is attainable only when the disparity between the challenge rates of African-American and non-African-American members of the venire is very large.

It should be emphasized that the probabilities of the possible outcomes under random selection or under a biased system with an odds ratio of $\theta>1$ can be calculated *before* one examines the data—that is, after the pool of prospective jurors is determined but *before* the peremptory challenges are made. Thus, in *Miller-El*, one might decide that requiring the prosecutor to challenge least 7 is too strict and might set the threshold at 6. Then, according to the probabilities summarized above, the probability a prosecutor using a fair system would be questioned would increase by nearly 0.07, so that the overall probability a fair prosecutor would need to explain his or her challenges would be 0.09, or 9%. Doing so would also increase the power of the test:

$$PW[\theta=2]=0.3589,\ PW[\theta=3]=0.5874,\ PW[\theta=4]=0.7362,\text{ and } PW[\theta=5]=0.8284.$$

This trade-off between the probability of questioning a prosecutor's challenges—or, in statistical parlance, rejecting the null hypothesis of fairness—and the probability of detecting an alternative procedure disadvantaging a protected group is inherent in statistical hypothesis testing.[48] The effect of requiring that the p-value is less than 0.05, say, when small data sets are analyzed is relevant to the analysis of peremptory challenge data, because the population of eligible potential jurors is rarely greater than 50 and the number of allowed strikes is typically a relatively small sample of that population. Thus, in situations in which the

48. For further examples and references to the literature, see Gastwirth & Xu, *supra* note 1.

parties have a small number—say 15 or fewer—of peremptory challenges, it is reasonable for courts to adopt a 0.10 level of significance in place of the usual 0.05 level as the threshold to support a prima facie case.[49] Put another way, the smaller the magnitude of the p-value, the more carefully courts should scrutinize the supposedly "neutral" explanations offered by the prosecutor.

B. Batson v. Kentucky

Batson, an African-American, was charged with burglary and the receipt of stolen goods.[50] At trial, after some potential jurors were excused for cause, the parties exercised their peremptory challenges on a list of qualified jurors. The number of potential jurors was equal to the number of jurors to be seated plus the number of allowable peremptory challenges. The offense was a felony, and the court needed 12 jurors and an alternate. The prosecutor was entitled to 6 challenges and the defendant 9.[51] The prosecutor removed all 4 African-Americans on the list.[52] The list contained a total of 28 potential jurors if the alternate was included. Thus, there were 24 non-African-Americans on the list along with the 4 African-Americans. Assuming the prosecutor used all 6 of his or her allowed challenges, the data are summarized in Table 2.

Table 2: Challenges in *Batson v. Kentucky*

Group	Number Struck	Number Kept	Total
African- American	4	0	4
Non-African-American	2	22	24
Total	6	22	28

Application of Fisher's exact test yields a p-value of 0.00073 or *less than 1 in 1,000*. The odds ratio is infinity, so it may be

49. Because the size of venires is usually small and minorities form a modest fraction of the venire, the power of Fisher's test using a strict 0.05 level to determine significance—to detect a biased system—says an odds ratio of three, will usually be low. The issue of low power does not arise in the context of cases concerning representation in a jurisdiction as a large sample of venires is available for analysis.

50. Batson v. Kentucky, 476 U.S. 79, 82 (1986).

51. *Id.* at 82 n.2.

52. *Id.* at 83.

preferable to consider the lower end of a 95% confidence interval[53] where the ratio of the odds an African-American was challenged to those of a non-African-American were 3.72, so that the odds a prosecutor would remove an African-American were nearly 4 times those of a non-African-American. If the prosecutor had used fewer challenges than his allotted 6, or if the number of non-African-Americans on the list was larger than 24, the p-value would be even smaller—that is, the data would provide stronger evidence of a non-neutral system of challenges used by the prosecution.

To appreciate the strength of this statistically significant result, one needs to consider the power of the test—that is, the probability that the test would classify as unfair a system in which the ratio of the odds the prosecutor challenged a African-American relative to those of a non-African-American equaled 2, 3, 4, 5 or 10. These probabilities, respectively, are 0.0866, 0.1651, 0.2429, 0.3148 and 0.5694, which are quite *low*. For example, when 6 individuals are chosen from a group of 4 African-Americans and 24 non-African-Americans, the probability of Fisher's test classifying a prosecutor's challenges as statistically significant when the odds an African-American is removed are 5 times those of a non-African-American is slightly less than 1/3. Indeed, one has less than a 60% chance of classifying an odds ratio of 10 as statistically significant.

The powers of Fisher's test to detect values of the odds ratio greater than one in *Batson* are *less* than the corresponding ones in *Miller-El*. This is a consequence of the smaller number of challenges made by the prosecutor as well as the more unbalanced jury pool (minorities formed one-seventh of the pool in *Batson* versus nearly one-fourth in *Miller-El*).

53. A confidence interval takes into account the "sampling error" or the normal variation occurring in random samples, often referred to as the "margin of error". The sampling variability of odds ratios in small samples is quite large. For example the lower end of a 95% confidence interval for the odds ratio in *Miller-El* is 5.83, much less than the odds ratio of 67 calculated from the raw data. Of course, when one is 95% confident that the odds an African-American would be removed by the prosecutor are nearly four times those of other members of the venire, it is difficult to believe that race was not entering into the prosecutor's decisions.

C. Purkett v. Elam

In *Purkett v. Elam*, the defendant raised a *Batson* claim after the prosecutor removed 2 of 3 African-Americans from the venire.[54] The trial judge accepted the reasons offered by the prosecutor. While the state appeals court affirmed the lower court, the Eighth Circuit did not think the reasons given were legitimate.[55] The U.S. Supreme Court reversed over dissents by Justices Stephens and Breyer.[56] Although the statistical data, given in Table 3, did not have a role in the Court's decision, the state appellate court observed that the state used 1/3 of its peremptory challenges against African-Americans, who formed 12% of the venire, but did not give much weight to the statistics and accepted the prosecutor's explanation.

Table 3: Prosecutor's Challenges in *Purkett v. Elam*

Group	Number Struck	Number Kept	Total
African- American	2	1	3
Non-African-American	4	18	22
Total	6	19	25

The expected number of African-Americans in a random sample of 6 from the venire is 0.5, so one would expect only 1 or none amongst the prosecutor's challenges. Fisher's exact test yields a p-value of 0.133, a *non-significant* result. It is instructive to present the probabilities of the 4 possible numbers of minorities in a sample of 6 from the venire. They are:

$$P[0]=0.4213,\ P[1]=0.4461,\ P[2]=0.1239,\ \text{and}\ P[3]=0.0087$$

Notice that the only outcome for which the p-value of Fisher's test is below the 0.05 level is 3—that is, only if the prosecutor were to remove all African-Americans from the venire. Assuming random selection, the probability of this outcome is slightly less than 0.01. Thus, statistical testing of the data in this case is actually using a 1 in 100 criterion for statistical

54. 514 U.S. 765, 766 (1995) (per curiam). The data are given in *Missouri v. Elem*, 747 S.W. 772, 773–74 (1988).

55. *Purkett*, 514 U.S. at 767.

56. *Id.* at 769–70.

significance, much more stringent than the commonly used 0.05 level. This implies that the test will have *low power* for this particular data set. Factors that contribute to the low power of the test are the small sample size, small African-American fraction of the venire, and the small number of peremptory challenges. The powers for the same set of possible odds ratios, θ, considered in our discussion of *Miller-El*, are:

$$PW[\theta=2]=0.0370,\ PW[\theta=3]=0.0755,\ PW[\theta=4]=0.1173,$$
$$PW[\theta=5]=0.1560, \text{ and } PW[\theta=10]=0.3348.$$

Notice that these probabilities or powers are less than those in *Batson* or *Miller-El*. In a situation where analyzing the available data only has a probability of 1/3 of detecting an unfair system of challenges in which the odds a member of a protected group being challenged are ten times that of another member of the jury pool, the fact that Fisher's test does not yield a significant result is *not* very informative. In cases like this, courts need to rely primarily on the non-statistical evidence in the case and, possibly, statistical data from similar cases. In other contexts, when a statistical test has a reasonably high probability—for example, 0.80—of detecting a meaningful difference, a non-significant result is good evidence of no effect or no difference in the proportions of the two groups challenged.

D. Eagle v. Linahan

Eagle v. Linahan is interesting, because the peremptory challenges made by both parties showed a strong pattern of unfairness; the prosecution appeared to target African-Americans, while the defense appeared to target non-African Americans.[57] The African-American defendant argued that the trial judge decided his Batson claim incorrectly when it was brought up at trial. The original venire included 16 African-Americans and 26 non-African Americans.[58] The prosecution used 8 of its 9 challenges to remove African-Americans, while the defendant used 18 of his 19 to remove non-African Americans.[59] The actual jury included 4 African-Americans and 8 non-African

57. 279 F.3d 926, 930–31 (11th Cir. 2001).

58. *Id.* at 930 n.3.

59. *Id.*

Americans.[60] The prosecutor argued that there was no *Batson* violation, as African-Americans formed 1/3 (33%) of the jury and 31% of the venire.[61] After agreeing with the prosecution's calculations, the trial judge denied the Batson claim because the jury reflected the African-American proportion of the venire; he also observed, "I think both of you were doing what you could to get the different races off."[62]

Table 4a: Prosecutor's Challenges in *Eagle v. Linahan*

Group	Number Struck	Number Kept	Total
African- American	8	8	16
White	1	25	26
Total	9	33	42

Table 4b: Defendant's Challenges in *Eagle v. Linahan*

Group	Number Struck	Number Kept	Total
African- American	1	15	16
White	18	8	26
Total	19	23	42

On appeal, the Eleventh Circuit emphasized that the focus of a *Batson* inquiry concerning the fairness of the prosecution's challenges is the pattern (8 of 9 challenges removed African-American members of the venire) of its challenges rather than the fact that the defendant's challenges against non-African Americans offset it.[63] While the court was troubled by the fact that the defendant might have used his challenges in a discriminatory manner, it held that such behavior did not justify the unconstitutional use of peremptory challenges by the prosecution, in part because *Batson* also protects the right of prospective jurors to fair treatment.[64]

Application of Fisher's exact test to Table 4a yields a p-value of 0.0002, or 1 in 5,000, clearly indicating that the prosecutor removed a statistically significantly greater number of

60. *Id.* at 930–31.

61. *Id.* at 931.

62. *Id.* We note that African-Americans formed 16/42=38.1%, rather than 33%, however, in a sample of 12, the difference between 38.1% and 33.3% is not significant. Perhaps the opinion has a typographical error or a few members of the pool were removed for cause.

63. *Id.* at 942.

64. *Id.*

African-American members of the venire. Applied to the data in Table 4b, the test yields a significant p-value of $7.65x\ 10^{-5}$, or slightly less than 1 in 10,000, showing that the defendant removed a statistically significantly greater number of non-African Americans from the venire. Thus, Fisher's test shows that the peremptory challenges of both parties were apparently discriminatory and provides strong support for the appellate decision reversing the lower court for its failure to find in favor of the defendant on his *Batson* claim.

IV. FISHER'S TEST UNCOVERS POSSIBLE DISCRIMINATION COURTS OVERLOOKED IN OTHER *BATSON* CASES

This Part discusses several cases in which the courts did not detect a pattern of bias in the prosecutor's challenges, in part, because the court compared the percentage of the jury formed by African-American members to their percentage of the venire. As noted previously, this can occur when the defendant uses all or nearly all their strikes to remove non-African-American members of the jury pool.

A. U.S. v. Forbes[65]

With respect to the *Batson* claim in *U.S. v. Forbes*, the issue was whether the prosecution violated the African-American defendant's equal protection rights when it used its peremptory challenges to remove 3 of 6 African-Americans from a panel of 33[66] Although the prosecutor was allowed 6 challenges, he only used 5.[67] The resulting jury contained 2 African-Americans, and the court noted that the African-American percentage of the jury (16.67%) was virtually identical to the African-American percentage (18.18%) of the jury pool.[68]

65. 816 F. 2d.1006 (5th Cir. 1987)
66. *Id.* 1006, 1008–09.
67. *Id.*
68. *Id.* at 1009.

Table 5: Prosecutor's Challenges in *U.S. v. Forbes*

Group	Number Struck	Number Kept	Total
African- American	3	2	5
Non-African-American	2	24	26
Total	5	26	31

In addition to these facts, the district court noted that the prosecutor could have removed all 6 African-Americans on the panel. Therefore, the court concluded that the defendant had not established a prima facie case.[69] The prosecutor did offer neutral explanations for their challenge of 2 African-Americans but did not explain the third.[70] Table 5 summarizes the data.

Fisher's exact test yields a statistically significant p-value of 0.02, indicating that the prosecutor's challenges included a disproportionate fraction (60%) of African-Americans, as they formed only 18.2% of the panel. Due to the fact that there were only 5 African-Americans on the venire and the prosecutor used 5 challenges, there are only 6 possible values (0- 5) for the number of African-Americans in a sample of 5 from the venire of 31. Their probabilities are:

$$P[0]=0.3871,\ P[1]=0.4399,\ P[2]=0.1530,\ P[3]=0.0191,$$
$$P[4]=0.00077, \text{ and } P[5]=5.8885\text{x}10^{-6}.$$

As the test can only reach significance when 3 or more minorities are among the 5, the actual significance cutoff of the statistical test is 0.02, rather than 0.05. This indicates that the power of the test to detect bias in the prosecutor's challenges of African-American members on the venire is low. The precise powers for various odds ratios are:

$$PW[\theta=2]=0.0809,\ PW[\theta=3]=0.1587,\ PW[\theta=4]=0.2368,$$
$$PW[\theta=5]=0.3105, \text{ and } PW[\theta=10]=0.5766.$$

Notice that the power of the test to detect an odds ratio of 5 is just under 1/3, which implies that a pattern of strikes disadvantaging African-Americans is unlikely to reach statistical significance unless it is quite stark.

69. *Id.* at 1009 n.5.
70. *Id.* at 1009.

To see whether the prosecutor could have removed a majority member with the unused challenge and changed our inference, one now assumes that 3 of the 26 non-African-Americans would have been challenged along with 3 of 5 African-Americans. Fisher's test yields a p-value of 0.038, which remains statistically significant at the usual 0.05 level.

Although the opinion does not report the challenges made by the defendant, since 2 African-Americans served on the jury, we know that the defendant removed 6 non-African-Americans with his challenges. Thus, after the peremptory challenges, the remaining panel was composed of 2 African-Americans and 19 non-African-Americans. If one considers the jury of 12 as a random sample of these 21 individuals, the probability that both African-Americans would be on the jury is nearly 1/2 (0.4857).

B. U.S. v. Allison

In his dissent in *U.S. v. Allison*, Judge Hatchett criticized the majority opinion, because it based its decision on a comparison between the African-American percentage of the jury and the African-American percentage of the panel along with the fact that the prosecutor had an unused peremptory challenge, which could have removed another African-American from the venire.[71] Apparently, Judge Hatchett focused on the high proportion (50%) of African-Americans on the panel struck by the prosecutor, relative to the African-American proportion (15%) of the venire.[72] The data are reported in Table 6.

Table 6: Prosecutor's Challenges in *U.S. v. Allison*

Group	Number Struck	Number Kept	Total
African- American	3	3	6
Non-African-American	2	32	34
Total	5	35	40

Analyzing the Allison data with Fisher's test yields a statistically significant p-value of 0.0178. In light of the fact that in this small data set the power of Fisher's test to detect a pattern of bias in the prosecutor's strikes is small, a significant result

71. 908 F.2d 1531, 1539 (11th Cir. 1990).
72. *Id.* at 1539.

indicates that the proportion (50%) of African-Americans removed by the prosecutor substantially exceeds the corresponding proportion (5.9%) of non-African-Americans.[73] Thus, formal statistical analysis supports Judge Hatchett's dissent and indicates that the court should have carefully examined the proposed explanations and compared the attributes of the minorities struck to those of the majority members who were not struck.

C. U.S. v. Grandison[74]

This case is somewhat unusual, because the African-American proportion (27.4%) of the venire exceeded their proportion in the state's population (22%). After individuals were removed for cause, the venire consisted of 14 African-Americans and 37 others. The prosecution used 9 of its allotted 10 challenges to remove six African-Americans. Table 7 reports the data.

Table 7: Prosecutor's Challenges in *U.S. v. Grandison*

Group	Number Struck	Number Kept	Total
African- American	6	8	14
Non-African-American	3	34	37
Total	9	42	51

Before analyzing the data, it should be noted that when the defendant raised the *Batson* claim at trial, the judge denied it without asking the prosecution to explain or justify its challenges. The majority of the appeals court noted that the jury consisted of 2 African-Americans and 10 non-African-Americans, but during the proceedings, the prosecution had accepted a jury consisting of 3 African-Americans, one of whom it later struck, but the defendant did not accept that jury. The court also considered the alternates, 3 out of 6 of whom were African-American, when it compared the African-American share (5/18=27.7%) of the total jury to their share of the venire or the state's population. Thus, the majority did not believe that the statistical evidence strongly supported a prima facie case.

73. As an example of the low power of the test in this case, the power of the test is only 0.31 when the odds a prosecutor removes an African-American are five times those of a White.

74. 885 F. 2d 143 (4th Cir. 1989).

In dissent, Judge Murnaghan stated, "With the number of peremptory strikes of African-American members of the venire disproportionate when measured against almost any criteria, it becomes incumbent upon the prosecution to justify its strikes as not racially motivated." Later, he noted that the prosecutor reduced a venire that started with nearly 30% African-Americans to a panel with 19% African-Americans.

Before formally analyzing the data, notice that the challenge rate (6/9=66.7%) is more than double the African-American share (27.4%) of the venire. Applying Fisher's exact test yields a p-value of 0.0085, less than 1 in 100. Thus, the data indicate that a statistically significantly larger fraction of the African-American members of the venire were struck by the prosecution and support the dissent's argument that the trial court should have required the prosecutor to justify those challenges.[75]

D. U.S. v. Robinson[76]

U.S. v. Robinson predates *Batson*, and an application of formal statistical analysis in that case illustrates how the *Batson* decision gives the defendant greater opportunity to show the prosecutor impermissibly removed members of a protected group from the venire than under *Swain v. Alabama*.[77] The decision focused on the proportion of criminal juries including at least one African-American in the jurisdiction over a two-year period.[78] The district court's opinion noted that 39 venires included at least one African-American member and that the average number of African-Americans was slightly over two.[79] The court then calculated that if 12 jurors were chosen from a venire of 28, one would expect that approximately 68% would contain at least one

75. The removal of five minority jurors would just meet the threshold or cut-off value of a 0.05 level test. The power or probability the test would detect a biased system in which the odds an African-American is removed by the prosecutor are 5 times those of a non-African American is 0.713, however, the probability of detecting a system when the odds ratio is three is only 0.443. These powers are somewhat larger than those in *Batson* are because the venire is larger, the minority forms a larger fraction of the venire and the sample size (number of peremptory challenges) is larger. From a statistical viewpoint, these probabilities of detecting a biased system of challenges are still relatively low. For example, medical studies are designed to have 80% power.

76. 421 F. Supp. 467 (D. Conn. 1976), *vacated sub nom*. United States v. Newman, 549 F.2d 240 (2d Cir. 1977).

77. Swain v. Alabama, 380 U.S. 202 (1965) (placing the burden of proof of purposeful discrimination upon the defendant), *overruled by* Batson v. Kentucky, 476 U.S. 79 (1986).

78. Robinson, 421 F. Supp. at 469.

79. *Id.* at 472.

African-American.[80] The data showed that African-Americans served on only 1/3 of the juries in the 39 venires analyzed, and the judge concluded that the prosecutor's challenges had a substantial impact on reducing the frequency of juries with at least one African-American member.[81]

The Second Circuit issued a writ of mandamus vacating the lower court's decision that directed that 4 challenged African-American members be restored to the final panel.[82] The Second Circuit further ordered that detailed statistics on the peremptory challenges made by the Government be preserved.[83] The appellate court observed that the statistical data submitted by the defendant combined data on the composition of juries for the Hartford area with similar data from the New Haven and Waterbury district, where the trial was held.[84] The opinion noted that in the New Haven division there were 15 trials with panels that included an average of 2 minorities and that minorities were on the final jury in 9 (60%) of them.[85] This percentage is close to the expected 68% and the court deemed the difference "*de minimis*."[86]

After voir dire, the original venire consisted of 37 individuals, 4 of whom were African-American. In addition to the jury of 12, 5 alternates were chosen. The prosecution had 7 challenges and removed *all* 4 African-American members of the panel. The data are given in Table 8. Applying Fisher's test to the data yields a p-value of 0.00053. In other words, only about 1 in 2000 random samples of seven from a venire of 4 African-Americans and 33 non-African-Americans would include all 4 African-Americans. The pattern of peremptory challenges in *Robinson* is very similar to that in *Batson* and strengthens the defendant's prima facie case.[87]

80. *Id.*

81. *Id.* at 473.

82. *Newman*, 549 F.2d at 250.

83. *Id.*

84. *Id.* at 242–43.

85. *Id.* at 245.

86. *Id.*

87. The p-value of Fisher's test in Robinson somewhat less than its p-value on the *Batson* data suggesting that the data in Robinson should be slightly more convincing evidence of a pattern disadvantaging African-Americans.

Table 8: Prosecutor's Challenges in *U.S. v. Robinson*

Group	Number Struck	Number Kept	Total
African- American	4	0	4
Non-African-American	3	30	33
Total	7	30	37

It is useful to examine the probability distribution of the number of African-Americans in a sample of 7 and the power of the test to detect meaningful odds ratios of 3 or more.

The probabilities of observing 0-4 African-American jurors in a sample of 7 from the venire are:

$$P[0]=0.415,\ P[1]=0.430,\ P[2]=0.138,\ P[3]=0.016,\ \text{and}\ P[4]=0.00053.$$

Therefore, even if the prosecutor had removed 3 of the 4 African-American members from the panel the result would be statistically significant.[88]

The power of Fisher's test to detect a prosecutor using a removal system where the odds of an African-American member being removed are a multiple of the odds that a majority member of the venire is removed are:

$$PW[\theta=2]=0.069,\ PW[\theta=3]=0.138,\ PW[\theta=4]=0.210,\ PW[\theta=5]=0.278,\ \text{and}\ PW[\theta=10]=0.533.$$

As expected, the power of Fisher's test in the data is *less* than 1/2, unless the odds an African-American is removed are at least 10 times those of a non-African-American. Again, finding a highly statistically significant result when the power of the test is low indicates that the pattern of strikes made by the prosecution in Robinson clearly diminished the chance of African-American members of the venire serving on the jury.

V. CASES WHERE STATISTICAL ANALYSIS DOES NOT SUPPORT A

88. The probability of observing three or more is $P[3]+P[4]=0.0165$. In addition, the probability of observing no African-Americans is 0.415, so the probability of observing at least one is 1–0.415=0.685, as Judge Newman calculated. This also means that a two-sided test reduces to a one-sided test as the probability of observing no African-Americans among those challenged is greater than 0.05 or even 0.10; the value 0.05 is the most frequently used criterion for determining statistical significance.

PRIMA FACIE *BATSON* CLAIM

There are a few cases where formal statistical analysis using Fisher's test or a related method would have helped the prosecution raise doubts about the significance of an alleged pattern. Three are discussed in this section.

A. U.S. v. Ochoa-Vasquez

In *U.S. v. Ochoa-Vasquez*, the Hispanic defendant alleged, along with several other claims, that the prosecutor peremptorily challenged a disproportionately high percentage of Hispanics.[89] The court considered both the challenges made during the selection of the jury and then included the challenges made in selecting the 5 alternates.[90] The data for the original jury are more favorable to the defendant and are given in Table 9.

Table 9: Prosecutor's Challenges in *U.S. v. Ochoa-Vasquez*

Group	Number Struck	Number Kept	Total
Hispanic	5	39	44
Non-Hispanic	1	37	38
Total	6	76	82

Because a defendant can also raise a *Batson* claim when the prosecution uses a disproportionate percentage of its strikes to remove individuals of a different race-ethnic group,[91] a two-sided Fisher-exact test is appropriate. The p-value of the test is 0.21, which is not close to significance. Although the challenge rate is 5/6, or 83.3%, Hispanics formed 53.6% of the venire, so a sample of 6 has a reasonable probability of including 5 or more Hispanics, while including only 0 or 1 would indicate a pattern disfavoring non-Hispanics. The exact probabilities of each of the 7 possible values for the number of Hispanics in a sample of 6 from the venire are:

$$P[0]=0.0078,\ P[1]=0.0631,\ P[2]=0.1992,\ P[3]=0.3191,$$
$$P[4]=0.2725,\ P[5]=0.1179,\ \text{and}\ P[6]=0.0202.$$

89. 428 F 3d 1015, 1033 (11th Cir. 2005).
90. *Id.* at 1040–42.
91. Powers v. Ohio, 499 U.S. 400, 402 (1991).

If one was only concerned with whether an excess number of Hispanics were removed by the prosecution, the p-value of Fisher's test is 0.138 ($P[5]+P[6]$), which exceeds 0.10, the largest commonly used threshold for statistical significance. Formal statistical analysis shows that although the data might initially appear to suggest an unfair pattern, they do not provide strong support for the defendant's claim.

In his dissent, Judge Barkett emphasized that the challenge rate (83.3%) was greater than the Hispanic portion (53.6%) of the venire.[92] He cited *U.S. v. Alvarado*[93] where the court found a pattern existed when the challenge rate was 172% of the minority fraction of the venire. In *Ochoa-Vasquez*, the corresponding ratio is 155%, somewhat less than in *Alvarado*.[94] Finally, if one includes the challenges made to possible alternate jurors, the data are more favorable to the prosecutor, who only removed 3 non-Hispanic potential alternates. Thus, the challenge rate becomes 5/9, or 55.5%, very close to the Hispanic percentage of the venire.[95]

B. Golphin v. Branker

In *Golphin v. Branker*, the Fourth Circuit rejected a habeas petitioner's *Batson* claim after the prosecution struck 5 of 7 or 71% of the African-Americans on the venire and 13 of 31, or 45%, of the non-African-Americans.[96] A two-sided Fisher's exact test yields a p-value of 0.222, and a one-sided test that checks to see if an excess of African-American members of the venire were excluded yields a p-value of 0.161. Put another way, about 1 in 6 samples of 18 of the venire members would include at least 5 African-Americans. Thus, Fisher's exact test supports the court's decision to deny the petitioner's *Batson* claim.

92. *Ochoa-Vasquez*, 428 F.2d at 1055 & n.15.

93. 923 F.2d 253 (2d Cir. 1991).

94. *Ochoa-Vasquez*, 428 F.2d at 1054–55.

95. Because an individual chosen to be an alternate has a much lower probability of actually being a member of the final jury deciding the case, this simple pooling of all the data is statistically inappropriate as the challenges to alternates are given the same weight in the calculation as the challenges made in selecting the original jury. Even if the challenges to alternates were down-weighted to reflect their lower probability of serving on the final jury, their inclusion would reduce the significance of any pattern in the *Ochoa-Vasquez* case.

96. 519 F.3d 168, 179, 187–88 (4th Cir. 2008).

C. U.S. v. Montgomery

The decision in *U.S. v. Montgomery* describes the defendant's *Batson* claim as follows:

> There were a total of four African-American persons available for selection as jurors, making up 14% of the venire. The government used 2 of its 6 strikes (33%) to eliminate 2 of the 4 African-American members of the venire. The defendant then used 1 peremptory challenge to strike 1 of the 2 potential remaining African-American jurors, so that the actual jury consisted of 11 non-African Americans and 1 African-American.
>
> Although the jury accepted by the government included two African-Americans, Montgomery asserts that these percentages indicate that African-American members of the jury panel were peremptorially struck at a rate in excess of double of that which a proportionate striking of African-Americans would have resulted in. He requests that his case be remanded pursuant to Batson for the district court to determine whether he has a prima facie case of purposeful discrimination and whether the government had permissible reasons for the strikes.[97]

The data is presented in Table 10.

Table 10: Prosecutor's Challenges in *U.S. v. Montgomery*

Group	Number Struck	Number Kept	Total
African-American	2	2	4
Non-African-American	4	20	24
Total	6	22	28

Applying Fisher's exact test to the data in Table 10 yields a p-value of 0.1915,[98] which indicates that about 1 in 5 random samples of 7 from the venire would include 2 African-Americans. The exact probabilities of each of the possible numbers of African-Americans in a random sample of 6 are:

$$P[0]=0.3573,\ P[1]=0.4513,\ P[2]=0.1692,\ P[3]=0.0215,\ \text{and}\ P[4]=0.000073.$$

Because Fisher's test would only classify a prosecutor removing 3 or more African-American members of the venire as statistically significant, one is actually using a significance level of 0.022, well below 0.05. This results from the relatively small

97. 819 F.2d 847, 850 (8th Cir. 1987).

98. The same p-value results if one conducts a two-tailed or one-tailed test. When the protected group forms a relatively small fraction of the venire because it may be impossible to have a statistically significant shortfall of minority selections as the probability of observing none of them in a small sample from the venire is greater than 0.05 or even 0.10.

African-American fraction of the venire and the small number of peremptory challenges. Consequently, the power of Fisher's test to detect a statistically significant pattern in the data from *Montgomery* is relatively low. For example, even if the prosecution used a system in which the odds an African-American member of the venire would be struck were 5 times those of a non-African-American, the probability that 3 or more African-Americans would be removed (the power of the test) is only 0.31. While a non-significant result when the power of the statistical test is low does not show that the prosecutor's challenges were fair, a p-value of almost 20%, or 1 in 5, is too large for the data to support a prima facie case. In cases like *Montgomery*, the data are insufficient to draw a firm conclusion, so courts need to rely on other evidence. Indeed, the opinion noted that the prosecution did not strike 2 African-Americans even though they could have.[99]

Instead of analyzing the data, the defendant advocated for a simple comparison of the proportion of the prosecution's challenges used to remove minorities to the minority proportion of the venire. However, this approach does not take into account the variation in characteristics of the potential jurors inherent in any random sample of the eligible population.[100]

VI. APPLYING FISHER'S EXACT TEST TO *FOSTER V. CHATMAN*

This Part focuses specifically on petitioner Timothy Foster's *Batson* claims in *Foster v. Chatman*, which is currently before the Supreme Court.[101] The jury selection process in the case is described in Foster's brief on the merits.[102] Although we focus on the statistics, which are not analyzed in the Joint Appendix, the petitioner's brief includes copies of notes by the prosecutor indicating the race of potential jurors.[103] After voir

99. *Montgomery*, 819 F.2d at 851.

100. As noted previously, the same probability model is applicable when the distributions of undesirable characteristics are the same in both racial groups. Because venire members are a random selection of the population, there will be substantial variation in the desirability of the individual members of the venire to the two parties.

101. No. 14-8349 (U.S. argued Nov. 2, 2015).

102. Brief of Petitioner at 4–9, Foster v. Chatman, No. 14-8349 (U.S. July 24, 2015).

103. *Id.* at 15–18. In *Avery v. Georgia*, the Court noted that the names of White and African-American potential jurors were printed on white and yellow cards, respectively. 345 U.S. 559, 561 (1955). As Justice Frankfurter noted in his concurrence, this practice makes it easier to discriminate and in fact no African-Americans were on the panel of 60 from which the jury was selected. *Avery*, 345 U.S. at 563–64. The State's brief in *Foster* alleges that the prosecution needed to know the race of potential jurors because the defendant had questioned the demographic

dire and excusals for cause, there were 42 prospective jurors, 4 of whom were African-American. The prosecution used 9 of its 10 challenges and removed all 4 African-American members, although only 0.857, or slightly less than 1, would be expected in a random drawing from the venire.

Table 11: Prosecutor's Challenges in *Foster v. Chatman*

Group	Number Struck	Number Kept	Total
African- American	4	0	4
Non-African-Americans	5	33	38
Total	9	33	42

Source: Petitioner's Brief on the Merits at 5-6

Applying Fisher's test to the data in Table 11 yields a p-value of 0.0011, which is statistically significant at both the 0.05 and 0.01 level. The lower end of a 95% confidence interval for the odds ratio is 3.11, indicating that we can have 95% confidence that the odds an African-American would be struck by the prosecutor are *at least 3* times those of a non-African-American.

It is instructive to examine the exact probabilities, $P[j]$, j=0–4 for the number of African-Americans appearing in a random selection of 9 from the venire. These are:

$$P[0]=0.3656,\ P[1]=0.4387,\ P[2]=0.1698,\ P[3]=0.0248,\ \text{and}\ P[4]=0.0011.$$

Notice that the number (4) of African-Americans on the venire is quite small and the p-value of a two-sided test at the usual 0.05 level of significance is the same as the p-value of a one-sided test.[104] This implies that the threshold of a two-sided or one-sided, 0.05-level test of significance is 3 and has a true probability level of $P[3]+P[4]=0.036$, which is less than 0.05. Thus, even if the prosecution had removed 3 African-Americans instead of 4, the data would still indicate that a statistically significantly greater

composition of the jury panel. Brief of Respondent at 6, Foster v. Chatman, No. 14-8349 (U.S. Sep. 8, 2015).

104. The probability of the smallest possible number (0) of minorities in a sample of four is 0.366, so observing no African-Americans among the four is consistent with a random selection from the venire.

number of African-Americans were challenged than non-African-Americans.

The power of Fisher's test—that is, the probability it would find a statistically significant difference—is the probability that at least three African-American members would appear in a random sample of 9 from the venire when the odds the prosecution challenges them are greater than those of a non-African-American. The power (*PW*) of Fisher's exact test to detect odds ratios (θ) in the range 2 through 5 and 10 are:

$$PW[\theta=2]=0.1011,\ PW[\theta=3]=0.1914,\ PW[\theta=4]=0.2791,$$
$$PW[\theta=5]=0.3583,\text{ and } PW[\theta=10]=0.6237.$$

Notice that these powers are quite *low*—less than those in *Miller-El*, although slightly higher than those in *Batson*. For example, the probability of detecting a prosecutor using a system in which the odds an African-American is removed are *3* times those of a non-African-American is about 0.20—that is, 80% of the time, the test will *not* classify the challenges of such a prosecutor as statistically significant. Even if the odds a prosecutor removes minorities are 10 times those of a non-African-American, there is nearly a 40% chance the test will not find the challenges statistically significantly different. Thus, the finding of a statistically significant disparity with a p-value of 0.0011 indicates that the difference in the challenge rates is substantial, so the explanations offered by the prosecution for removing the African-American members deserve careful scrutiny.

Statistical tests have *low* power in small data sets, regardless of whether the data refers to a small random sample from a large population or a small sample of a modest fraction of a small population. This is a consequence of keeping the significance level, or probability of making a fair prosecutor explain their challenges, low—for example, at 0.05 or 0.10.[105] This problem is more acute in situations where minority groups

105. In statistical textbooks, the hypothesis of fairness is the null or no-effect hypothesis. In scientific applications there often is evidence supporting the null hypothesis, so one does not want to question it unless the observed data have a low probability of occurring when it is true. In other applications—for example, clinical trials of new drugs—the null hypothesis might be there is no difference in the 5-year survival rates of individuals given the new drug and those on a placebo or an existing drug. If patients on the new drug have a statistically significantly higher rate of survival, then we conclude that this difference is unlikely to have arisen by chance and the new drug is beneficial. In the present context, one does not want to require a fair prosecutor to explain their challenges very often.

form a small fraction of the overall data, as the set of possible outcomes is very small. In *Foster* or *Batson*, the possible numbers of African-Americans that could be struck were only 0–4. In this low-power situation, when a test reaches statistical significance, courts should realize that the odds of a member of the protected group being challenged by the prosecutor are substantially larger than those of the majority group; otherwise, the result would not be significant. Therefore, such a disparity in the challenge rates of the two groups should be legally meaningful, and the explanations provided by the prosecution and the "side-by-side" comparison of characteristics of the minorities struck with the majority members retained should be examined carefully.

VII. INCREASING THE POWER OF THE TEST BY INCORPORATING DATA FROM SIMILAR TRIALS.

When the statistical data suggest possible discrimination, as in *Purkett v. Elem*, but the p-value is not statistically significant even at the 0.10 level of a one-sided test,[106] in the low power setting inherent in the analysis of peremptory challenges, courts will need to place greater reliance on the non-statistical evidence. In some situations, there may be data from related cases—for example, in cases involving retrial, data on the challenges in both trials could be used to study the fairness of the prosecution's challenges. Peremptory challenge data from similar cases involving lawyers from the same District Attorney's office or law firm could also be used in some cases. The appropriate statistical method is the Cochran-Mantel-Haenszel ("CMH") test, which combines the differences between the numbers of minority members challenged and their expected number under random selection in each trial.[107] This enables the analyst to check that

106. Because in many cases the defendant will be concerned with the prosecution-removing members of their race-ethnic group a one-sided test is considered. In cases where it is plausible for a defendant to argue that members of another protected group are treated unfairly, a two-sided test is appropriate and consistent with *Powers v. Ohio*, 499 U.S. 400 (1991). When the protected group forms a small fraction of the prospective jurors, the outcomes that lead to declaring statistical significance of either a one- or two-sided test are the same. This occurred in *Batson*, *Purkett v. Elam*, *U.S. v. Forbes*, and *Foster v. Chatman* because the probability a small sample (the allowed peremptory challenges) of the venire would have no members of the protected group is greater than 0.05 or even 0.10. *See supra* Part III.

107. *See* Hogan v. Pierce, No. 79-2124, 1983 WL 30295, at *3 (D.D.C. 1983) (accepting the results of a CMH analysis of data on a sequence of promotions made from eligible minority and majority employees). The data and analysis are given in Gastwirth, *supra* note 36, at 265–67. The data was also analyzed by Robert L. Strawderman & Martin T. Wells, *Approximately Exact Inference for the Common Odds Ratio in Several 2x2 Tables*, 93 J. AMER. STATIST. ASSOC. 1294,

the data are reasonably homogenous, so that data from trials where minorities were disadvantaged and data from trials where prosecutors removed a higher fraction of non-minorities are not analyzed together.

Gastwirth & Yu and Gastwirth & Xu illustrate the use of the CMH method and the increased power one obtains from it when analyzing peremptory challenge data from several related cases, and Agresti describes its statistical properties.[108] To appreciate the increased information available and the consequent increased power of the CMH test to detect an odds ratio of 3 or more when one has data from a few related cases, suppose that there were 2 additional trials related to *Purkett*. Imagine, for example, that there were 2 codefendants being tried for the same crime, the compositions of the venires were the same, and the prosecution removed 2 of the 3 individuals of the defendant's race-ethnic group with peremptory challenges in all 3 cases. In each trial, 2 African-Americans were challenges, although only 0.5 were expected to be challenged under random selection without regard to race. Using the data from all 3 trials, the CMH test compares the observed number 6) with the expected number (1.5) and yields a p-value of 0.0054, or about 1 in 200, a highly significant result. The three trials provide a larger set of potential jurors and a larger number of challenges, which form the sample size. An appropriate test used to analyze larger sample sizes has greater power to detect a biased system.

One might think that a simpler way to analyze data from several trials would be to combine the data from each 2x2 table into a single table and apply Fisher's test. While there are situations where this is a valid procedure, the probability model underlying it is *not* appropriate for peremptory challenges because analyzing the pooled data assumes that the peremptory challenges in each trial are selected from the pool of all venires. The CMH test, however, sums the differences between the observed and expected number of minorities challenged in each one of the trials that relate to the behavior of the prosecutor and

1301–02 (1998) (confirming the data showed African-Americans were statistically significantly under-represented in promotions).

108. Joseph L. Gastwirth & Binbing Yu, *Appropriate Statistical Methodology Yields Stronger Evidence of Discriminatory Peremptory Challenges in North Carolina: Application to the Randolph County Data in* North Carolina v. Rouse *and Related Cases*, 12 L. PROBABILITY & RISK 155 (2013); Gastwirth & Xu, *supra* note 1; AGRESTI, *supra* note 44, at 231–34.

takes into account differences in the demographic composition of the venires.[109]

CONCLUSION

Legal scholars have questioned the effectiveness of *Batson* to achieve its objective of ensuring defendants are tried by a jury that represents the community, including members of minority groups.[110] Diamond and Rose recommend reducing the number of peremptory challenges,[111] while others advocate for the abolition of peremptory challenges altogether.[112] Justice Breyer suggested that he agreed with those calling for complete elimination of peremptory challenges in his concurrence in *Miller-El*.[113] Some who recommend eliminating peremptory challenges also suggest that the criteria for excusing potential jurors for cause be less stringent.[114] Recently, Professor Morrison proposed an alternative process to allow the parties to negotiate for removal of potential jurors from the venire.[115]

This Article recommends a statistical method for analyzing and interpreting the data on the peremptory challenges in a case. The procedure should assist courts when they determine whether there is a "pattern of strikes". Indeed, one can calculate the

---

109. Although we illustrated the CMH method by assuming we had three identical venires and patterns of peremptory challenges, in most cases the minority proportion of the venires in a jurisdiction will vary from trial to trial.

110. *See* Russell D. Covey, *The Unbearable Lightness of* Batson*: Mixed Motives and Discrimination in Jury Selection*, 66 MD. L. REV. 279, 311 (2007) (critiquing *Batson*'s treatment of instances where a prosecutor has "mixed motives"—some legitimate, others not—for striking a juror); Leonard L. Cavise, *The* Batson *Doctrine: The Supreme Court's Utter Failure to Meet the Challenge of Discrimination in Jury Selection*, 1999 WISC. L. REV. 501, 501 (describing *Batson* as "toothless") ; Melilli, *supra* note 14.

111. Shari Seidman Diamond & Mary R. Rose, *Real Juries*, 1 ANN. REV. L. & SOC. SCI. 255 (2005).

112. *See* Morris B. Hoffman, *Peremptory Challenges: Lawyers Are from Mars, Judges are from Venus*, 3 GREEN BAG 2D 135, 136 (2000); Charles J. Ogletree, J*ust Say No!: A Proposal to Eliminate Racially Discriminatory Uses of Peremptory Challenges*, 31 AM. CRIM. L. REV. 1099, 1104–05 (1994); Raymond J. Broderick, *Why the Peremptory Challenge Should Be Abolished*, 65 TEMP. L. REV. 369, 374–75 (1992),

113. 545 U.S. 231, 273 (2005) (Breyer, J., concurring).

114. *See* Nancy S. Marder, *Beyond Gender: Peremptory Challenges and the Roles of the Jury*, 73 TEX. L. REV. 1041, 1137 (1995); Matt Haven, Note, *Reaching* Batson*'s Challenge Twenty-Five Years Later: Eliminating the Peremptory Challenge and Loosening the Challenge for Cause Standard*. 11 UNIV. OF MD. L. J. OF RACE, RELIGION, GENDER & CLASS 97, 119–23 (2011).

115. Caren Myers Morrison, *Negotiating Peremptory Challenges*, 104 J. CRIM. L. & CRIMINOLOGY 1 (2014).

necessary statistic at the time the *Batson* issue is raised.[116] When the data indicate that the procedure has *low* power to detect a legally meaningful odds ratio, a non-significant finding should not receive much weight. In contrast, a significant finding in this instance is quite informative, as Fisher's exact test can only reach significance when there is a substantial disparity between the challenge rates. In situations where one has data on the same prosecutor's challenges in similar cases, incorporating those data into the CMH procedure, which essentially combines the results of Fisher's exact test from each trial into an overall analysis, increases the power of the statistical test. Only a *few* related cases are needed, so the focus of the inquiry does not change from the challenges in the particular case to whether there was a long-term systematic pattern of excluding minority members from jury service. In other settings where one may have a very large sample—class action cases, for example—small differences in rates or proportions can be deemed statistically significant at the 0.05 level. In this large-sample situation, one can reduce the threshold required for significance from 0.05 to 0.01 or even 0.005 and still have high power (at least .90) for detecting a legally meaningful difference.[117]

Finally, the demographic composition of the venire depends on the fairness of the process by which the jurisdiction assembles its jury pools. This is important, as having 1 or 2 additional minority members on a venire increases the power of Fisher's exact test to detect a meaningful disparity in odds ratios of the challenge rates of the different groups in the context of the jury pool at large.[118] Gastwirth et al. recommend that judges consider the effect of 1 or 2 additional minority individuals in the prospective jury pool when they consider the legal implications of a statistically significant difference between the minority fraction of the venires over several months or even a year and their fraction of the jury-eligible population.[119]

---

116. With modern statistical packages both Fisher's exact test and the distribution of the number of minority members in a sample of size equal to the number of peremptory challenges made be either party can be done in a few minutes and assist courts in their analysis of a Batson claim before the trial begins. This would avoid the current situation where courts review the events that occurred during jury selection several years after they happened.

117. *See* Gastwirth & Xu, *supra* note 1, at 302–03.

118. *See* Joseph L. Gastwirth et al., *Statistical Measures for Evaluating Protected Group Under-Representation: Analysis of the Conflicting Inferences Drawn from the Same Data in* People v. Bryant *and* Ambrose v. Booker, 14 L. PROBABILITY & RISK, 279, 296–99 (2015).

119. *Id.* at 299–301.

Currently, courts use a variety of intuitive approaches to decide whether a party has used their peremptory challenges improperly to deny members of a protected group an opportunity to serve on the jury. Using a sound statistical test would provide a more uniform and coherent approach to the problem. By analyzing data sets from a number of cases, I have shown that Fisher's exact test detects bias in the cases where courts found the prosecutor's challenges were biased and in some cases where the courts may have failed to detect biased behavior. By adopting a systematic procedure for analyzing peremptory challenge data, courts could better implement the protections afforded to citizens of all races and genders by the Court in *Batson* and its progeny.

APPENDIX: ANALYSIS OF DATA FROM OTHER CASES CONCERNING THE FAIRNESS OF PEREMPTORY CHALLENGES CONSIDERED BY THE SUPREME COURT WITH FISHER'S EXACT TEST

In this Appendix, Fisher's exact test is used to analyze the data sets from the other cases concerning the fairness of peremptory challenges considered by the Supreme Court. The data for all but two cases, *Powers v. Ohio* and *Georgia v. McCollum*, are reported either in the Supreme Court's opinion or a lower court decision.

A. J.E.B. v. Alabama ex rel. T.B.

In *J.E.B. v. Alabama ex rel T.B.*, the state filed a claim for child support on behalf of a mother of a child who allegedly was fathered by J.E.B.[120] At trial, the prosecution used 9 of its 10 challenges to remove 9 of the 10 males who remained on venire.[121] Table A.1 reports the data. Notice that one expects 3.03 males to be included in a random selection of 10 from the prospective jurors, so the prosecutor challenged 3 times as many males as would be expected if random selections were made. Fisher's exact test yields a highly significant p-value of 2.5×10^{-6}, or 1 in 400,000. The lower end of a 95% confidence interval for the ratio of the odds a male was challenged to those of a female is 8.57. This is extremely strong statistical evidence against the assumption of a gender-neutral or random selection of strikes.

Table A.1: Preemptory Challenge Data from *J.E.B. v. Alabama*

Group	Number Struck	Number Kept	Total
Male	9	1	10
Female	1	22	23
Total	10	23	33

120. 511 U.S. 127, 129 (1994).
121. *Id.*

B. Snyder v. Louisiana

In *Snyder v. Louisiana*, each party had 12 peremptory challenges; we assume that each side took full advantage of them.[122] The data are reported in Table A.2. Because there were 5 African-Americans on a panel of 36, under a "fair" or random selection process only (5/36)*12=1.667, or just under 2, African-Americans would be expected to be struck. Since the prosecutor struck all 5 African-American members of the venire, Fisher's exact test yields a p-value of 0.0021, or one in 500 for the observed outcome.[123] Thus, the observed data were unlikely to have occurred if a random selection or chance process was used by the prosecutor. As such, the data support the defendant's *Batson* claim. After a careful side-by-side comparison of the African-American jurors struck and the majority jurors not struck, the Court found in favor of the defendant.[124]

Table A.2: Prosecutor's Challenges in *Snyder v. Louisiana*

Group	Number Struck	Number Kept	Total
African- American	5	0	5
Non-African-American	7	24	31
Total	12	24	36

C. Edmonson v. Leesville Concrete Inc.

In *Edmonson c. Leesville Concrete*, a civil case, the plaintiff, an African-American who worked at Fort Polk, a federal facility, was injured when a truck operated by the defendant's firm rolled backward, pinning him against construction equipment.[125] The defendant used 2 of his three challenges to remove the 2 African-Americans on the panel from which the jury was chosen.[126]

122. 552 U.S. 472, 475 (2008).
123. *Id.* at 475–76.
124. *Id.* at 486.
125. 500 U.S. 614, 616–17 (1991).
126. *Id.*

Although the full data set is not reported, Judge Rubin's dissent from the en banc decision of the Fifth Circuit notes that each party in civil cases heard in federal court is allowed 3 peremptory challenges.[127] Thus, the minimum size of the panel of prospective jurors was 18, 15 of whom were non-African-American. The data for this panel is reported in Table A.3. The expected number of African-Americans in a random sample of 3 is 0.5.

Table A.3: Defendant's Challenges from the Smallest Possible Panel in *Edmonson v. Leesville Concrete, Inc.*

Group	Number Struck	Number Kept	Total
African- American	2	1	3
Non-African-American	1	14	15
Total	3	24	18

Fisher's exact test yields a p-value of 0.056, indicating that the probability a random draw of 3 from the pool would include 2 or 3 African-Americans is slightly over the 0.05 level of significance. Given the low power of the test in this situation, a p-value of .056 should suffice to show that the data are inconsistent with the assumption of a random or race-neutral selection process.[128] The analysis supports Justice Kennedy's opinion, which reversed the appeals court and remanded the case for consideration of the plaintiff's *Batson* claim.[129]

127. Edmonson v. Leesville Concrete, Inc., 895 F.2d 218, 232 (1990) (en banc) (Rubin, J., dissenting).

128. The probabilities of the 4 possible number of minorities amongst the prosecutor's challenges under random selection are: $P[0]=0.6676$, $P[1]=0.3860$, $P[2]=0.0551$, and $P[3]=0.00012$. If one insisted on requiring a p-value less than .05 for significance, one really would be testing using a criterion of 0.0012, which is extremely stringent because a disparity equivalent to three standard deviations from expected, noted by the Court in *Castaneda*, is 0.0027. The probability of observing 3 minorities in 3 random selections from this pool when the odds an African-American would be removed by the prosecutor were five times those of a White would only be 0.0381, or less than 4%, demonstrates the very low power Fisher's test would have if one adopted the 0.0012 level criterion. Even when one considers a p-value of 0.056 as significant, the probability of detecting an odds ratio of five would only be 0.3811, or about 40%. The determination of a legally meaningful disparity in peremptory challenges is in the province of the courts, rather than statisticians; however, it is difficult to imagine that an odds ratio of 5 would not be meaningful.

129. *Edmonson*, 500 U.S. at 631.