Group: Paine’s Protectors (SF RAI - Civil Rights Groups)

Members: Anushka Mehta, Ben Farrah, Pablo Zavala Reina, Pallavi Rajan, Youyou Huang

**Judgment by Algorithm: Exploring AI Fairness in Criminal Justice**

**Synthetic Data Analysis**

**Summary**

For this project, we focus on evaluating the use of AI in criminal justice, with an emphasis on analyzing AI and human bias and fairness across three racial groups (predominantly black, predominantly white, and racially balanced which is annotated as others) and three counties (Claiborne, Copiah, and Warren).

1. **Demographic Distribution**

Figure 1.1.1. Sample Racial Composition by County

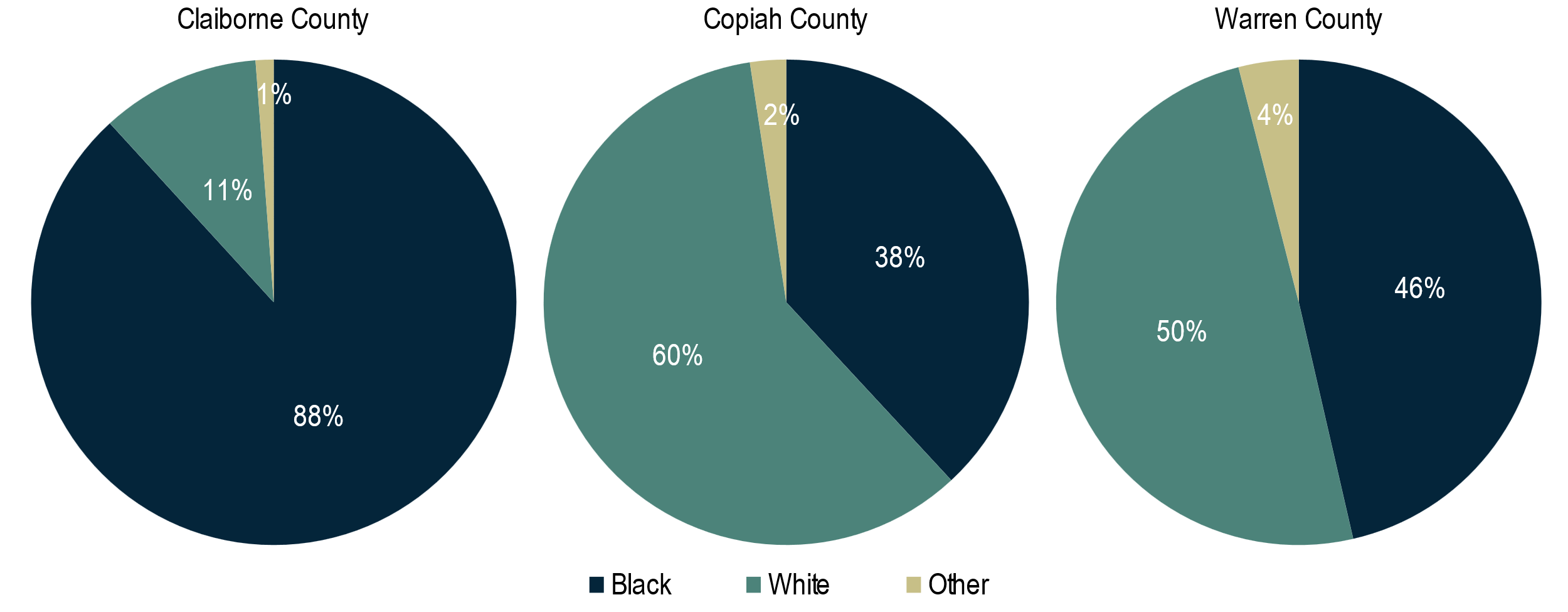


Figure 1.1.2. Sample Population by Race and County

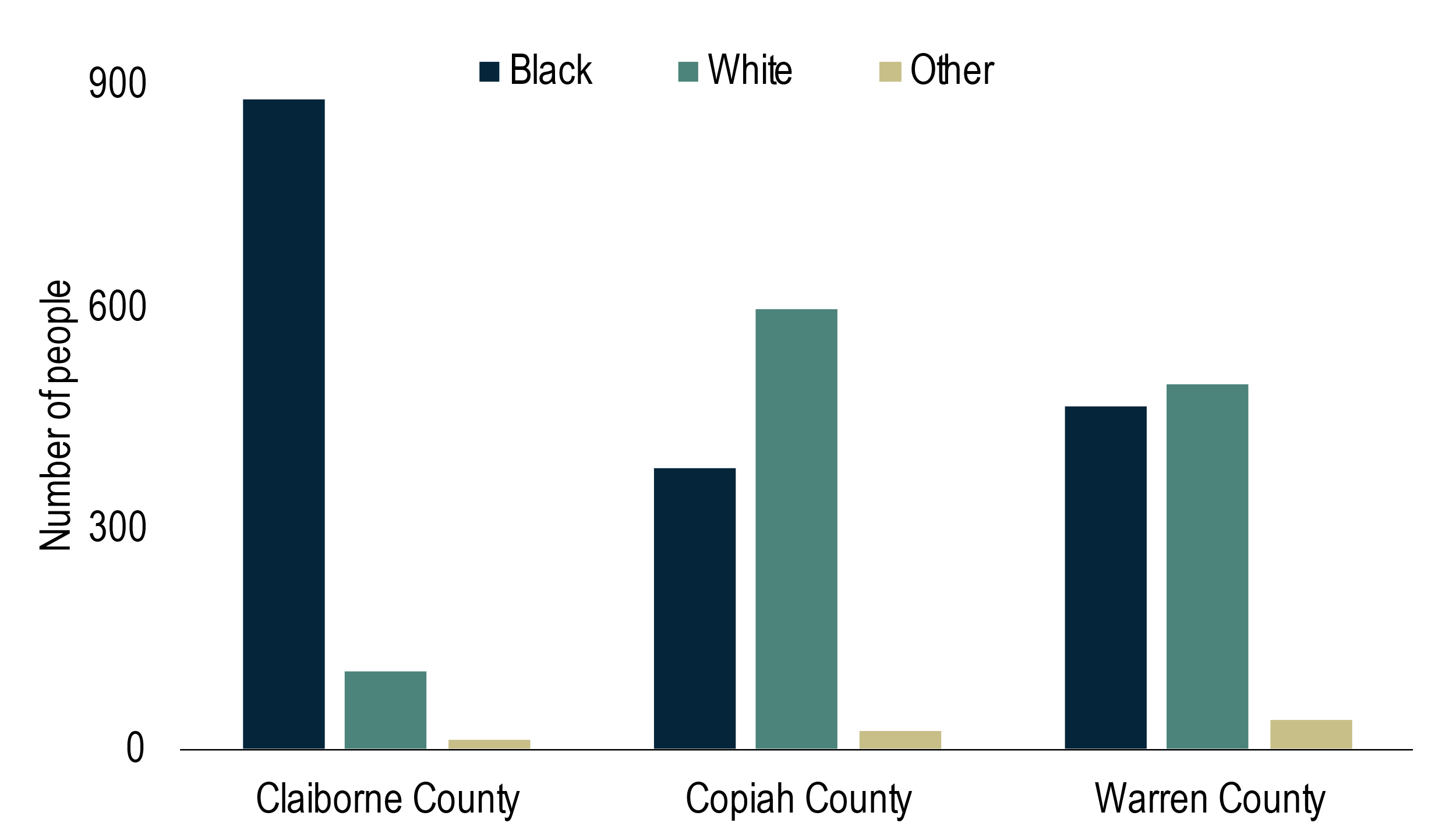


Table 1.1.1. Sample Population by Race and County

| **County** | **Black** | **White** | **Other** | **Total** |
| --- | --- | --- | --- | --- |
| Claiborne County | 882 | 106 | 12 | 1000 |
| Copiah County | 381 | 595 | 24 | 1000 |
| Warren County | 464 | 496 | 40 | 1000 |
| **Total** | 1727 | 1197 | 76 | 3000 |

The total sample population of the datasets across 3 counties is balanced, each containing 1000 observations (Table 1.1.1). There are noticeable disparities in the racial composition, both in the overall population and across the 3 counties. Black defendants make up the majority of the sample population (57.6%), followed by White defendants (39.9%%) and others (2.5%). Claiborne county has a significantly larger black population, representing about 88% of the county total (Figure 1.1.1) and over half of the total black population for all 3 counties (Table 1.1.1, Figure 1.1.2). Due to the disparate racial distribution, the conclusions drawn from these datasets may be vulnerable to statistical bias, especially for Others whose data is very scarce.

Figure 1.2.1. Sample Gender Composition by County

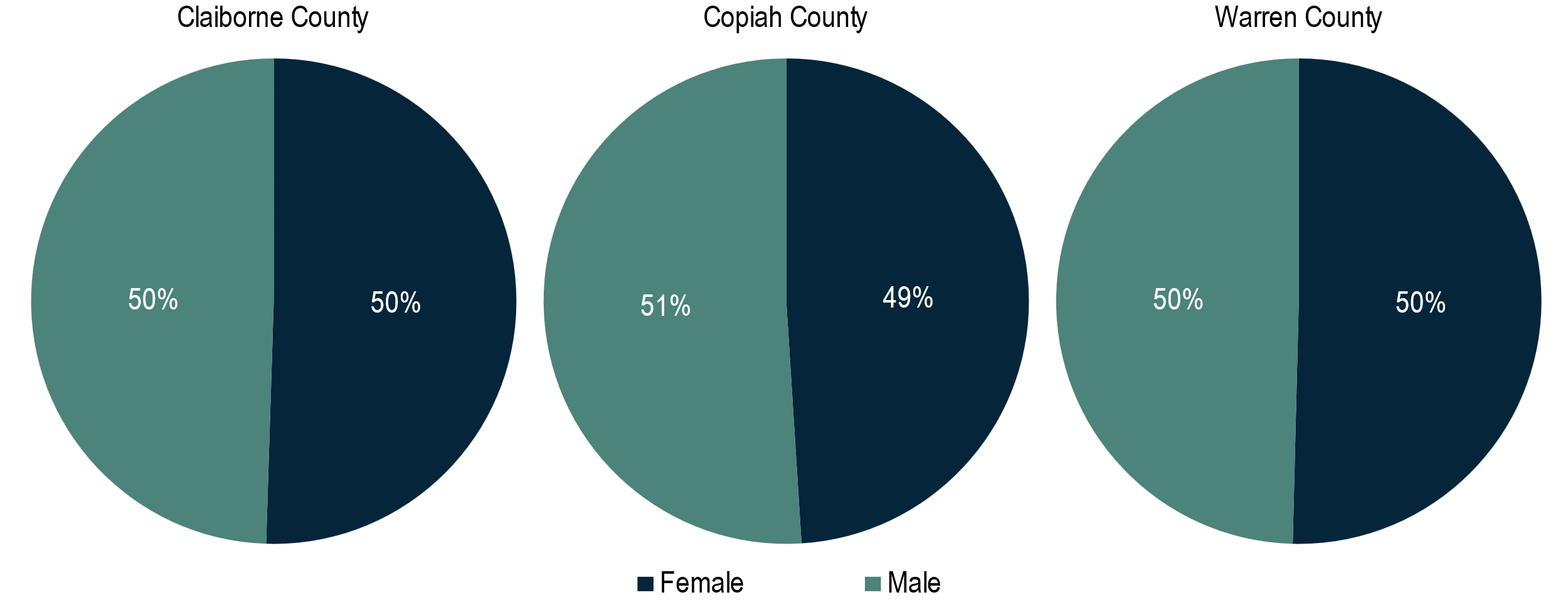


Figure 1.2.2. Sample Population by Gender and County

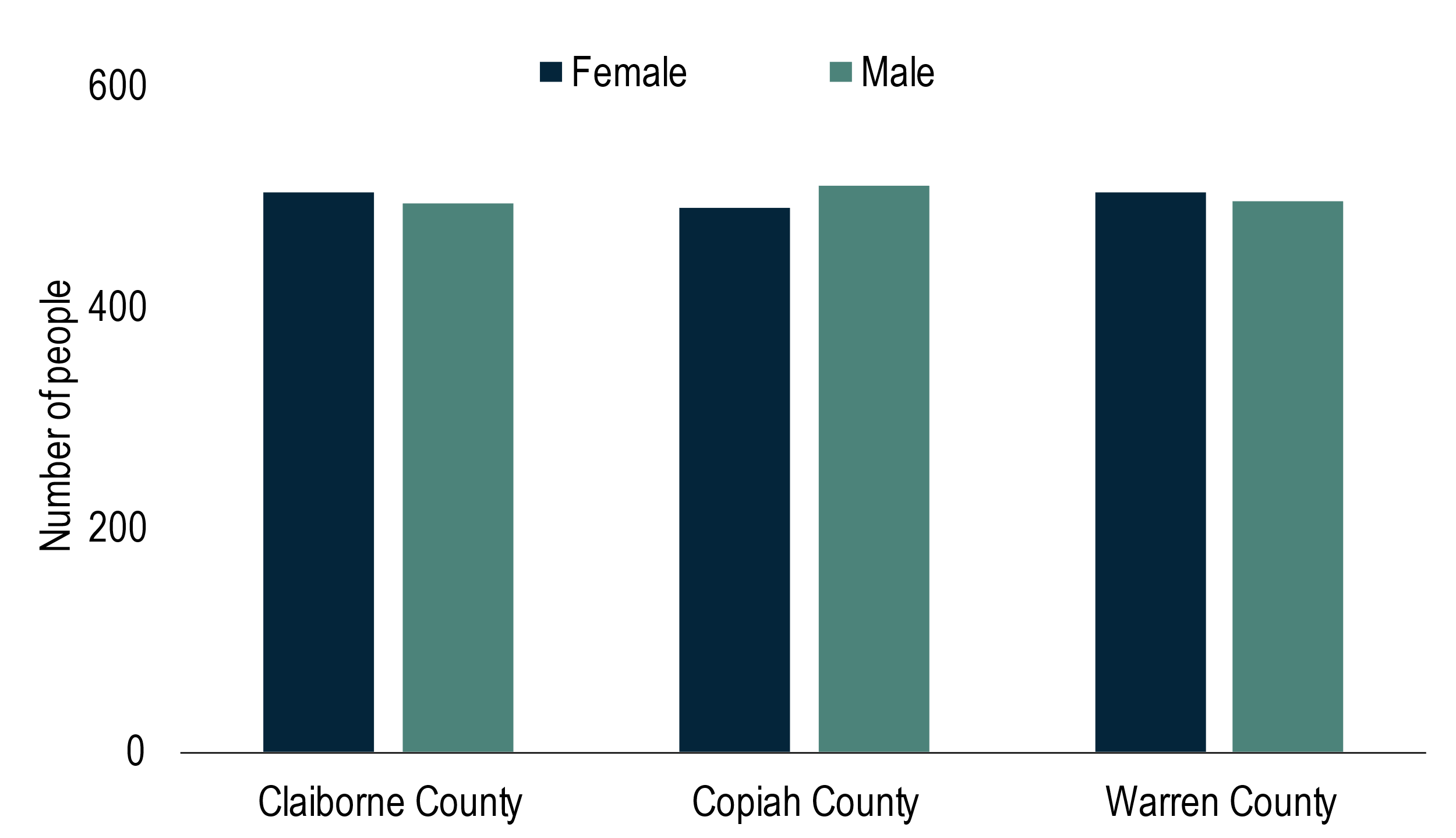


Figure 1.3.1. Sample Education Level Composition by County

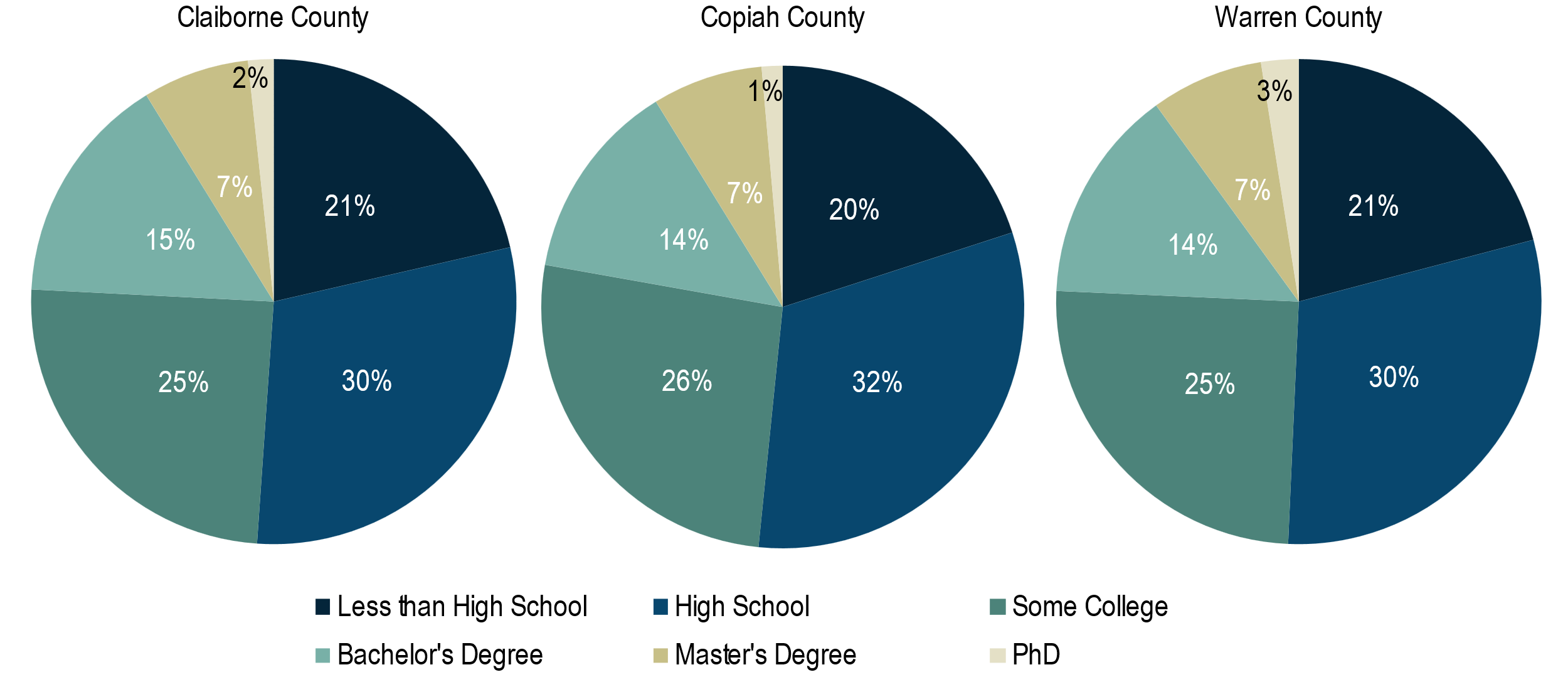
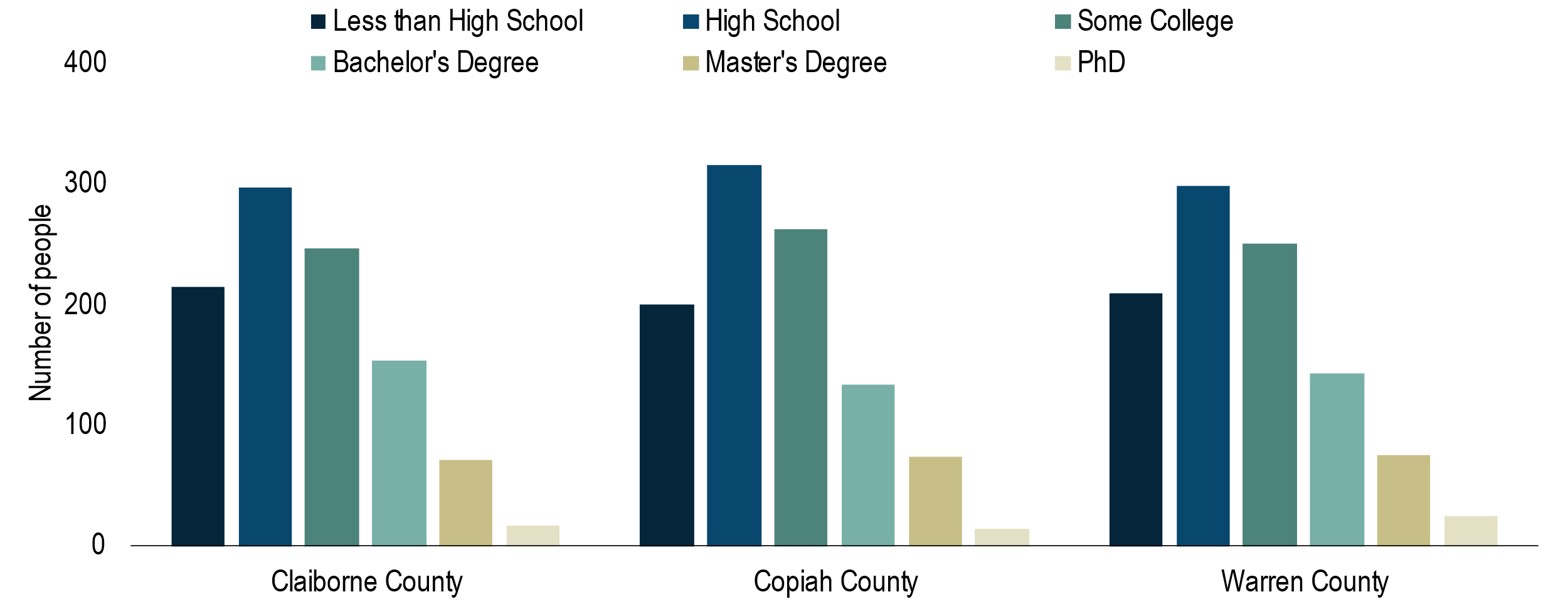


Figure 1.3.2. Sample Population by Education Level and County



Female and Male populations are almost equally distributed both in the overall population and across the 3 counties. As for education, individuals whose highest level of education is High School has the highest number, followed by Some College and Less than High School. Individuals with more advanced degrees only consist of less than 25% of the population. This trend is consistent across all 3 counties.

1. **Risk scores among different demographic groups**

Figure 2.1. Average Risk Score by Race and County

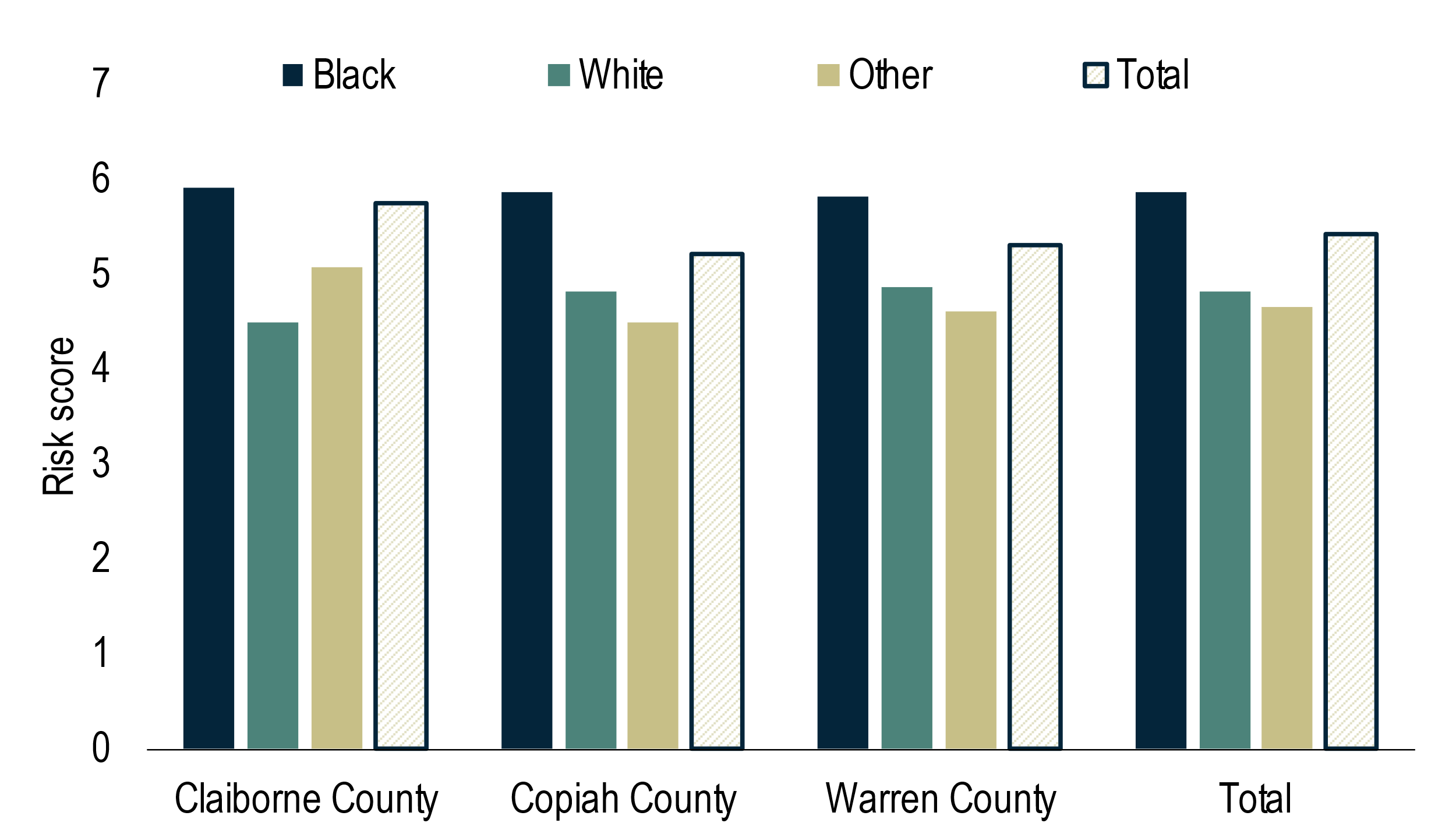
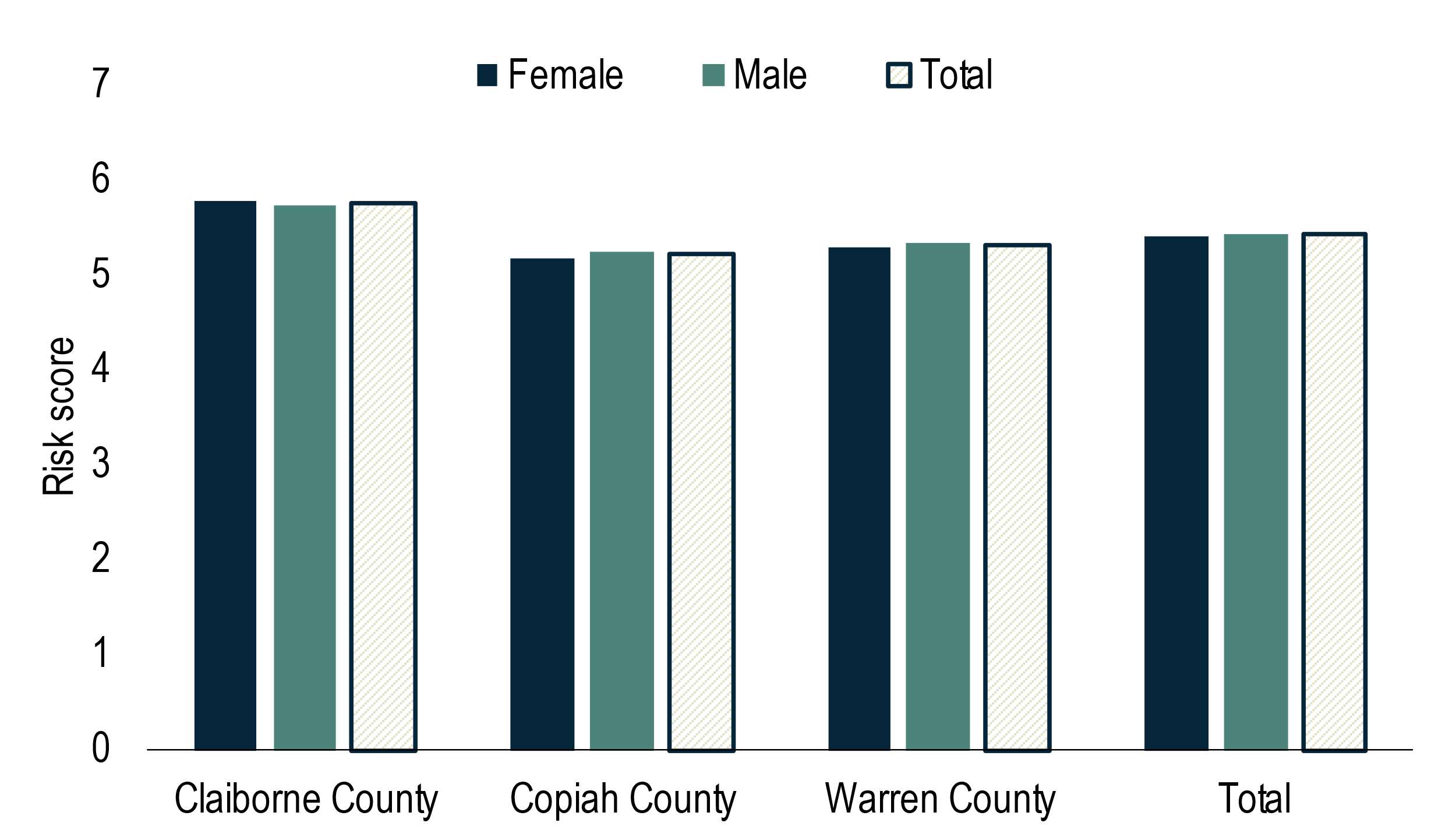


Table 2.1. Difference with Average Risk Score of Total Population

| **County** | **Black** | **White** | **Other** |
| --- | --- | --- | --- |
| Claiborne County | +0.16 | -1.25 | -0.67 |
| Copiah County | +0.65 | -0.39 | -0.71 |
| Warren County | +0.51 | -0.42 | -0.68 |
| **Total** | +0.45 | -0.60 | -0.77 |

Figure 2.2. Average Risk Score by Gender and County



Overall, the average risk score derived by AI is higher for individuals in Claiborne county than those in the other 2 counties. Among different racial groups, risk scores are on average higher for Black individuals than for White and Other individuals by at least 1. This trend is consistent across all 3 counties. In Claiborne county with the highest density of Black individuals, the racial disparity is more significant, where the average risk score for Black is the highest and that for White is the lowest among the 3 counties. Individuals categorized as Other have the lowest risk score on average, but have slightly higher average score than White in Claiborne county. However, due to the small sample size of Other, nuanced differences may not be significant and cannot generalize the larger racial group. In contrast, the average risk scores are well balanced and present little differences among Female and Male groups.

1. **Risk score and judge’s decision on bails among different demographic groups**

Figure 3.1.1. Risk Score Distribution by Judge’s Bail Decision and Race, Size of Population

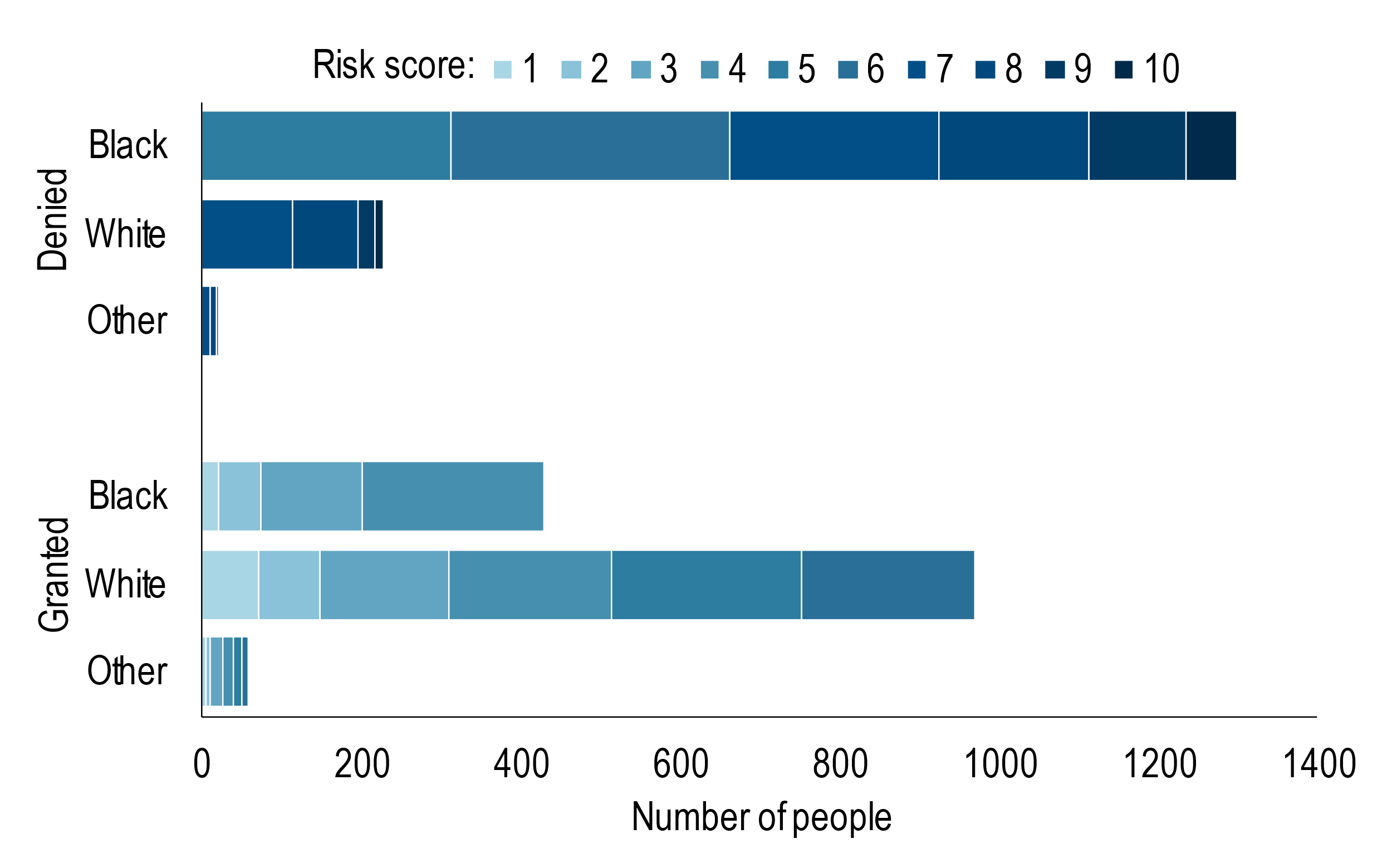


Figure 3.1.2. Risk Score Distribution by Judge’s Bail Decision and Race, Share of Population

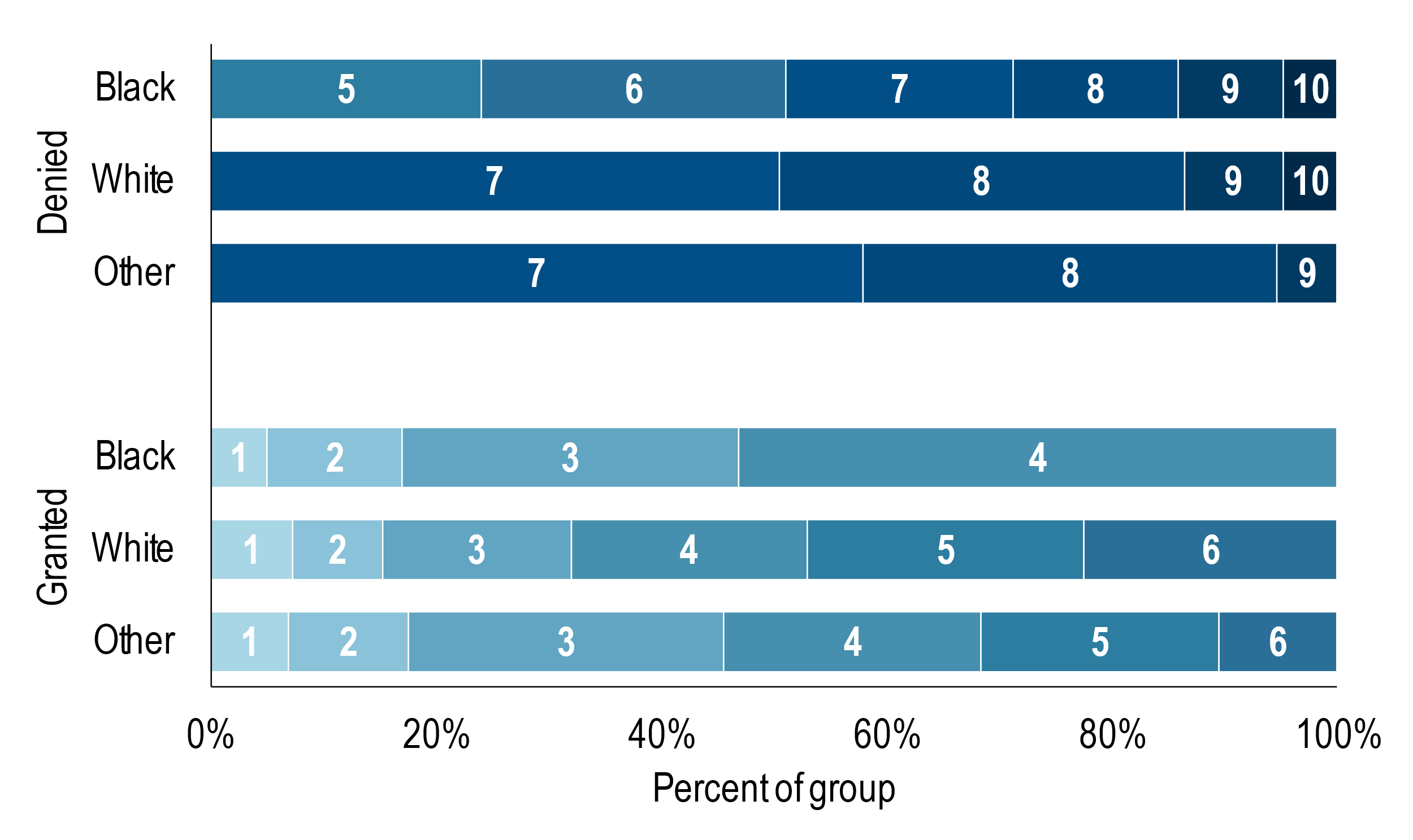


Figure 3.2.1. Risk Score Distribution by Judge’s Bail Decision and Gender, Size of Population

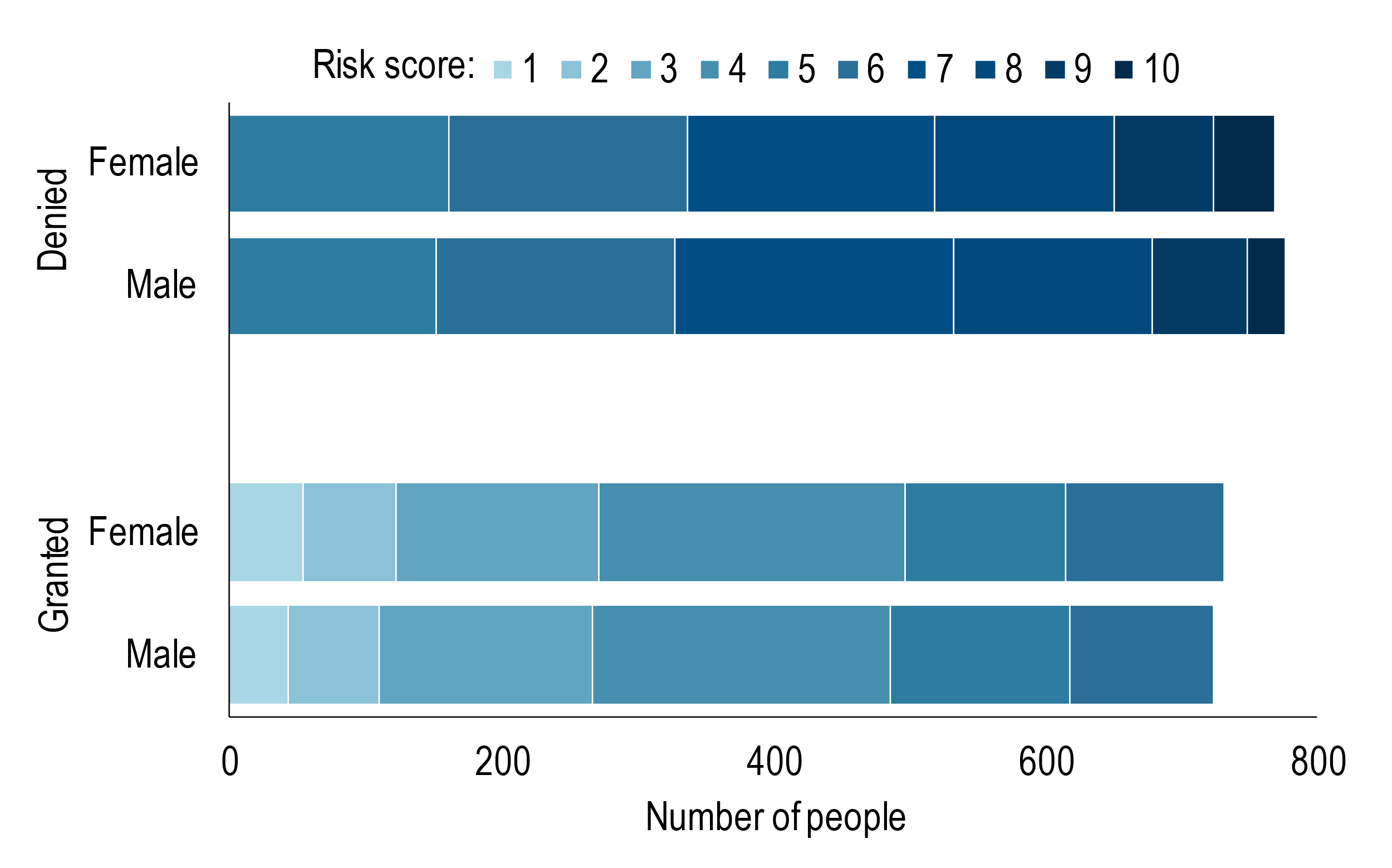
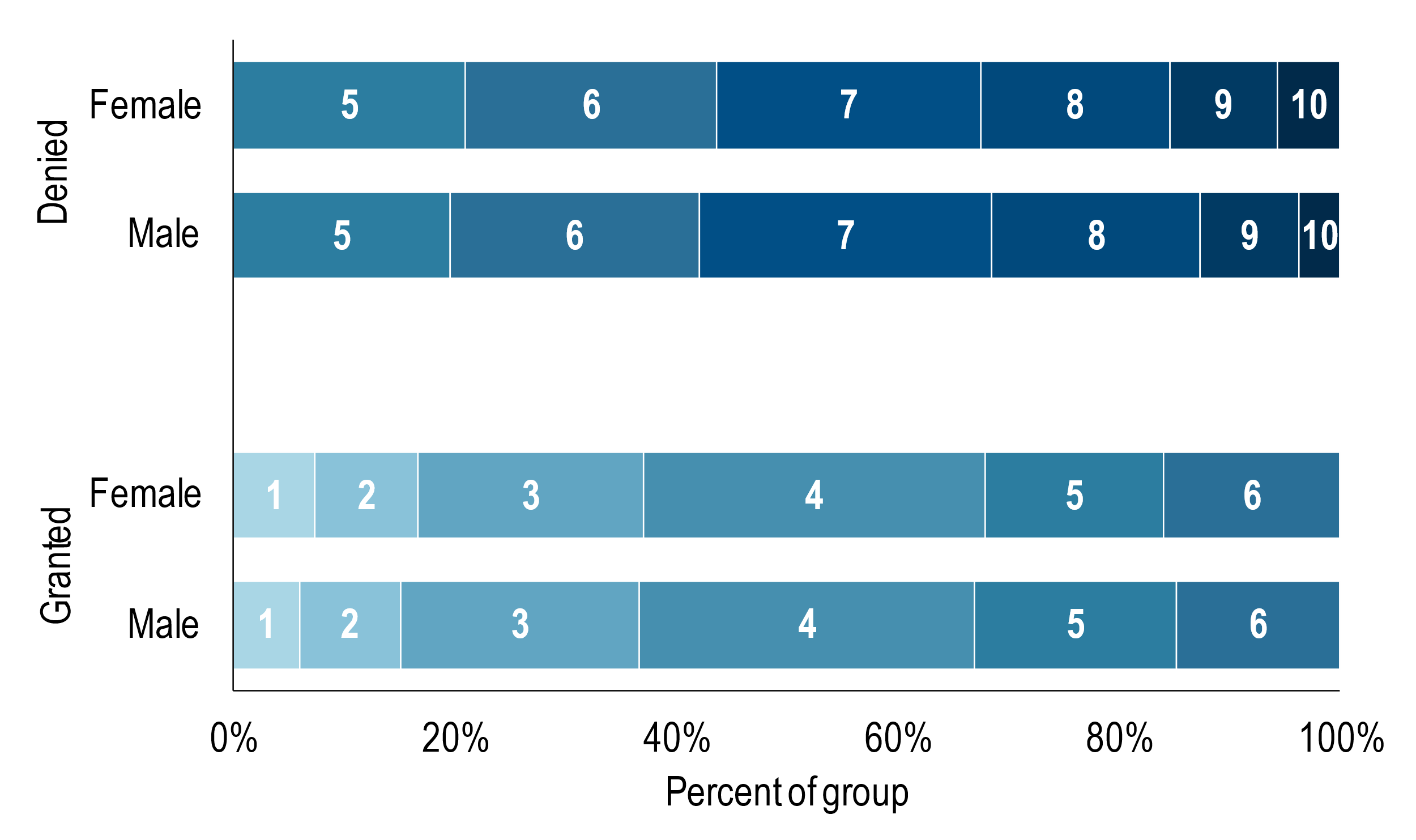


Figure 3.2.2. Risk Score Distribution by Judge’s Bail Decision and Gender, Share of Population



Among the group who were denied bail, there were far more Black individuals than White or Other, whereas most individuals granted bail were White. Black individuals given a risk score of 5 or higher were more likely to be denied bail by the judge. White or Other individuals were much more likely to be granted bail than Black individuals, however, since people who were White or Other and were granted bail had risk scores of up to 6. When comparing gender groups who were denied or granted bail, however, no major differences were evident. The proportion of risk-score groups were generally similar across males and females who were denied or granted bail.

1. **Re-offense rate and fairness metrics among different racial groups**

Figure 4.1. Re-offense, False Positive, and False Negative Rates by Race

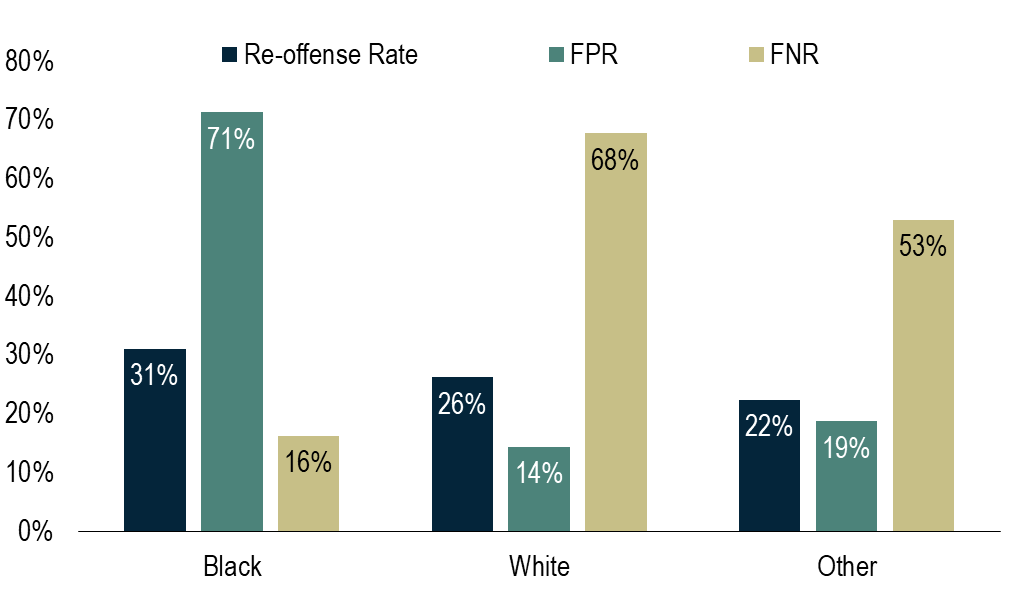


Table 4.1. Fairness Metrics

| **Race** | **Judge Decision** | **Re-offended** | **Did Not Re-offend** |
| --- | --- | --- | --- |
| Black | Denied | 449 | **849** |
| Granted | 87 | 342 |
|  |  |  |  |
| White | Denied | 101 | **127** |
| Granted | 212 | 757 |
|  |  |  |  |
| Other | Denied | 8 | **11** |
| Granted | 9 | 48 |

The reoffense rate was not found to be very different across racial groups, but the false positive rates (FPR) and false negative rates (FNR) do differ significantly. The FPR for White individuals was over 4 times higher than Black individuals. The FNR for Black individuals was over 5 times higher than White individuals.

**Appendix.**

Figure 3.1.1.alt. Risk Score Distribution by Re-offense and Race, Size of Population

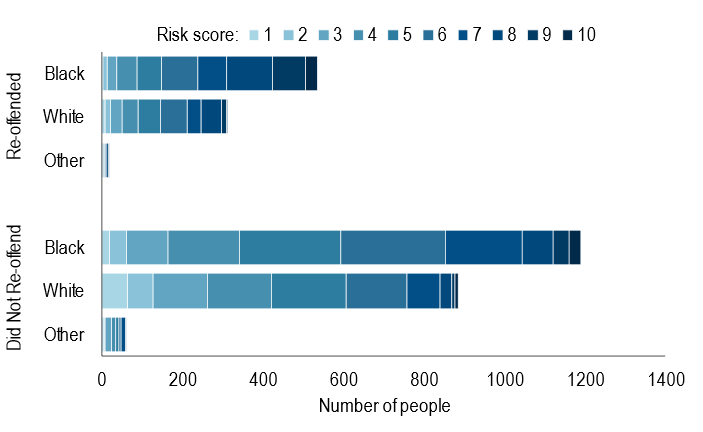


Figure 3.1.2.alt. Risk Score Distribution by Re-offense and Race, Share of Population

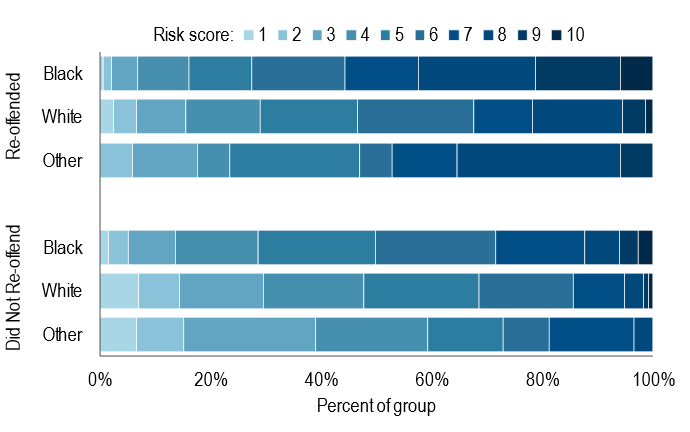


Table 4.1.alt. Fairness Metrics by Risk Score Thresholds by Race

| **Rate** | **Race** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **FPR** | **Black** | 100% | 98% | 95% | 86% | **71%** | 50% | 28% | 12% | 6% | 3% |
| **White** | 100% | 93% | 86% | 70% | 52% | 31% | **14%** | 5% | 2% | 1% |
| **Other** | 100% | 93% | 85% | 61% | 41% | 27% | **19%** | 3% | 0% | 0% |
|  |  |  |  |  |  |  |  |  |  |  |  |
| **FNR** | **Black** | 0% | 1% | 2% | 7% | **16%** | 28% | 44% | 58% | 79% | 94% |
| **White** | 0% | 3% | 7% | 16% | 29% | 47% | **68%** | 78% | 95% | 99% |
| **Other** | 0% | 0% | 6% | 18% | 24% | 47% | **53%** | 65% | 94% | 100% |
|  |  |  |  |  |  |  |  |  |  |  |  |
| **ACC** | **Black** | 31% | 32% | 34% | 38% | **46%** | 57% | 67% | 74% | 71% | 69% |
| **White** | 26% | 31% | 35% | 44% | 54% | 65% | **72%** | 76% | 74% | 74% |
| **Other** | 22% | 28% | 33% | 49% | 63% | 68% | **74%** | 83% | 79% | 78% |

The table shows the fairness metrics of False Positive Rate, False Negative Rate, and Accuracy Rate for different racial groups if we apply different risk scores as thresholds for judicial decision on bail denial and as a predictor of re-offense (i.e. At a score of 6, AI will deny a bail). It shows that 1) there are biases negatively impacting Black Individuals as the errors are higher at each threshold; 2) However, human bias of the judge is larger than machine bias in this case as the judges, on average, is acting as if denying bails for Black individuals with a risk score of 5+ while only denying bails for White and Other individuals with a risk score of 7+. Although AI algorithms have not satisfied fairness metrics, adopting risk score as a universal standard for bail decisions may improve judge’s current decisions.