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*

JOSEPH L. GASTWIRTH,† WENJING XI AND QING PAN

Department of Statistics, George Washington University, Washington, DC 2006, USA

[Received on 2 October 2014; revised on 13 May 2015]

Due to an error in the computer program used to summon potential jurors for service, the African-American percentage (4.17%) of jury pools was only half their percentage (8.25%) of the age-eligible population for about fifteen months. In June 2012, in *Ambrose v. Booker*, the Sixth Circuit accepted the statistical evidence as sufficient to meet the requirements for a *prima facie* case of unfair representation the U.S. Supreme Court established in *Duren*. On the same day, the Michigan S. Ct. in *People v. Bryant* said that the statistics were insufficient. The Michigan Court accepted a new measure, the disparity of the risk (DR), of under-representation. The DR measure is shown to be extremely stringent and an alternative measure based on the probability that a jury randomly selected from the jury pool will have fewer minorities than a jury selected from the age-eligible population is proposed. The minority fraction of individuals ultimately serving on juries also depends on the fairness of the peremptory challenges made by the parties. A method for detecting unfairness is reviewed; its effectiveness depends on the number of minorities on the venire. A reanalysis of the Michigan data shows that if the proportion of African-Americans in the jury pool equaled their proportion in the age-eligible population, prosecutors could only reduce their proportion on juries to about 80% of their proportion in the population. In contrast, the criteria for adequate representativeness based on the DR measure adopted by the Michigan Court could lead to virtually no minorities serving on actual juries as only three percent of the venires would have a sufficient number of minorities to classify a prosecutor’s peremptory challenging all of them as significant. Our results indicate that when assessing statistics on the demographic mix of jury pools for legal significance, courts should consider the possible reduction in minority representation that can occur in the peremptory challenge proceedings.

*Keywords:* absolute disparity; disparity of the risk; fair representation; jury discrimination; PD-measure; peremptory challenges; selection ratio.

1. Introduction

From April 2001 until August 2002 an error in a computer program was responsible for a substantial shortfall in the number of African American residents being summoned for jury service on trials in Kent County, Michigan. Consequently, several African American defendants whose trials were held during this period appealed their convictions on the grounds that their Sixth amendment right to a jury composed of a fair-cross section of the community was violated. The same statistical studies, which compared the African American proportion, $p$, of individuals summoned for jury service when the defective computer program was in use to the African American fraction, $\pi$, of the age-eligible population for about fifteen months. In June 2012, in *Ambrose v. Booker*, the Sixth Circuit accepted the statistical evidence as sufficient to meet the requirements for a *prima facie* case of unfair representation the U.S. Supreme Court established in *Duren*. On the same day, the Michigan S. Ct. in *People v. Bryant* said that the statistics were insufficient. The Michigan Court accepted a new measure, the disparity of the risk (DR), of under-representation. The DR measure is shown to be extremely stringent and an alternative measure based on the probability that a jury randomly selected from the jury pool will have fewer minorities than a jury selected from the age-eligible population is proposed. The minority fraction of individuals ultimately serving on juries also depends on the fairness of the peremptory challenges made by the parties. A method for detecting unfairness is reviewed; its effectiveness depends on the number of minorities on the venire. A reanalysis of the Michigan data shows that if the proportion of African-Americans in the jury pool equaled their proportion in the age-eligible population, prosecutors could only reduce their proportion on juries to about 80% of their proportion in the population. In contrast, the criteria for adequate representativeness based on the DR measure adopted by the Michigan Court could lead to virtually no minorities serving on actual juries as only three percent of the venires would have a sufficient number of minorities to classify a prosecutor’s peremptory challenging all of them as significant. Our results indicate that when assessing statistics on the demographic mix of jury pools for legal significance, courts should consider the possible reduction in minority representation that can occur in the peremptory challenge proceedings.

*Keywords:* absolute disparity; disparity of the risk; fair representation; jury discrimination; PD-measure; peremptory challenges; selection ratio.

1. Introduction

From April 2001 until August 2002 an error in a computer program was responsible for a substantial shortfall in the number of African American residents being summoned for jury service on trials in Kent County, Michigan. Consequently, several African American defendants whose trials were held during this period appealed their convictions on the grounds that their Sixth amendment right to a jury composed of a fair-cross section of the community was violated. The same statistical studies, which compared the African American proportion, $p$, of individuals summoned for jury service when the defective computer program was in use to the African American fraction, $\pi$, of the age-eligible population for about fifteen months. In June 2012, in *Ambrose v. Booker*, the Sixth Circuit accepted the statistical evidence as sufficient to meet the requirements for a *prima facie* case of unfair representation the U.S. Supreme Court established in *Duren*. On the same day, the Michigan S. Ct. in *People v. Bryant* said that the statistics were insufficient. The Michigan Court accepted a new measure, the disparity of the risk (DR), of under-representation. The DR measure is shown to be extremely stringent and an alternative measure based on the probability that a jury randomly selected from the jury pool will have fewer minorities than a jury selected from the age-eligible population is proposed. The minority fraction of individuals ultimately serving on juries also depends on the fairness of the peremptory challenges made by the parties. A method for detecting unfairness is reviewed; its effectiveness depends on the number of minorities on the venire. A reanalysis of the Michigan data shows that if the proportion of African-Americans in the jury pool equaled their proportion in the age-eligible population, prosecutors could only reduce their proportion on juries to about 80% of their proportion in the population. In contrast, the criteria for adequate representativeness based on the DR measure adopted by the Michigan Court could lead to virtually no minorities serving on actual juries as only three percent of the venires would have a sufficient number of minorities to classify a prosecutor’s peremptory challenging all of them as significant. Our results indicate that when assessing statistics on the demographic mix of jury pools for legal significance, courts should consider the possible reduction in minority representation that can occur in the peremptory challenge proceedings.

*Keywords:* absolute disparity; disparity of the risk; fair representation; jury discrimination; PD-measure; peremptory challenges; selection ratio.

1. Introduction

From April 2001 until August 2002 an error in a computer program was responsible for a substantial shortfall in the number of African American residents being summoned for jury service on trials in Kent County, Michigan. Consequently, several African American defendants whose trials were held during this period appealed their convictions on the grounds that their Sixth amendment right to a jury composed of a fair-cross section of the community was violated. The same statistical studies, which compared the African American proportion, $p$, of individuals summoned for jury service when the defective computer program was in use to the African American fraction, $\pi$, of the age-eligible population for about fifteen months. In June 2012, in *Ambrose v. Booker*, the Sixth Circuit accepted the statistical evidence as sufficient to meet the requirements for a *prima facie* case of unfair representation the U.S. Supreme Court established in *Duren*. On the same day, the Michigan S. Ct. in *People v. Bryant* said that the statistics were insufficient. The Michigan Court accepted a new measure, the disparity of the risk (DR), of under-representation. The DR measure is shown to be extremely stringent and an alternative measure based on the probability that a jury randomly selected from the jury pool will have fewer minorities than a jury selected from the age-eligible population is proposed. The minority fraction of individuals ultimately serving on juries also depends on the fairness of the peremptory challenges made by the parties. A method for detecting unfairness is reviewed; its effectiveness depends on the number of minorities on the venire. A reanalysis of the Michigan data shows that if the proportion of African-Americans in the jury pool equaled their proportion in the age-eligible population, prosecutors could only reduce their proportion on juries to about 80% of their proportion in the population. In contrast, the criteria for adequate representativeness based on the DR measure adopted by the Michigan Court could lead to virtually no minorities serving on actual juries as only three percent of the venires would have a sufficient number of minorities to classify a prosecutor’s peremptory challenging all of them as significant. Our results indicate that when assessing statistics on the demographic mix of jury pools for legal significance, courts should consider the possible reduction in minority representation that can occur in the peremptory challenge proceedings.

*Keywords:* absolute disparity; disparity of the risk; fair representation; jury discrimination; PD-measure; peremptory challenges; selection ratio.
community, were submitted in virtually all of these trials. On 28 June 2012, in Ambrose v. Booker et al., the Sixth Circuit Federal Court of appeals found that the data and studies were sufficient to meet the requirements, set out by the U.S. Supreme Court in Duren v. Missouri for a defendant to establish a prima facie case that their rights to a fair jury were violated, i.e. the representation of a legally protected group on venires from which juries in Kent County were selected was not fair and reasonable in relation to their numbers in the community and this was a result of the faulty computer program. On the same day, in a 4–3 decision in People v. Bryant, the Michigan Supreme Court decided the statistical data were inadequate to show that the proportion of African Americans on juries in the County was too low relative to their representation in the community. This decision relied on the disparity of the risk (DR) measure, proposed by Detre (1994) and advocated by Re (2011), which had not previously been accepted by a court, nor examined in the statistical literature concerning jury discrimination.

The DR measure is defined as the maximum difference between the cumulative distribution functions (CDFs) of a binomial distribution with parameters $\pi$, the protected group’s fraction of the age-eligible population in the jurisdiction and their fraction, $p$, in the relevant jury pool for all values of the number, $k$, of jurors from the protected group no larger than its expected value ($n\pi$) under equal representation. If one does not restrict the range of possible values to those no larger than $n\pi$, the measure would be the classical Kolmogorov–Smirnov (KS) distance between the two binomial CDFs. Gastwirth et al. (2014) showed that when $p$ is less than $\pi$, the value of $k$, at which the maximum difference between the two binomial CDFs is attained is always less than or equal to $n\pi$, which implies that the DR measure is the KS distance. The size $n$, of the two binomial distributions depend on whether one focuses on a single jury of 12, as did Re (2011) and the Michigan Supreme Court, a grand jury of 23, as in Detre (1994), or the venire from which the defendant’s jury was selected, which in Kent County was 45 but can be higher in other jurisdictions, or on the pool of individuals who were called for service during the time (day, week, month or year) of the trial.

In Section 2, measures for assessing whether the jury pool is a fair cross-section of the community will be reviewed and illustrated on data from Duren. Absolute measures focus on the difference between the proportion, $p$, of the venire formed by the protected group and their fraction, $\pi$, of the age-eligible population. Comparative measures focus on the ratio of the probability a member of the protected group serves on a venire to either the probability a member of the non-protected group serves (selection ratio, SR) or the probability a member of the eligible population serves (comparative...
disparity, CD). Detre (1994) argues that a proper fair cross-section analysis should focus on the degree to which the defendant’s ex ante chance of a representative jury have been affected by the under-representation of a group on the jury wheel and proposed the DR measure. A related measure, probability less (PL), the probability that a venire (or jury) selected from a jury wheel on which the protected group is underrepresented will include fewer protected group members than one chosen from the broad community will also be examined here. This measure is an analog of the Mann-Whitney form of the Wilcoxon test, which has been used to examine fairness of pay (Gastwirth, 1975) and the size of an effect (difference between the two distributions) in social science (Grissom, 1994). Section 3 illustrates the use of these measures for the data submitted in the Kent County cases.

After the venire is sent to the courtroom, individuals can be removed for cause and both parties are allowed to remove potential jurors using their peremptory challenges. Bellin and Semitsu (2011 at n. 6), Burke (2012) and Justice Breyer’s concurrence in Miller-El v. Dretke, 545 U.S. 231, 267-70 (2005) cited a number of studies showing that race-based jury selection occurs frequently. Thus, even when the demographic composition of the original venire is a random selection from the community, after those proceedings, it is not clear that the demographic mix of the final jury will continue to resemble a random sample of 12 from the age-eligible community. Contrary to some legal decisions, e.g. Jenkins, 496 F.2d 57 (2d Cir. 1975) that discount the potential effect of one or two additional protected group members on venire, it will be seen that the probability of detecting discriminatory peremptory challenges often is ‘noticeably increased’ when one or two more protected group members are on the venire. In Batson, the case where the Court ordered that the defendant should receive a new trial because prosecutors peremptorily challenged all ‘four’ African Americans, had there been only two or three of them on the venire it would have been easier for the prosecutors to come up with an ‘explanation’, especially if they challenged Whites with similar characteristics. Thus, the potential effect on the probability a defendant could utilize a formal statistical test to demonstrate that a protected group was disproportionately reduced by the prosecutor’s peremptory challenges is described in Section 4.

2. Review of the statistical measures and tests used in evaluating the representativeness of jury pools

In order to illustrate the statistical procedures, it is helpful to recall the data from the major precedential case, Duren v. Missouri, which concerned the effect of the state’s jury selection process on the representation of women. As a detailed re-analysis of the data from both Duren and Berghuis v. Smith,8 a previous case concerning fair representation of African Americans on Kent County venires is given in Gastwirth and Pan (2011), this review applies the various measures and statistical tests to the data from Duren.

2.1 The statistical measures and their application to the data from Duren

The fraction, \( \pi \), of jury-eligible residents in the jurisdiction from the legally protected group is obtained from the latest Census data. According to the 1970 Census, 54% of the adult inhabitants of Jackson County were women. Data on the jury pools during June–October 1975 and January–March 1976 indicated that 11,197 persons were summoned, 2,992 or 26.7% of whom were women. In Duren,
π = 0.54, while the female fraction, \( p \), of the individuals summoned for jury service, i.e. the jury wheel, was 0.267.

The absolute disparity (AD) is the difference, \( p - \pi \), indicating the fraction by which women were underrepresented on the jury panels examined. For the jury wheel in Duren, \( 0.267 - 0.54 = -0.273 \), i.e. if the jury wheel fully reflected the female fraction of the age-eligible population, an additional 27.3\% of the jury wheel would have been women.\(^9\) Many courts require an AD of 0.10 or more before finding under-representation, however, many commentators and some cases have noted that this criterion is not appropriate for small protected groups as a protected group with less than 10\% of the adult community could be totally excluded. The absolute impact (AI) is the expected number of additional women that would have been on the venire if the proportion of women in the jury wheel equalled their proportion in the age-eligible community. Numerically, the AI is the product of the size of the venire and the AD. The CD is the ratio of the AD (\( p - \pi \)) to the protected group proportion (\( \pi \)) of the eligible population. For the jury wheel in Duren, \( \frac{0.273}{0.54} = 0.506 \). The CD 'measures the diminished likelihood that members of an underrepresented group, when compared to the population as a whole, will be called for jury service'.\(^10\) More simply, the CD measure is the shortfall in the protected group called for jury service, expressed as a percentage of their share (\( \pi \)) of the age-eligible community.

Gastwirth and Pan (2011) suggested that the ratio of the probability a member of the protected group is selected for jury service to the corresponding probability of a non-protected group member being selected, called the SR be considered. A similar ratio of the pass rate of job candidates from the protected group to the pass rate of other applicants is used to assess the possible disparate impact of a test or job requirement in Equal Employment cases. As explained in (Gastwirth and Greenhouse, 1987), the formula for the SR is:

\[
SR = \frac{p/(1-p)}{\pi/(1-\pi)}.
\]

For the Duren wheel data,

\[
SR = \frac{0.267/0.733}{0.54/0.46} = \frac{0.3643}{1.1739} = 0.310.
\]

In equal employment cases, government guidelines consider an SR less than 0.80 or ‘four-fifths’ evidence of a possible disparate impact of a test.\(^11\) Thus, a SR of 0.31 clearly indicates that women were disproportionately affected by the county’s process of summoning potential jurors. The reason the SR calculated on the Duren data is noticeably less than the CD, is because the CD is based on the ratio of the probability a member of the disadvantaged group has of being selected for the jury wheel to the corresponding probability of any member of the entire population. As women formed over half the age-eligible population if they are included in the comparison group, the probability that any member of the age-eligible population has of being on the wheel or panels is reduced. It is not statistically sound to compare a group under study with a comparison group, which also includes it. For example, in

\(^9\) Usually the negative sign is ignored so the AD is shortfall in protected group members of the jury wheel (or venires) expressed as a fraction of the eligible population in the jurisdiction.

\(^{10}\) Ramseur v. Beyer, 983 F.2d 1215, 1231 (3rd Cir. 1992).

\(^{11}\) Uniform Guidelines, 29 C.F.R. 1607 (D) (2000). If a test or job requirement has a disparate impact on a legally protected group, the employer needs to demonstrate that it is job-related.
clinical trials, patients given the new treatment are compared to a control group of similar health status who receive a placebo or older treatment. The SR avoids this problem by comparing the probability a protected group member has of being placed on the jury wheel or pool of venires to the corresponding probability of the other group.

The Batson decision also stated ‘The harm from discriminatory jury selection extends beyond that inflicted on the defendant and the excluded juror to touch the entire community. Selection procedures that purposefully exclude black persons from juries undermine public confidence in the fairness of our system of justice’. The SR measure, the ratio of the probability a member of the protected group has of being called for jury service to the corresponding probability others have, is consistent with this Batson criteria.

The Duren opinion emphasized that in order to establish a prima facie case the defendant needs to show that the representation of this group on venires from which juries are selected is not fair and reasonable in light of their proportion in the community. It examined data on the composition of venires over a period of eight months. Lower courts have not allowed a defendant to rely on data for just one venire to establish a prima facie case.

The main statistical technique used in equal protection and fair representation cases tests the hypothesis that the observed jury composition is consistent with a random sample of size n (the number of individuals on the venires or jury pools examined) from a jury-eligible population of much larger size, N, in which a fraction, π, belong to the distinct subgroup in question. Technically, the number, x, is a random variable (X) that follows the binomial model, i.e. the probability that one observes exactly x minorities in a sample of size n is given by

\[ P(X = x) = \binom{n}{x} \pi^x (1 - \pi)^{n-x}. \]  

In a random sample of n members of the eligible population, one expects \( n\pi \) to be from the protected group. On account of the randomness inherent in a sample, the actual number will deviate from the expected number by an amount measured in standard deviation units. The binomial variable, X, defined by (2) has standard deviation \( \sqrt{n\pi(1 - \pi)} \). Thus, the statistic used to test whether the observed number (x) of protected group members in the jury pool of size n is consistent with a random sample of the eligible population, is the ratio of the difference between the observed and expected numbers, \( x - n\pi \), to its standard deviation, i.e.

\[ z = \frac{x - n\pi}{\sqrt{n\pi(1 - \pi)}} \]  

In large samples the statistic (3) is approximately normally distributed. Therefore, about 95% (99) of the time, the observed number (x) will differ from the expected number (n\pi) by no more than two (three) standard deviation units. The U.S. Supreme Court accepted statistical hypothesis testing in


13 Singleton v. Lockhart, 871 F.2d 1395, 1398 (8th Cir. 1989) states that systematic under-representation cannot be demonstrated by data for a single venire; rather data on other venires must also be included. Similarly, Commonwealth v. Arriaga, 781 N.E. 2d 1253 (Mass. 2003) rejected statistics pertaining to the defendant’s venire.

14 The application of the method to data in jury discrimination cases was suggested in a classic article by Finkelstein (1966) and is now used in virtually all statistical analyses (Kaye, 1985; Gastwirth, 1988).
Castaneda v. Partida 430 U.S. 482 (1977) and indicated that values of \( Z \) or the number of ‘standard deviations’ from expected larger than two or three are suspect to social scientists. These criteria approximately correspond to the 0.05 or 0.01 levels of significance, i.e. assuming the venires are a representative random sample, the probability that a shortfall in the number of minorities on the venires at least as large as the one observed should be less than 0.05 or 0.01. The Court actually considered two-sided tests, i.e. the form of the test statistic used would detect a system in which the non-protected group was under-represented. For two-sided tests, the \( p \)-value is the probability a random sample of the jury eligible community would produce a ‘shortfall’ or ‘excess’ in the number of the protected group on the venires at ‘least as large’ as the observed shortfall or excess.\(^{15}\)

The methodology will be illustrated on the data on the jury wheel from Duren. Recall that the jury wheel consisted of 11,197 people, of whom \( x = 2992 \) were women, although females formed the fraction, \( \pi = 0.54 \), of the adult community. Notice that one expects \( n\pi \) or 11,197 \( \times \) 0.54 = 6046.38 females to be on the jury wheel, if the wheel were a random sample of the community. Substituting these values in (3) yields

\[
Z = \frac{2992 - 11197 \times 0.54}{\sqrt{11197 \times 0.54 \times 0.46}} = \frac{2992 - 6046.38}{52.738} = \frac{-3054.38}{52.738} = -57.916.
\]

This value of \( Z \) corresponds to a \( p \)-value less than one in a billion. The \( p \)-value or statistical significance of a test of a hypothesis depends on two factors. The first, the numerator of \( Z \), is the absolute difference between the observed number (2992) of women on the wheel and the expected number (6046) assuming the wheel were a random selection of the community. The second, the denominator, the standard deviation reflects the variability of the number of individuals from the protected group in the sample. Both the numerator and denominator depend on \( n \), the size of the sample examined.

Critics, e.g. Detre (1994) and King (2007), of the use of statistical testing note that in very large samples, one will classify small and legally meaningless differences between the protected group fraction (\( p \)) on the wheel or venires to their fraction (\( \pi \)) of the age-eligible community. The large sample (11 197) in Duren illustrates this issue because one standard deviation is 52.74 so a shortfall of three-standard deviations would be 158. As the expected number of females in a sample of 11 197 from the adult community was 6046.36, would a shortfall of 158 women on the jury wheel be meaningful? Probably not, as women would still form over one-half of the jury wheel. Gastwirth and Xu (2014) show that by requiring a lower probability threshold for statistical significance, one can resolve this problem and still detect meaningful shortfalls.\(^{16}\)

\(^{15}\) It is important to note that the calculations involved in determining the \( p \)-value of a statistical test and whether it is sufficiently small to be statistically significant at the 0.05 (or 0.01) level are made assuming the null hypothesis is true, i.e. the government is selecting jurors in accordance with a random sample of the community. The \( p \)-value is ‘not’ the probability that an unusually low number of minorities appeared on the venires examined occurred by chance. The \( p \)-value is the probability that if the venires were chosen by chance or random sampling of the adult community, one would observe a difference between the observed and expected numbers of the protected group at least as great as in the data. Common sense tells us that events of low probability do not occur very often, so the occurrence of a ‘rare’ event makes us doubt that the venires were a random sample of the adult community.

\(^{16}\) They demonstrate that if one has a very large sample, as in Duren one can set a lower threshold such as a \( p \)-value less than 0.001 and still have a high probability, e.g. 0.99, of detecting a meaningful disparity.
2.2 The DR measure

Detre (1994) introduced the DR measure after observing that neither AD nor AI measure how the defendant’s chances of obtaining a representative jury panel were affected by the under-representation. He argued that statistical testing is not appropriate for a fair cross section analysis because the probability the composition of a jury wheel arose by random selection from the community is not directly related to the defendant’s chances of drawing a jury of a certain composition. As emphasized previously, the p-value of a statistical test is ‘not’ the probability the composition of the jury wheel arose by random selection.\(^{17}\) It is the probability a random sample of the age-eligible residents of the jurisdiction would result in a ‘shortfall’ on the jury wheel as large as or larger than the actual shortfall.\(^{18}\)

To define the DR measure, recall that when the protected group forms the fraction, \(\pi\), of the community, the number from the protected group in a random sample of size \(n\) from that community follows a binomial model, denoted by \(B(n, \pi)\), where the probability of observing exactly \(k\) protected group members among the \(n\), is given by (2). When the protected group fraction on the large jury wheel is \(\mu\), less than \(\pi\), the number of minorities on a grand jury or venire of size \(n\) now follows the binomial distribution \(B(n, \mu)\) as each potential juror now has probability \(\mu\) of being a protected group member. For each possible value, \(k\), of the number from the protected group in a sample of \(n\), Detre considers the ‘difference’ between the probability that \(k\) ‘or fewer’ jurors from the protected group are randomly selected when a member of the sample has probability \(\mu\) of being in the protected group and the corresponding probability, when they have probability \(\pi\), of being in the protected group. For each value of \(k\) up to \(n/\pi\) (the expected number of minorities in a sample of \(n\) from the community), one computes the difference \(D(k)\) of these two probabilities. The maximum value of \(D(k)\) is the DR and occurs at \(k^*\), the value of \(k\) for which the difference in the two binomial probabilities of obtaining \(k\) or fewer protected group members is greatest. Both Detre and Re assert that this measures the amount by which under-representation on the wheel increases the defendant’s risk of an unrepresentative jury. Based on his study of a grand jury of size, \(n = 23\), Detre recommends a value \(D(k^*)\) of 0.37 or 37% as a threshold. Re recommends a threshold value of 0.50 or 50%, based on studies of petit juries of size 12 and states that this threshold parallels the preponderance of the evidence standard, i.e. eliminating a DR of at least 50% would more likely than not change the composition of the defendant’s jury.\(^{19}\) In contrast to the preponderance of the evidence standard, which is the criteria the jury should apply when it considers the ‘totality’ of the evidence in a civil case, when a defendant establishes a prima facie case of under-representation, the government is required to produce an explanation, i.e. in a fair representation case that the process that led to the under-representation served a significant state interest.\(^{20}\) To require the defendant to produce statistical evidence of such strength that by itself would meet the ‘preponderance’ of the evidence standard just to establish a prima facie case, based on his study of a grand jury of size, \(n = 23\), Detre recommends a value \(D(k^*)\) of 0.37 or 37% as a threshold. Re recommends a threshold value of 0.50 or 50%, based on studies of petit juries of size 12 and states that this threshold parallels the preponderance of the evidence standard, i.e. eliminating a DR of at least 50% would more likely than not change the composition of the defendant’s jury.\(^{19}\) In contrast to the preponderance of the evidence standard, which is the criteria the jury should apply when it considers the ‘totality’ of the evidence in a civil case, when a defendant establishes a prima facie case of under-representation, the government is required to produce an explanation, i.e. in a fair representation case that the process that led to the under-representation served a significant state interest.\(^{20}\) To require the defendant to produce statistical evidence of such strength that by itself would meet the ‘preponderance’ of the evidence standard just to establish a prima facie case,

\(^{17}\) In U.S. v. Hernandez-Estrada, 749 F. 3d 1154 (9th Cir. 2014) at 1164, unfortunately, the court quoted this statement from the article by Detre.

\(^{18}\) This definition is for a one-tailed test, which focuses on under-representation of the protected group. It is appropriate when there has been a history of discrimination. Otherwise, courts use a two-tailed test as the majority group could have been under-represented in the jury pool. Then the p-value is the probability that the absolute value of the difference between the observed and expected number of protected group members in the jury pool is equal to or greater than the observed absolute difference.

\(^{19}\) Re (2011) at 542–43. As noted previously this formulation ignores the voir dire process and the removal of members of the venire for cause or by peremptory challenges.

\(^{20}\) The Duren opinion, supra n. 3 at n. 25, observed that the federal court in the jurisdiction allowed for the exemption of individuals working in several occupations as well as a child-care exemption. This led to a female representation on jury panels of
would allow jurisdictions to limit protected group participation in the jury process to a substantial degree, although not completely, with impunity.

Neither proponent of the DR measure discusses the meaning of \( k^* \), the number of the protected group on the jury or venire at which the difference between the two cumulative probability distributions is largest. Thus, \( k^* \) can be considered the dividing point; values of \( k \leq k^* \) have higher probability of occurring in a sample following a binomial \( B(n, p) \) variable than in a sample following a \( B(n, \pi) \) one, while the reverse is true for values of \( k > k^* \). This reflects the fact that a smaller number from the protected group is more likely to be on a venire when it is chosen from a pool with a protected group fraction, \( p \), than a venire chosen from a pool having a larger protected group fraction, \( \pi \).

To illustrate the DR measure, consider a situation similar to Duren where \( \pi = 0.50 \) and \( p = 0.30 \), i.e. the protected group forms 50% (30%) of the jury eligible in the jurisdiction. The probabilities, \( D(k) \) for \( k = 0–6 \) when \( n = 12 \) are given in Table 1.

Table 1 shows that the maximum value of \( D(k) \) is 0.53, which exceeds both the 0.37 and 0.50 thresholds advocated by Detre and Re. Of course, the absolute disparity is 0.20, clearly exceeding the value 0.10, which suffices to establish a prima facie case in many jurisdictions. The CD is 0.40, and the SR is 0.43. This SR implies that the probability a woman has of being on a venire is only 43% of the probability a man has. When the number \( n \), of individuals on the wheel or potential jury pool from which \( p \) was calculated was at least 30, the disparity would be statistically significant. Thus, all of the measures and methods will classify the system as unrepresentative.

More interesting is the situation considered by Detre, where \( p = 0.40 \) and \( \pi = 0.5 \), which might occur in a case concerning the fair representation of women. If a jury of 12 is examined, the DR is 0.278, well below 0.37 and 0.50. The point, \( k^* \), of maximum difference is 5, i.e. the probability of having 5 or fewer women is 0.665 if women form 40% of the jury wheel but is just under 40%. Thus, a modest reduction shortfall in the representation of a protected group on jury pools that arose from a legitimate need did not violate the Sixth Amendment.

21 If one observed only 9 protected group members in a sample of 30 members of a population in which 50% were protected members, the shortfall of 6 protected group members from the expected number 15, would be statistically significant at the 0.05 or two-standard deviation level. Clearly, 30 is a much smaller sample size than the number of individuals on the venires that are compared to the demographic composition of the jury eligible community.
0.387 if women form 50% of the wheel. The absolute disparity, $0.4 - 0.5 = -0.10$ or 10%, is just at the value courts relying on this measure have considered meaningful.\(^{22}\) The CD, $(p - \pi)/\pi = -0.1/0.5 = 0.20$, while the SR is two-thirds or 0.667. Thus, adult women in the area have only ‘two-thirds’ the chance of becoming a juror as a man, which would be considered meaningful in most applications.

To guide his choice of a threshold value of the DR measure, Detre (1994) considered a grand jury of 23 being chosen from a jury pool with 40% female but their fraction of the age-eligible population was 50%. While the AD, CD and SR values remain the same as before, the DR becomes 0.37 and the point of maximum disparity, $k^* = 10$. Next consider a venire of 45, as in the Kent County cases. Again, the AD, CD and SR measures remain 0.10, 0.20 and 0.67, respectively. Now the DR is 0.50, exceeding both recommended thresholds (0.37 and 0.50) and $k^*$ is 20.

These examples show that the values of the maximum disparity and the corresponding point ($k^*$) where it occurs depend on whether data about a petit jury of 12, a grand jury of 23 or a venire of 45 is examined. Thus, the sample size or number of individuals from the large jury wheel examined affects the value of the DR measure. In large samples, a small difference between two probabilities will often be statistically significant at the commonly used 0.05 level and will satisfy the DR > 0.50 criteria. For example suppose a case concerned the fair representation of women, i.e. $\pi = 0.50$, and suppose $p = 0.45$, so the AD = 0.05, the CD = 0.10, while the SR = 0.818; none of them indicating substantial unfairness. Assume there were 1000 potential jurors on the venires examined. The usual statistical test (3) indicates that the 450 women among the 1000 are statistically significantly less than the 500 expected ($Z = 3.16, p\text{-value} = 0.002$) and the DR measure for this jury pool is 0.887, clearly exceeding 0.50. The point ($k^*$) of maximum disparity is 474, which is close to the expected number (500) of women when the jury pool is drawn from the community. Indeed, if 474 of the 1,000 members of the jury pool were from the protected group, corresponding to a shortfall of 26 from their expected number 500, the shortfall would ‘not’ be statistically significant.\(^{23}\) Thus, the value, $k^*$, does not have relevance for assessing whether women or minorities were so under-represented in the jury pool that a defendant could not obtain a fair trial.

Gastwirth et al. (2014) show that in the jury discrimination context, the DR measure is just the KS distance between two binomial distributions with the same sample size, $n$, but different success probabilities. In one, the success probability is $p$, the protected group fraction in a large sample of individuals either summoned for jury service or served on a venire, while in the other the success probability is $\pi$, the protected group fraction of the residents of the jurisdiction who are age-eligible for jury duty. Consequently, DR is ‘not’ a probability or risk of obtaining an unfair jury or venire. It is the maximum difference in the probabilities of obtaining $k$ or fewer jurors from the protected group over all possible values of $k$, of the two different binomial distributions. Statistical investigations of the power of the KS test show that as the sample size increases $k^*$ will approach

\(^{22}\) The Court found an AD of 23% sufficient to establish a prima facie case in Turner v. Fouche, 396 U.S. 346 (1970) and an AD of 14% sufficed in Hernandez v. Texas, 347 U.S. (1954). An AD of 8.69% between the Hispanic proportion of the age-eligible population and their proportion in venires, however, was insufficient to establish a prima facie case in Butler v. United States, 611 F.2d. 1066, 1069–70 & n. 9 (5th Cir. 1980). Recently, in U.S. v. Hernandez-Estrada, 749 F.3d 1154, 1164 (9th Cir. 2014), US App LEXIS 8139, the Ninth Circuit changed its policy of using the absolute disparity criteria as the sole measure of under-representation, noting that any minority group forming less than 7.7% (the largest minority in Montana) of the community could be completely excluded.

\(^{23}\) Now equation (2) becomes $-26/15.81 = -1.64$, which is less than the two standard deviation criterion.
$n\pi$, its expected value in a sample from the community,\(^ {24}\) confirming that in large samples, the value of $k^*$ is not informative.

When focusing on a jury of 12, the special case where $k^* = 0$, i.e. the maximum disparity occurs when there are ‘no’ jurors from the protected group may have greater practical importance. Here, defendants from the protected group would benefit by a modest increase, say 0.20 in the probability of having ‘at least’ one member from the protected group on their jury. In this situation, a DR of 0.20 implies that in an additional 20 out of 100 trials there will be at least one member of the jury who belongs to the protected group when juries are drawn from the community, with a protected group fraction $\pi$ rather than a jury wheel with a fraction of the protected group, $p$. King (1993) showed that the race and ethnicity of jurors does influence jury decisions and a mock jury study (Sommers, 2006) found that the decisions of Whites are influenced by the presence of Blacks on the jury, so a moderate increase in the probability of obtaining ‘at least’ one juror from the protected group should be important, not only for increasing public confidence in the fairness of the system but also in rendering better decisions.

### 2.3 The probability a jury has fewer jurors from the protected group than one chosen from a sample of the jury-eligible population

The well-established Wilcoxon test (Lehmann, 1975) for the comparison of two distributions is based on the probability, $P$, that a randomly selected member of the first population is at least large as a randomly selected member of the second population. When the characteristic being measured has a continuous distribution, e.g. height, blood pressure, which is the same in both populations, this probability is 0.50.\(^ {25}\) If the first population is shifted to the right of the second, i.e. tends to be larger, then this probability is ‘greater’ than 0.5. The reverse is true when the first population is shifted to the left.

In the context of jury discrimination, the probability a jury of 12 taken from the general age-eligible population follows a binomial distribution with $n = 12$ and success probability $\pi$, the protected group fraction of the age-eligible population. The actual jury is selected from a large pool, with a protected group fraction $p$, typically less than $\pi$ and the number of protected group members on the actual jury follows a Binomial (12, $p$) distribution. Protected group under-representation will be reflected by the probability a jury of 12 randomly selected from the jury pool contains ‘fewer’ minorities than a jury of 12 chosen from the age-eligible population, defined as the probability less (PL) measure.\(^ {26}\) Because the binomial distribution is discrete, i.e. the number from the protected group on a jury of 12 can only be an integer between 0 and 12 and it is possible for a random sample of 12 from the large jury pool, with a fraction of the protected group, $p$, to
have the same number as a random sample of 12 from the age-eligible population of the community. Because a ‘tie’ in the outcomes of the two binomial distributions would not indicate protected group under-representation, they are excluded in calculating PL.

It is possible for a sample of 12 taken from a large pool with protected group fraction, \( p \), to have more protected group members than an independent sample taken from a pool with protected group fraction, \( \pi \), even if \( p \) is less than \( \pi \). This probability, \( PM \), is the expected fraction of juries where the protected group is over-represented. The difference \( PD = PL - PM \) is the fraction of juries of 12 with fewer protected group members than would occur if the jury pool represented the community, after accounting for the smaller fraction of juries that would have more minorities. When the protected group is fully represented on the jury wheel, i.e. \( p = \pi \), \( PM = PL \), so \( PD = 0 \). Values of \( PD \) greater than zero indicate the net fraction of juries with fewer minorities than would occur if the juries were chosen from the community.\(^{27} \) The choice of threshold value of PL or PD should be made by the courts and might depend on whether one is considering a jury of 12, a grand jury of 23 or the entire venire.\(^{28} \)

Because the protected group’s proportion, \( p \), of the jury pool is considered to be a random sample from the community, it is a random variable. Consequently, PL and PD also are random variables and ordinarily one would take this sampling error into account. In most jury discrimination cases, courts examine the demographic composition of a large number of potential jurors, so the sampling error of PL or PD will be small and is ignored here.

It is useful to compare the PL and PD measures to the DR, SR, CD and AD measures when \( \pi = 0.5 \) and when \( \pi = 0.20 \), say for values of \( p \) less than \( \pi \). These values are given in Tables 2 and 3, respectively. Detre (1994) considered the case where \( p = 0.4 \) and \( \pi = 0.5 \). Table 2 shows that \( PL = 0.616 \), implying that 60% of defendants had juries with fewer minorities than they would have had if minorities were fully represented in the jury pool. The PD of 0.375 shows that a net shortfall of the protected group occurs in three-eighths of the trials. These large values of PL and PD are consistent with the values, 0.667 of SR and an AD of 0.10 that also indicate under-representation.

The results in Table 2 show that when the protected group is one-half of the eligible population, the DR measure does not reach the 0.50 threshold adopted by the Michigan Supreme Court until \( p \) is below one-third. When \( p = 0.333 \), the AD = 0.167, and the SR = 0.500, implies that members of the protected group have ‘one-half’ the probability of serving on a venire as those from the majority. The value 0.738 of the PL measure indicates that nearly three-fourths of the juries of 12 chosen from the actual jury pool will have ‘fewer’ protected group members than a jury of 12 randomly selected from the eligible population and the PD of 0.591 indicates that there will be a net shortfall in nearly 60% of the juries. Therefore, courts requiring a DR of at least 0.50 are imposing a far more stringent statistical standard than requiring an AD of at least 0.10 and would substantially diminish the possibility of detecting under-representation of the protected group on juries.

When the protected group forms 20% of the eligible population, Table 3 shows that requiring a DR of at least 0.50 or an AD of 0.10 in this situation would ‘not’ classify a system in which only 12% of those called for jury service were from the protected group as unrepresentative. In contrast, the SR of 0.55 and a PD nearly 0.40 indicate that the protected group is noticeably disadvantaged with respect to jury service. Again, the extreme stringency of the requirement that the DR exceed 0.50 is illustrated

\(^{27} \) The net fraction reduces PL by the probability that \( X_1 > X_2 \).

\(^{28} \) As noted previously, the jury selection process may further reduce the protected group fraction of actual jurors, so using a binomial distribution with \( n = 12 \) and protected group proportion \( p \) to compare with the corresponding distribution with protected group proportion, \( \pi \), is likely to under-estimate the unrepresentativeness of actual juries. The effect of peremptory challenges on the composition of the final jury is discussed in Section 4.
when \( p = 0.10 \). Now even though the \( AD = 0.10 \), the \( SR = 0.44 \), \( PL = 0.66 \) and \( PD = 0.50 \), the \( DR \) measure is only 0.38.

### 2.4 Application of the DR measure for comparing two normal distributions and implications to disparate impact cases

Most people are more familiar with the normal or ‘bell-shaped’ distribution than the binomial, e.g. the distribution of IQ scores is standardized to have mean 100 and standard deviation 15. Thus, it is insightful to use the DR measure to compare two normal distributions. In this subsection we consider two random variables \( X_1 \) and \( X_2 \), which follow \( N(0,1) \) and \( N(\mu,1) \) \((\mu \geq 0)\) distributions, respectively.\(^{29}\)

Re’s criteria that the DR or \( \sup_x |F(x) - G(x)| \) occurs at the value of \( \mu \) satisfying \( \Phi\left(\frac{\mu}{\sigma}\right) = .5 \) which is \( \mu = 1.349 \).\(^{30}\) The density functions corresponding to standard normal distribution with mean 0 and one with mean 1.349 are plotted in Fig. 1. Visually one can see that these two densities are quite far apart. A similar calculation shows that Detre’s requirement that the DR is at least 0.37 or

\(^{29}\) A normal distribution with mean \( m \) and standard deviation \( s \) can be transformed into a variable, \( Z \) that is \( N(0,1) \) by subtracting the mean and dividing by the standard deviation. For an IQ score of 130, say, \( Z = (130 - 100)/15 = 2 \).

\(^{30}\) See Lemma 2 of Gastwirth et al. (2014) giving the proof that the DR measure is \( 2\Phi(\theta/2) - 1 \) and occurs at \( x = \theta/2 \).
sup \| F(x) - G(x) \| = 0.37 implies that the mean, \( \mu \), of the second distribution is 0.963. This criterion corresponds to a difference between the means of the two distributions of nearly one standard deviation. This value (0.963) of an ‘effect size’\(^{31}\) is also quite large.

In the context of IQ scores, if one used the DR criteria to decide whether the scores of two groups were different, the difference in the averages scores corresponding to a DR = 0.50 (0.37) is 20.24 (14.5). These values either exceed or are near the limits of the difference between the IQ scores of racial group found in a number of controversial studies.\(^{32}\)

The stringency of the proposed values, 0.37 or 0.50, of the DR measure can be seen by considering their implications for evaluating whether a pre-employment test has a disparate impact. For many years the government has done this by comparing the SR, i.e. the ratio of the pass rate of the protected group to the pass rate of the majority applicants, to ‘four-fifths’ or 0.80.\(^{33}\) In most cases, only when the SR is below 0.80, will the employer need to demonstrate that the requirement or test is job-related, i.e. predictive of performance of the job.

Suppose applicants must pass an exam and that the scores of each group follow a ‘bell curve’ or normal distribution with the same standard deviation (\( \sigma \)) but different means. After standardization, the scores in the protected group follow a \( \mathcal{N}(0,1) \) distribution, while the majority scores have a \( \mathcal{N}(\theta,1) \) distribution. Thus, \( \theta \) is the difference in means of the two score distributions, expressed in terms of the standard deviation of the exam scores. Because the SR depends on the cut-off value for passing the exam, Table 5 presents the SRs for a range of pass rates of the non-protected group when the differences (\( \theta \)) between the average scores of the two groups are 1.349 and 0.963, which correspond to DR

\(^{31}\) A (population) effect size \( \tau = \frac{\mu_1 - \mu_2}{\sigma} \) is the standardized difference between the means of two populations, where \( \mu_1 \) is the mean for one population, \( \mu_2 \) is the mean for the other population, and \( \sigma \) is a standard deviation based on either or both populations. The effect size is a statistical measure of the effect of a new treatment or method in the psychological and education literature (Ellis, 2010, p. 41). Effect sizes of 0.8, 0.5 and 0.2 are considered large, moderate and small, respectively.

\(^{32}\) For example, Rushton and Jensen (2005) report differences in IQ scores of racial groups up to 1.1 standard deviations or 16.5 points. Hartigan and Wigdor (1989) note, at page 263, that the difference in averages of minority and majority individuals taking the General Aptitude Test Battery is about 0.9 standard deviations. They note, at page 106, that part of the difference in average scores is likely due to minority examinees having less experience with standardized tests and less opportunity to practice.

\(^{33}\) Uniform Guidelines, 29 C.F.R. 1607 (D), (2000).
values of 0.50 and 0.37, respectively. Notice that for common situations, where the pass rate of the non-protected group is in the 20–90% range, the SRs corresponding to a DR of 0.50 are less than 0.60, while the SRs corresponding to a DR of 0.37, are all less than 0.70. Throughout this wide range of pass rates of the non-protected group, the SR ratios corresponding to the DR measure utilizing the criteria of either Detre or Ray are ‘always less’ than the government’s ‘four-fifths’ guideline.

To appreciate how much more stringent the proposed DR criteria are in the disparate impact context, consider the data in Connecticut v. Teal\(^{34}\) and Lewis v. Chicago.\(^{35}\) In Teal, the Court found that the difference between the 54.2% pass rate of Blacks and the corresponding 79.5% rate of Whites sufficed to establish a \textit{prima facie} case of disparate impact.\(^{36}\) The results in Table 4 show that when the pass rate of the White is 80%, the DR > 0.37 (0.50) criteria implies that the Black’s pass rate would need to be ‘less’ than 45.2% (30.6%) to meet those values of the DR. Disparities of this magnitude are far ‘greater’ than the one the Supreme Court accepted as sufficient in Teal as well as the ‘four-fifths’ rule. Although the legal issue in Lewis focused on when an individual who has been affected by an exam with a disparate impact can file a suit, the underlying data reported by the District court is illuminating. The City established two ‘passing’ scores; one (89) for an applicant to be ‘highly qualified’ and one (65) for being qualified. While the scores of 12.6% of White applicants met the threshold for being ‘highly qualified’ only 2.2% of the Blacks did. In contrast, 93.45% (72.3%) of Whites (Blacks) met the ‘qualified’ standard. The rows of Table 4 for a majority pass rate of 0.125 show that the clear disparity in the percentages of ‘highly qualified’ scores in the two groups, with a SR of only 0.175, would ‘not’ meet either the DR >= 0.50 or DR >= 0.37, criteria as 2.2% ‘exceeds’ the values, 0.6% and 1.7% required.

\(^{34}\) 457 U.S. 440 (1981).
\(^{35}\) 560 U.S. 205; 130 S. Ct. 2191 (2010), reversing 528 F. 3d 488 (7th Cir. 2008).
\(^{36}\) The data is also statistically significant at the usual 0.05 or 0.01 levels as normal form of the standard test for the difference between two proportions yields a \(Z = -3.76\) (\(p\)-value 0.01), See Gastwirth (1988) at pages 204 and 213 for the data and details of the calculation.

### Table 4 Selection ratios corresponding to DR meeting the criteria of Detre and Ray

<table>
<thead>
<tr>
<th>Disparity Risk (DR)</th>
<th>Majority pass rate ((p_1))</th>
<th>Minority pass rate ((p_2))</th>
<th>Selection ratio ((\Psi = \frac{p_2}{p_1}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.50</td>
<td>1.349</td>
<td>0.95</td>
<td>0.616</td>
</tr>
<tr>
<td></td>
<td>0.90</td>
<td>0.473</td>
<td>0.526</td>
</tr>
<tr>
<td></td>
<td>0.80</td>
<td>0.306</td>
<td>0.383</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
<td>0.250</td>
<td>0.333</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>0.089</td>
<td>0.178</td>
</tr>
<tr>
<td></td>
<td>0.20</td>
<td>0.014</td>
<td>0.070</td>
</tr>
<tr>
<td></td>
<td>0.125</td>
<td>0.006</td>
<td>0.048</td>
</tr>
<tr>
<td>0.37</td>
<td>0.963</td>
<td>0.95</td>
<td>0.752</td>
</tr>
<tr>
<td></td>
<td>0.90</td>
<td>0.625</td>
<td>0.694</td>
</tr>
<tr>
<td></td>
<td>0.80</td>
<td>0.452</td>
<td>0.565</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
<td>0.386</td>
<td>0.515</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>0.168</td>
<td>0.336</td>
</tr>
<tr>
<td></td>
<td>0.20</td>
<td>0.036</td>
<td>0.180</td>
</tr>
<tr>
<td></td>
<td>0.125</td>
<td>0.017</td>
<td>0.136</td>
</tr>
</tbody>
</table>
Table 5 presents the value of the DR measure and the difference, \( \theta \), between the mean scores of the two groups that correspond to pass rates that just meet the ‘four-fifths’ rule or just meet a stricter ‘three-fifths’ rule. For \( \theta = 0.80 \), considered a large ‘effect size’, for ‘none’ of the pass rates combinations considered does the DR reach 0.37, much less 0.50. Only when a very high percentage of White applicants pass the test, will the DR criteria suggested by Re or Detre be met. For example, if \( \theta = 1.468 \), an effect size of almost 1.5 standard deviations, which corresponds to the situation when White (95%) applicants pass an exam but only 57% of the Black members pass, would the DR exceed 0.50.

The stringency of the requirement of a DR of at least 0.37 (0.50) can also be illustrated by the ‘effect size’ it corresponds to. Recall that social scientists define the size of an effect of a treatment by the standardized difference between the means of two groups (Ellis, 2010). Although there are several variants of the definition, because the two normal distributions used for illustration have the same standard deviation, they are equivalent. The standardized differences 0.963 and 1.349 can be translated into the probability a White member has a higher score than a randomly selected Black member. Grissom (1994) shows that the standardized difference 0.963 (1.349) correspond to values of the PL measure 0.75 (0.83). Clearly, if ‘three-fourths’ of the juries in a jurisdiction have fewer Blacks on them than would be expected if juries were random samples of the community, this would be a very large effect.

In some contexts much smaller effect sizes or disparities of the risk are quite meaningful. Motivated by the negative effect on children’s mental abilities of exposure to lead, suppose another toxic substance children were exposed to had the effect of reducing their average adult IQ by five points; this corresponds to an effect size of 0.333. From the formula in footnote 31 this leads to a DR of just 0.132. The impact of a reduction in the average IQ of the population by five points, however, has important consequences. For example, the fraction of individuals with an IQ of 70 or less would

![Table 5](image-url)

Table 5 presents the value of the DR measure and the difference, \( \theta \), between the mean scores of the two groups that correspond to pass rates that just meet the ‘four-fifths’ rule or just meet a stricter ‘three-fifths’ rule. For \( \theta = 0.80 \), considered a large ‘effect size’, for ‘none’ of the pass rates combinations considered does the DR reach 0.37, much less 0.50. Only when a very high percentage of White applicants pass the test, will the DR criteria suggested by Re or Detre be met. For example, if \( \theta = 1.468 \), an effect size of almost 1.5 standard deviations, which corresponds to the situation when White (95%) applicants pass an exam but only 57% of the Black members pass, would the DR exceed 0.50.

The stringency of the requirement of a DR of at least 0.37 (0.50) can also be illustrated by the ‘effect size’ it corresponds to. Recall that social scientists define the size of an effect of a treatment by the standardized difference between the means of two groups (Ellis, 2010). Although there are several variants of the definition, because the two normal distributions used for illustration have the same standard deviation, they are equivalent. The standardized differences 0.963 and 1.349 can be translated into the probability a White member has a higher score than a randomly selected Black member. Grissom (1994) shows that the standardized difference 0.963 (1.349) correspond to values of the PL measure 0.75 (0.83). Clearly, if ‘three-fourths’ of the juries in a jurisdiction have fewer Blacks on them than would be expected if juries were random samples of the community, this would be a very large effect.

In some contexts much smaller effect sizes or disparities of the risk are quite meaningful. Motivated by the negative effect on children’s mental abilities of exposure to lead, suppose another toxic substance children were exposed to had the effect of reducing their average adult IQ by five points; this corresponds to an effect size of 0.333. From the formula in footnote 31 this leads to a DR of just 0.132. The impact of a reduction in the average IQ of the population by five points, however, has important consequences. For example, the fraction of individuals with an IQ of 70 or less would

---

37 Even controversial research, summarized by Rushton and Jensen (2005), on differences in IQ scores of racial groups, suggests that the gap is only 1.1 standard deviations.

38 The three most commonly used measures of the size of the effect corresponding to a difference between the means of two groups are due to Cohen, Glass and Hedges (Ellis, 2010, p. 10). They differ in the way they compute the estimate of the standard deviation from the data. Cohen and Hedges combine the standard deviations of the two groups slightly differently, while Glass uses the standard deviation computed from the ‘control’ group, which would be the majority in the discrimination context.
change from 0.0228 to 0.0478; while the fraction of individuals with an IQ of 130 or more would decrease from 0.0228 to 0.01. The net impact on society of having the fraction of highly talented individuals drop by more than 50% and having the fraction of mentally retarded\textsuperscript{39} double would be substantial. The value 0.132 of the DR measure is about ‘one-fourth’ of the 0.50 value adopted in Bryant and would discount a very noticeable impact on the cognitive ability of the next generation.

3. Analysis of the data submitted in the Kent County cases

In most cases concerning the fairness of representation of protected groups on venires or the entire jury wheel from which venires are chosen, one knows the ‘actual’ number of protected group members and the total size of the venires or wheels examined. Then the protected group fraction ($p$) of the composition of venires or wheels is compared with the corresponding fraction ($\pi$) of minorities in the eligible population of the jurisdiction, which is obtained from the latest decennial Census.

In the Kent county cases, the actual number of African Americans serving on the venires examined was ‘not known’ because the master list of potential jurors was created from individuals in the Department of Motor Vehicle list of people with a driver’s license or identification card. The statistical experts in the ‘computer glitch’ cases followed the approach defendant’s expert used in Berghuis v. Smith, which was described in Gastwirth and Pan (2011). Here we will analyze the data using the standard binomial model summarized in Section 2.

Defendant’s expert analysed data for the 3898 individuals who were summoned for jury service during the first 3 months of 2002, a sub-period of the time the computer error affected the jury system in Kent County. According to the 2000 Census, Blacks form 8.25% of the population 18 years or older in Kent County. The expected number of Blacks among the 3898 summoned was 321.16; however, only 4.17% of those called for jury service were African American. Thus, there were 163 African Americans on the venires during January–March, 2000. The usual test statistic (2) is:

$$Z = \frac{163 - 3898 \times .0825}{\sqrt{3898 \times .0825 \times .9175}} = -9.23.$$ 

This difference of nine standard deviations clearly exceeds the two to three standard deviation units criterion used in equal protection cases, e.g. Castaneda v. Partida (1977). Indeed, the probability that a random sample of the age-eligible community would result in as few or fewer than the 163 African Americans summoned when 321 would be expected, is less than one in a million. In addition to the results of a statistical test, the basic measures (AD, CD and SR) of the differential should be calculated. Here, the AD = −0.0408, the CD = −0.4945 and the SR = 0.4839. The SR tells us that African Americans had just under ‘one-half’ the probability of being ‘summoned’ for jury service as Whites. By definition a random sample of a population is one in which every member of the population has the same probability of being included. Of course, in the real world it is difficult to ensure that every member of a large population has exactly the same probability of being included in a particular sample, however, when some individuals have ‘twice’ the probability of being in the sample as others, the resulting sample is not representative. This is especially so, when the individuals with the much lower probability of being sampled belong to an identifiable subset of the larger population.

\textsuperscript{39} In Atkins v. Virginia, 536 U.S. 304 (2002) the U.S. Supreme Court decided that a state could not execute a defendant who was seriously mentally retarded. He had an IQ of 59. Recently, in Hall v. Florida, 134 S. Ct. 1986 (2014) US LEXIS 4373, the Court disallowed the state’s classifying a defendant as mentally retarded solely on the basis that his IQ was 70 or less.
Table 6 reports the statistical measures for several data sets relevant to defendant’s trial in People v. Bryant, i.e. a jury of 12, the actual venire of 45 and the 132 who appeared for jury duty on the day of defendant Bryant’s trial. As above, the African American fraction, \( \pi \), of the age-eligible population is 0.0825; their corresponding fraction of those summoned during the three months studied was 0.0417. The opinion reports that on the defendant’s venire, there was one African American and one Hispanic.

The results in Table 6 show that the DR for a jury of 12 or a venire of 45 is less than 0.50, the value adopted by the Michigan Court for a jury of 12, however, it exceeds 0.50 for the jury pool on the day of the trial. Notice that the size of the sample analysed has a strong influence on the values of DR, PL and PD, while AD, CD and SR are not affected. For a jury of 12, the point of maximum disparity occurs at \( k = 0 \), i.e. when there are ‘no’ African Americans are on the jury. If a jury of 12 were randomly drawn from a large wheel in which 8.25% were African American, the probability of obtaining no African Americans is 0.356. When only 4.17% of the wheel is African American, the probability of obtaining no African Americans is 0.600. This implies that the probability a defendant has at least one member of their race-ethnic group on their jury, increases by 0.24, which seems substantial. Furthermore, the PL measure is 0.473, implying that ‘nearly one-half’ of the juries will contain ‘fewer’ African Americans when they are chosen from a pool with an African-American fraction of 4.17% than they would have had if the jurors were chosen from the community, with an African-American fraction of 8.25. The PD of 0.295 tells us that after accounting for juries where they are over-represented African Americans will be under-represented in about ‘thirty percent’ of trials.

For a venire of 45, the SR of 0.4839 means that during the period when there was a computer problem, African Americans had slightly less than ‘one-half’ the chance of being called for jury service than Whites. The PD measure of 0.573 implies that defendants from the protected group will be at a disadvantage in nearly ‘sixty’ percent of the venires selected from a pool that is 4.17% African American instead of one that is 8.25% African American. The DR equals 0.44 and the point of maximum disparity occurs when the venire has two African-American members. The probability of observing two or fewer African Americans on a venire of 45 is only 0.2710, when they form 8.25% of the large jury pool but equals 0.7108 when they form 4.17%. The difference of 0.44 in these probabilities is the increase in the probability that a venire would contain ‘at least’ three African Americans. The decision in Hassan v. State of Texas,\(^{40}\) stated that only rarely could a \textit{prima facie} case of discrimination in peremptory challenges be established when ‘fewer than three’ members of the

distinctive group were struck. Thus, defendants would have statistical support for a Batson claim if
the prosecutor challenged all African Americans on their venire in an additional 44 out of 100 cases,
if the African-American fraction of the large jury wheel equalled their fraction of the age-eligible
community. Thus, the requirement of a DR of 0.50, even for a venire of 45 is far too stringent and
would deprive members of protected groups forming 5–10% of the community from a reasonable
opportunity to serve on a jury.

4. Accounting for the potential effect of under-representation of a protected group on
venires on a defendant’s chance of questioning unfairness in the peremptory challenges made
by the prosecution

After a venire is sent to a courtroom, the potential jurors are asked questions to assess whether they
have formed an opinion about the case, know one of the parties, lawyers or witnesses or have some
relationship or interest that indicates they might not be unbiased. These individuals are excused for
cause. Then the prosecutor and defendant can remove a set number, say P, of potential jurors by
peremptorily challenging them. No reason needs to be given; however, in 1986 the Batson decision
prohibited the lawyers from removing potential jurors because of race, gender or ethnicity. There are
several statistical procedures that have been proposed (DiPrima, 1995; Gastwirth, 2005; Barrett, 2007)
for the analysis of peremptory challenge data. In the present context, the easiest one for the courts to
apply is Fisher’s exact test, which is described in the texts of Gastwirth (1988, p. 217) and Finkelstein
and Levin (2001, p. 154) and applied to peremptory challenge data in DiPrima (1995) and Gastwirth
(2005). It is based on comparing the number of protected group members challenged by the prosecutor
(or defendant) to the number expected if the lawyer had taken a random sample of size P from the
venire. If the number of protected group members is statistically significantly higher than what would
likely occur in a random sample, the lawyer should provide a legitimate reason for removing the
protected group from the venire. When one can obtain peremptory challenge data from several similar
cases, the results of the individual Fisher exact tests can be combined to provide a more powerful test.
The procedure is illustrated in Gastwirth and Yu (2013) on data from nine death penalty cases in one
county in North Carolina. In this section we demonstrate how having one or two ‘fewer’ protected
group members on the venire can noticeably ‘decrease’ the power of the Fisher exact test to classify a
biased system of peremptory challenges as statistically significant.

4.1 The reduction in the probability of detecting an unfair system of peremptory challenges in
Kent County resulting from the under-representation of African Americans in the jury pool

The Code of Criminal Procedure (Act 175 of 1927 and updates) in Michigan allows different
numbers of peremptory challenges in accordance with the seriousness of the charge. For of-
fenses that may be punished by death or life imprisonment each side is allowed 12 peremptory
challenges (section 768.13 of the Code). For offenses with less severe punishment both sides are
allowed 5 peremptory challenges (section 768.12). The law also permits the judge to grant either
party or both extra peremptory challenges when they are justified. This sub-section describes the
reduced chance defendants in Kent County had when their share of venire members was 4.17%
instead of 8.25%.

Tables 7 and 8 report the conditional probability that Fisher’s test will find a significant difference, at
either the 0.05 level in both types of criminal cases, where each side has 5 or 12 challenges. The
\textit{Table 7} The probability that a venire of 45 has \( k \) African Americans and whether (1) or not (0) Fisher’s exact test would be statistically significant if the prosecution removes all or all but one of the \( k \) African American members from a venire of 45 in 5 peremptory challenges

<table>
<thead>
<tr>
<th>Proportion</th>
<th>( k )</th>
<th>prob.</th>
<th>cum. prob.</th>
<th>significant if remove all</th>
<th>( p )-value</th>
<th>significant if remove all but one</th>
<th>( p )-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0825</td>
<td>1</td>
<td>0.084</td>
<td>0.105</td>
<td>0</td>
<td>0.111</td>
<td>0</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.166</td>
<td>0.271</td>
<td>1</td>
<td>0.010</td>
<td>0</td>
<td>0.212</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.214</td>
<td>0.485</td>
<td>1</td>
<td>0.001</td>
<td>1</td>
<td>0.029</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.202</td>
<td>0.687</td>
<td>1</td>
<td>0.000</td>
<td>1</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.149</td>
<td>0.837</td>
<td>1</td>
<td>0.000</td>
<td>1</td>
<td>0.000</td>
</tr>
<tr>
<td>0.0417</td>
<td>1</td>
<td>0.288</td>
<td>0.435</td>
<td>0</td>
<td>0.111</td>
<td>0</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.276</td>
<td>0.711</td>
<td>1</td>
<td>0.010</td>
<td>0</td>
<td>0.212</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.172</td>
<td>0.883</td>
<td>1</td>
<td>0.001</td>
<td>1</td>
<td>0.029</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.079</td>
<td>0.961</td>
<td>1</td>
<td>0.000</td>
<td>1</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.028</td>
<td>0.989</td>
<td>1</td>
<td>0.000</td>
<td>1</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Note: The probability a venire of 45 has \( k \) or fewer African Americans is given in the column cum. prob. (cumulative probability).

\(p\)-values of Fisher’s exact test for the various situations are also given. Both tables assume a venire of 45 individuals sampled from a population with protected group proportions of 8.25% or 4.17%, respectively.\(^{41}\) When each side has five peremptory challenges and the African-American fraction of the jury-eligible population is 0.0825, the probabilities of having exactly \( k \) or less than or equal to \( k \) minorities on a venire of 45 are given in the columns prob and cumprob. Once there are at least ‘two’ African Americans from the protected group, the prosecution ‘cannot’ remove all of them without triggering a statistically significant result as the \( p\)-values of Fisher’s exact test are less than 0.05. This means that if the prosecutor’s office plans to remove all African Americans, defendants have the possibility of supporting a Batson claim in 89.5% of the cases.\(^{42}\) If the effective African American fraction of the population is 0.0417, defendants have the possibility of raising a Batson claim when the prosecutor removes all African Americans in only 56.5% of cases because the probability of observing zero or just one African American is 0.435.\(^{43}\) The difference of 0.333 in the probabilities of having sufficiently many African Americans on the venire to raise a statistically based Batson claim if the prosecutor challenges all of them is substantial. To account for situations where a prosecutor leaves one African American member on the venire to argue that they did not discriminate because they could have used another peremptory challenge to remove that individual, the corresponding probabilities

\(^{41}\) Recall that African Americans formed 8.25% of the age-eligible members of Kent County but the computer glitch caused their percentage of individuals serving on venires during the relevant time period to be only 4.17%.

\(^{42}\) This follows from the first and second rows of Table 7 because only when there is no more than one African American on the venire Fisher’s test will not be statistically significant when the prosecutor removes all of them. The probability of observing either 0 or 1 African American on a venire of 45 is given in the cumprob column and equals 0.105. The probability of having two or more African Americans on the jury is 1 – 0.105 = 0.895 or 89.5%. The \( p\)-value of Fisher’s test when there are 2 African American jurors is 0.01, i.e. the probability that a random sample of 5 drawn from a venire of 2 African Americans and 43 Whites would contain both African Americans is only 1 in a 100.

\(^{43}\) When the probability of observing no more than one African American on a venire is 0.435, the probability of observing at least two is 1 – 0.435 = 0.565.
Table 8: The probability that a venire of 45 has k African Americans and whether (1) or not (0) Fisher’s exact test would be statistically significant if the prosecution removes all or all but one of the k African American members from a venire of 45 in 12 peremptory challenges.

<table>
<thead>
<tr>
<th>Proportion</th>
<th>k</th>
<th>prob.</th>
<th>cum. prob.</th>
<th>significant if remove all</th>
<th>p-value</th>
<th>significant if remove all but one</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0825</td>
<td>1</td>
<td>0.084</td>
<td>0.105</td>
<td>0</td>
<td>0.267</td>
<td>0</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.166</td>
<td>0.271</td>
<td>0</td>
<td>0.067</td>
<td>0</td>
<td>0.467</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.214</td>
<td>0.485</td>
<td>1</td>
<td>0.016</td>
<td>0</td>
<td>0.169</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.202</td>
<td>0.687</td>
<td>1</td>
<td>0.003</td>
<td>0</td>
<td>0.052</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.149</td>
<td>0.837</td>
<td>1</td>
<td>0.001</td>
<td>1</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>0.089</td>
<td>0.926</td>
<td>1</td>
<td>0.000</td>
<td>1</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>0.045</td>
<td>0.971</td>
<td>1</td>
<td>0.000</td>
<td>1</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>0.019</td>
<td>0.990</td>
<td>1</td>
<td>0.000</td>
<td>1</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>0.007</td>
<td>0.997</td>
<td>1</td>
<td>0.000</td>
<td>1</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.002</td>
<td>0.999</td>
<td>1</td>
<td>0.000</td>
<td>1</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>0.001</td>
<td>1.000</td>
<td>1</td>
<td>0.000</td>
<td>1</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>0.000</td>
<td>1.000</td>
<td>1</td>
<td>0.000</td>
<td>1</td>
<td>0.000</td>
</tr>
<tr>
<td>0.0417</td>
<td>1</td>
<td>0.288</td>
<td>0.435</td>
<td>0</td>
<td>0.267</td>
<td>0</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.276</td>
<td>0.711</td>
<td>0</td>
<td>0.067</td>
<td>0</td>
<td>0.467</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.172</td>
<td>0.883</td>
<td>1</td>
<td>0.016</td>
<td>0</td>
<td>0.169</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.079</td>
<td>0.961</td>
<td>1</td>
<td>0.003</td>
<td>0</td>
<td>0.052</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.028</td>
<td>0.989</td>
<td>1</td>
<td>0.001</td>
<td>1</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>0.008</td>
<td>0.998</td>
<td>1</td>
<td>0.000</td>
<td>1</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>0.002</td>
<td>1.000</td>
<td>1</td>
<td>0.000</td>
<td>1</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>0.000</td>
<td>1.000</td>
<td>1</td>
<td>0.000</td>
<td>1</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>0.000</td>
<td>1.000</td>
<td>1</td>
<td>0.000</td>
<td>1</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.084</td>
<td>0.105</td>
<td>1</td>
<td>0.000</td>
<td>1</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>0.166</td>
<td>0.271</td>
<td>1</td>
<td>0.000</td>
<td>1</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>0.214</td>
<td>0.485</td>
<td>1</td>
<td>0.000</td>
<td>1</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Note: For each possible value, k, of African Americans on a venire of 45, the probability a venire has k or fewer African American members is given in the column cum. prob. (cumulative probability).

that the statistical test will be significant are 72.9%, when venire members are selected from a population in which African Americans form 8.25% but only 28.9% when African Americans form only 4.17%. Notice that the difference of 44% in the probabilities a defendant could have statistical support for a Batson claim, is ‘larger’ than in the case where the prosecutor plans to remove all African Americans.

Next, consider criminal cases in which both sides are allowed 12 peremptory challenges. If there is only one African American member on a venire of 45, then the first row of Table 8 shows there is a probability of 0.267 that they would be included in a random sample of 12 from the venire. Thus, Fisher’s test would ‘not’ be near significance and would not support a Batson challenge. When there are 2 African American members on the venire, the probability that both would be included in a random sample of 12 from a pool of 43 Whites and 2 African Americans is 0.067; while small this probability exceeds 0.05, the value typically required for statistical significance. Thus, for serious offenses even when there are two African Americans on a venire
of 45 and both are challenged, a defendant ‘cannot’ provide ‘strong’ statistical support for a Batson claim.

When there are three African Americans on the venire, the third line of Table 8 gives the probability that all three would be included in a random sample of 12 as 0.016, which is a statistically significant result. Thus, when there are three African Americans on the venire and the prosecutor challenges all three, the defendant has statistically significant evidence showing that the prosecutor’s challenges appear to be related to race. Recall that in a venire of 45 randomly drawn from a large population of whom 8.25% are African American, the probability is 0.729 that the venire would include at least three African American members. However, if the venire is randomly selected from a pool where African Americans only form 4.17% of the pool, the probability is only 0.289 that there will be three or more African Americans on the venire. Thus, the probability that the defendant has an opportunity to provide statistical support for a Batson claim declines by 0.44, which implies that in an ‘additional’ 44 out of 100 serious criminal cases, prosecutors could peremptorily challenge ‘all’ the African American members on the venire without creating a statistically significant result.

In the situation where a prosecutor plans to remove all but one African American, Table 8 shows that the venire needs to have at least 5 African Americans if courts strictly require statistical significance at the 0.05 level or 4, if they accept a \( p \)-value of 0.052 as statistically significant. The probability of having at least 5 (4) African Americans on a venire of 45 are 0.3126 (0.5148) when African Americans form 8.25% of the population called for jury service but only 0.0106 (0.1172) when they form 4.17%. Again the differences of 30.2% and 39.76% in the proportion of venires a defendant would have statistical support for a Batson claim if a prosecutor removed all but one African American member of the venire are substantial.

4.2 The effect of ‘statistically undetectable’ biased peremptory challenges on representation of the protected group on juries

Following Re (2011) the Bryant opinion considered the DR measure for a jury of 12, by comparing the composition of a jury chosen from the share of the protected group, \( p \), of the jury pool with the corresponding share, \( \pi \), of the community. The results in Table 8 show that there needs to be a minimum number of African Americans on the venire in order for Fisher’s exact test to detect a prosecutor’s unfairly removing them. Thus, it is worth calculating the expected African American proportion of actual jurors, or equivalently their expected proportion of the venire remaining after the peremptory challenges are made when the prosecutor removes as many African Americans as possible without triggering a statistically significant result.

For the data in Table 8 Fisher’s test will ‘not’ be significant if there are no more than two African Americans on the venire and they are peremptorily challenged. This occurs with probability 0.7108 (0.2710) when African Americans form 4.17% (8.25%) of the source pool. When the venire of 45 has three African Americans, the prosecutor can remove two without Fisher’s test reaching significance. When the venire has four African Americans, the prosecutor can remove two without Fisher’s test being significant at the 0.05 level; however, if three are removed the test

\[\text{This is obtained from Table 8 by subtracting the cumulative probability (0.271) of obtaining 2 or fewer African Americans on a venire of 45 from 1.}\]

\[\text{Because the final jury can be regarded as a random sample of 12 from the venire, after individuals have been removed for cause or by peremptory challenges, the protected group proportion of the remaining venire is their expected proportion on the jury.}\]
almost reaches significance.\textsuperscript{46} Similarly, when the venire has five or six (seven, eight or nine) African Americans, the prosecutor can safely remove three (four) etc. To calculate the protected group fraction of the venire after the challenges, one needs to know how many were removed during the process. The African American proportion of the remaining venire is largest, if one assumes that both sides use all 12 of their allowed challenges and the defendant only removes Whites.\textsuperscript{47} By using all 12 challenges, a prosecutor decreases the probability that Fisher’s test will find the number of African Americans they challenged statistically significant.\textsuperscript{48} Thus, there will be 21 venire members remaining. If the original venire of 45 had 3(4) African Americans, after the prosecutor removes 2, the African American fraction of the remaining venire is 1/21 (2/21).\textsuperscript{49} Taking the weighted average of these fractions, where the weights are the probabilities that the original venire had the assumed number of African Americans (see the lower part of Table 7) yields a protected group proportion of the remaining venire from which the actual jurors are chosen of 0.02 (if the judge includes the situation when the prosecutor removes three of four African Americans out of a panel of 45 as significant) or 0.016, if the judge does not.\textsuperscript{50}

Thus, a system of peremptory challenges that would ‘not’ be classified as statistically significant could reduce the African American share of the pool from which the actual jurors are chosen from a group forming 4.17\% of venire members to just 2\%. Consequently, the expected African American proportion of actual jurors would only be ‘one-fourth’ of their proportion of the eligible population. On the other hand, had African American proportion of the venires in Kent County equalled their 8.25\% of the eligible population, a similar calculation shows that if the prosecutor removed as many African Americans as possible without triggering a statistically significant Fisher exact test, the African American share of actual jurors would be 6.89\% if the judge included the borderline Fisher p-value of 0.052, when three of four African American venire members were challenged or 5.9\%, if the judge rigidly adhered to the 0.05 cut-off value. Notice that the prosecutor’s peremptory challenges could reduce African American share of actual jurors to between 71.5\% and 83.5\% of their proportion in the community when they are fairly represented on the venire, substantially greater than 25\% of actual jurors when their proportion of the venire was already reduced to a little over one-half their fraction of the eligible population. These calculations demonstrate that when determining a ‘reasonable or allowable’ shortfall in African American representation on venires, courts should also consider the further effect the peremptory challenge process is likely to have on the demographic mix of the actual jury.

\textsuperscript{46} When three of the four are removed, the p-value of Fisher’s test is 0.052. While technically, this probability is larger than 0.05, given the relatively small sample available it would be reasonable for a judge to conclude that this event is close enough to formal significance that it would help the defendant establish a prima facie case.
\textsuperscript{47} While the prosecution will likely remove some Whites, if only to counter a Batson challenge, it might not use all of their allowed 12, which would lead to a higher fraction of Whites on the remaining venire.
\textsuperscript{48} To see this consider the situation where there are 5 minorities and 40 Whites on the venire. If the prosecution removes three minorities and nine Whites, i.e. 60\% of the minorities and 22.5\% of the Whites, Fisher’s test has a p-value of 0.109, which exceeds 0.05; the commonly used cut-off for statistical significance. Thus, the prosecutor can remove three of the five minorities without triggering a statistically significant result. Had the prosecutor challenged three minorities but only three Whites (7.5\%), Fisher’s test would be statistically significant (p-value = 0.028).
\textsuperscript{49} Similarly, when the venire originally had 5 or 6 African Americans, after three were removed, the African American fraction of those remaining are 2/21 and 3/21. When the original venire had 7 or 8 and the prosecutor removes four, African Americans form 3 of 21 and 4 of 21, respectively. Taking the weighted average of these fractions, where the weights are the probabilities that the original venire had the assumed number of African Americans (see the lower section of Table 7) yields a protected proportion of actual jurors of 0.02.
\textsuperscript{50} See n. 46, supra, and the surrounding text for the mathematical details.
To appreciate the combined effect of the Michigan Supreme Court’s requirement that the DR is at least 0.50 and the prosecutor’s removing as many African Americans of the venire as possible without triggering a statistically significant Fisher’s exact test, recall that the DR must exceed 0.50 criteria would have accepted an African American proportion on venires of 1.5%, i.e. if African Americans composing 8.25% of the eligible community only formed 1.5% of a large sample of venires, The Michigan Court would ‘not’ find they were under-represented.51 A calculation of the effect of the peremptory challenge process, similar to the ones presented here, shows that African Americans could almost always be eliminated from those who serve on actual juries as only three percent of the venires remaining after the challenge process would include a sufficient number of them.52

5. Summary and discussion

The analysis demonstrates that the DR measure is not a measure of how the defendant’s chances of a representative jury are affected by smaller fraction of the protected group actually summoned for jury service than their proportion in the eligible community. Rather, it is a measure of the statistical distance between the corresponding binomial distributions. On the other hand, the PL measure tells us the probability that a jury or venire chosen from the population with the smaller protected group fraction will have ‘fewer’ protected group members than one chosen from the community. This probability tells us the fraction of all juries in the jurisdiction that will have fewer protected group members than they would have if the jury wheel truly represented the community. The PD measure adjusts the PL measure to account for the small proportion of juries where the protected group is over-represented even though they are under-represented in the jury pool. In conjunction with the SR, which compares the probability a protected group member has of being called for jury service to that of others and statistical hypothesis testing adopted in Castaneda, courts have statistically sounder procedures for the examination of jury under-representation of the protected group than the DR.

Another implication of this article is that the Michigan Supreme Court should have inquired about the statistical properties of DR measure before accepting it. As seen in the discussion of data from Duren as well as our examination of the difference between two normal distributions, the suggested criteria requiring a DR of 0.37 or 0.50 are far too stringent and would allow jurisdictions to severely restrict minority or female participation on juries. The results in Table 8 show that when the Bryant court used the DR >0.50 criteria and accepted an African American share of 4.17% of the jury pool as consistent with their 8.25% share of the eligible community, only 29% of the venires would have a sufficient number of African Americans to question a prosecutor who peremptorily challenged all of them. Thus, in 71% of trials during the period prosecutors could remove ‘all’ African Americans from a venire, without triggering a statistically significant disparity.

51 Although the DR measure of 0.478 is less than 0.50, the selection ratio, however, would equal 0.169, implying that an eligible protected group member would have less than one-fifth the probability of being on a venire as others. The PL measure is 0.578, implying that in nearly 60% of venires; there would be fewer protected group members than there would be if the venires were drawn from the eligible community. All three measures do not include the additional reduction in representation of the protected group that can occur when the peremptory challenges are made.

52 The probability that a venire or binomial random variable with $n = 45$ and $p = 0.015$ has two or fewer African Americans on it is 0.97. Thus, only three percent of venires will include a sufficient number of African Americans to detect a significant disparity when the prosecutor removes all of them. Consequently, the expected African American percentage of a venire remaining after the prosecution’s challenges is only 0.146%, or about one-fiftieth of their percentage, 8.25%, of the community.
Most of the legal decisions concerning fair representation focus on the demographic mix of the large jury pool from which venires are selected. Recently, in *U.S. v. Hernandez-Estrada*, the Ninth Circuit noted that ‘if a statistical analysis shows under-representation, but the under-representation does not substantially affect the representation of the group in the actual jury pool, then the under-representation does not have legal significance in the fair cross-section context’. The results presented here show that when courts assess the legal significance of a statistically significant under-representation and meaningful as indicated by the values of SR and PD measures, of a protected group on the venires during the relevant period, they should consider the additional decrease on actual juries due to the increased probability that prosecutors could remove all or nearly all protected group members from the venire but not a statistically significant number of them.

In the Kent County cases, there was only a difference of two in the expected number of African Americans on a venire of 45 between jury pools with African American shares of 8.25% and 4.17%. While a difference of two on a venire might appear small, it led to a substantial increase in the proportion of venires that would allow a prosecutor to remove all African Americans from them without triggering a statistically significant finding that African Americans were disproportionately struck. By incorporating the effects of the under-representation of African Americans in the large jury pool and the potential decrease in African American membership of the venire remaining after peremptory challenges the methodology described here addresses the reservations the opinion in *United States v. Biaggi* expressed about the absolute impact criteria, i.e. ‘The risk of using this approach is that it may too readily tolerate a selection system in which the seemingly innocuous absence of small numbers of a protected group from an average array creates an unacceptable probability that the protected group members of the jury ultimately selected will be markedly deficient in number and sometimes totally missing.’

The reanalysis of the jury composition data in Kent County presented in Section 3 and the related analysis reported in Gastwirth and Pan (2011) show that the statistical evidence presented by the experts for the affected defendants concerning the ‘computer glitch’ led to a systematic and significant decrease in representation of African Americans on venires and juries from April 2001 thru August 2002. Thus, all the statistical analyses of the relevant data are consistent with the conclusion of the 6th Circuit in *Ambrose* that the statistical evidence helped establish a *prima facie* case of under-representation of African Americans during the relevant period.

---

53 Supra n. 20 at 1165, citing *United States v. Kleifgen*, 557 F.2d 1293, 1297 (9th Cir. 1977). This opinion noted that in an array of 100 jurors the under-representation would lead to 2.9 fewer African Americans and 4.4 fewer males and concluded that this was not legally meaningful. This degree of under-representation of males is very similar to the one in *Bryant*, which we found led to a noticeable reduction in the probability of detecting unfairness in peremptory challenges.

54 Beale (1983) cites a number of cases where courts have said that an absolute impact of one or two minorities on a grand jury or a jury venire is not substantial. See, for example, *U.S. v. Goff*, 509 F. 2d 825, 826 ((5th Cir. 1975) (under-representation of 1.4 members of a grand jury of 23 not substantial), *U.S. v. Potter*, 552 F.2d 901, 906 (9th Cir, 1977) (under-representation of less than one African American on a grand jury of 23 or one person in a jury array of 50 not substantial) or *United States v. Test*, 550 F. 2d 577 (10th Cir. 1975) (two persons on a jury panel is not a ‘gross disparity’). The 10th Circuit’s opinions in Test and *U.S. v. Yazzie*, 660 F. 2d 422 (10th Cir. 1981) also evaluate the absolute disparity on a petit jury of 12 by multiplying 12 by the absolute disparity. Thus, a two person shortfall on a venire of 50 becomes a shortfall of about 0.5 or one-half a person on the petit jury. This procedure ignores the potential impact of peremptory challenges on the ultimate jury.

55 909 F. 2d 662, 678 (2nd Cir. 1990). The court observed that the impacts of failing to include two more African Americans or two to three more Hispanics on a venire of 60 stretched the absolute impact approach it had accepted in *Jenkins*, 496 F.2d 57 (2d Cir. 1975) to the limit and would be of greater concern had the city not randomly selected potential jurors from the voter registration list.
REFERENCES


