Chapter 6.1

Machine Bias*

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner

> There's software used across the country to predict future criminals. And it's biased against blacks.

ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

Just as the 18-year-old girls were realizing they were too big for the tiny conveyances—which belonged to a 6-year-old boy—a woman came running after them saying, "That's my kid's stuff." Borden and her friend immediately dropped the bike and scooter and walked away.

But it was too late—a neighbor who witnessed the heist had already called the police. Borden and her friend were arrested and charged with burglary and petty theft for the items, which were valued at a total of $80.

Compare their crime with a similar one: The previous summer, 41-year-old Vernon Prater was picked up for shoplifting $86.35 worth of tools from a nearby Home Depot store (Figure 6.1.1).

Prater was the more seasoned criminal. He had already been convicted of armed robbery and attempted armed robbery, for which he served five years in prison, in addition to another armed robbery charge. Borden had a record, too, but it was for misdemeanors committed when she was a juvenile.

Yet something odd happened when Borden and Prater were booked into jail: A computer program spat out a score predicting the likelihood of each committing a future crime. Borden—who is black—was rated a high risk. Prater—who is white—was rated a low risk.

Two years later, we know the computer algorithm got it exactly backward. Borden has not been charged with any new crimes. Prater is serving an eight-year prison term for subsequently breaking into a warehouse and stealing thousands of dollars' worth of electronics.

Scores like this—known as risk assessments—are increasingly common in courtrooms across the nation. They are used to inform decisions about who can be set free at every stage of the criminal justice system, from assigning bond amounts—as is the case in Fort Lauderdale—to

* Julia Angwin, Jeff Larson, Surya Mattu, Lauren Kirchner, "Machine Bias," *ProPublica* (May 23, 2016). Reprinted with permission.

DOI: 10.1201/9781003278290-37

even more fundamental decisions about defendants' freedom. In Arizona, Colorado, Delaware, Kentucky, Louisiana, Oklahoma, Virginia, Washington and Wisconsin, the results of such assessments are given to judges during criminal sentencing.

Rating a defendant's risk of future crime is often done in conjunction with an evaluation of a defendant's rehabilitation needs. The Justice Department's National Institute of Corrections now encourages the use of such combined assessments at every stage of the criminal justice process. And a landmark sentencing reform bill currently pending in Congress would mandate the use of such assessments in federal prisons.

In 2014, then U.S. Attorney General Eric Holder warned that the risk scores might be injecting bias into the courts. He called for the U.S. Sentencing Commission to study their use. "Although these measures were crafted with the best of intentions, I am concerned that they inadvertently undermine our efforts to ensure individualized and equal justice," he said, adding, "they may exacerbate unwarranted and unjust disparities that are already far too common in our criminal justice system and in our society."

The sentencing commission did not, however, launch a study of risk scores. So ProPublica did, as part of a larger examination of the powerful, largely hidden effect of algorithms in American life.

Figure 6.1.1 Comparison of risk scores for two petty theft arrests.

We obtained the risk scores assigned to more than 7,000 people arrested in Broward County, Florida, in 2013 and 2014 and checked to see how many were charged with new crimes over the next two years, the same benchmark used by the creators of the algorithm.

The score proved remarkably unreliable in forecasting violent crime: Only 20 percent of the people predicted to commit violent crimes actually went on to do so.

When a full range of crimes were taken into account—including misdemeanors such as driving with an expired license—the algorithm was somewhat more accurate than a coin flip. Of those deemed likely to re-offend, 61 percent were arrested for any subsequent crimes within two years.

We also turned up significant racial disparities, just as Holder feared. In forecasting who would re-offend, the algorithm made mistakes with black and white defendants at roughly the same rate but in very different ways (Figure 6.1.5).

- The formula was particularly likely to falsely flag black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants.
- White defendants were mislabeled as low risk more often than black defendants.

Could this disparity be explained by defendants' prior crimes or the type of crimes they were arrested for? No. We ran a statistical test that isolated the effect of race from criminal history and recidivism, as well as from defendants' age and gender. Black defendants were still 77 percent more likely to be pegged as at higher risk of committing a future violent crime and 45 percent more likely to be predicted to commit a future crime of any kind. (Read our analysis.)

The algorithm used to create the Florida risk scores is a product of a for-profit company, Northpointe. The company disputes our analysis.

In a letter, it criticized ProPublica's methodology and defended the accuracy of its test: "Northpointe does not agree that the results of your analysis, or the claims being made based upon that analysis, are correct or that they accurately reflect the outcomes from the application of the model."

Northpointe's software is among the most widely used assessment tools in the country. The company does not publicly disclose the calculations used to arrive at defendants' risk scores, so it is not possible for either defendants or the public to see what might be driving the disparity. (On Sunday, Northpointe gave ProPublica the basics of its future-crime formula — which includes factors such as education levels, and whether a defendant has a job. It did not share the specific calculations, which it said are proprietary.)

Northpointe's core product is a set of scores derived from 137 questions[1] that are either answered by defendants or pulled from criminal records. Race is not one of the questions. The survey asks defendants such things as: "Was one of your parents ever sent to jail or prison?" "How many of your friends/acquaintances are taking drugs illegally?" and "How often did you get in fights while at school?" The questionnaire also asks people to agree or disagree with statements such as "A hungry person has a right to steal" and "If people make me angry or lose my temper, I can be dangerous." (Figure 6.1.2).

The appeal of risk scores is obvious: The United States locks up far more people than any other country, a disproportionate number of them black. For more than two centuries, the key decisions in the legal process, from pretrial release to sentencing to parole, have been in the hands of human beings guided by their instincts and personal biases.

If computers could accurately predict which defendants were likely to commit new crimes, the criminal justice system could be fairer and more selective about who is incarcerated and for how long. The trick, of course, is to make sure the computer gets it right. If it's wrong in one direction,

Figure 6.1.2 Comparison of risk scores for two drug possession arrests.

a dangerous criminal could go free. If it's wrong in another direction, it could result in someone unfairly receiving a harsher sentence or waiting longer for parole than is appropriate.

The first time Paul Zilly heard of his score—and realized how much was riding on it — was during his sentencing hearing on Feb. 15, 2013, in court in Barron County, Wisconsin. Zilly had been convicted of stealing a push lawnmower and some tools. The prosecutor recommended a year in county jail and follow-up supervision that could help Zilly with "staying on the right path." His lawyer agreed to a plea deal.

But Judge James Babler had seen Zilly's scores. Northpointe's software had rated Zilly as a high risk for future violent crime and a medium risk for general recidivism. "When I look at the risk assessment," Babler said in court, "it is about as bad as it could be."

Then Babler overturned the plea deal that had been agreed on by the prosecution and defense and imposed two years in state prison and three years of supervision.

CRIMINOLOGISTS HAVE LONG TRIED to predict which criminals are more dangerous before deciding whether they should be released. Race, nationality and skin color were often used in making such predictions until about the 1970s, when it became politically unacceptable, according to a survey of risk assessment tools by Columbia University law professor Bernard Harcourt.

In the 1980s, as a crime wave engulfed the nation, lawmakers made it much harder for judges and parole boards to exercise discretion in making such decisions. States and the federal

government began instituting mandatory sentences and, in some cases, abolished parole, making it less important to evaluate individual offenders.

But as states struggle to pay for swelling prison and jail populations, forecasting criminal risk has made a comeback.

Dozens of risk assessments are being used across the nation—some created by for-profit companies such as Northpointe and others by nonprofit organizations. (One tool being used in states including Kentucky and Arizona, called the Public Safety Assessment, was developed by the Laura and John Arnold Foundation, which also is a funder of ProPublica.)

There have been few independent studies of these criminal risk assessments. In 2013, researchers Sarah Desmarais and Jay Singh examined 19 different risk methodologies used in the United States and found that "in most cases, validity had only been examined in one or two studies" and that "frequently, those investigations were completed by the same people who developed the instrument."

Their analysis of the research through 2012 found that the tools "were moderate at best in terms of predictive validity," Desmarais said in an interview. And she could not find any substantial set of studies conducted in the United States that examined whether risk scores were racially biased. "The data do not exist," she said.

Since then, there have been some attempts to explore racial disparities in risk scores. One 2016 study examined the validity of a risk assessment tool, not Northpointe's, used to make probation decisions for about 35,000 federal convicts. The researchers, Jennifer Skeem at University of California, Berkeley, and Christopher T. Lowenkamp from the Administrative Office of the U.S. Courts, found that blacks did get a higher average score but concluded the differences were not attributable to bias.

The increasing use of risk scores is controversial and has garnered media coverage, including articles by the Associated Press, and the Marshall Project and FiveThirtyEight last year.

Most modern risk tools were originally designed to provide judges with insight into the types of treatment that an individual might need—from drug treatment to mental health counseling.

"What it tells the judge is that if I put you on probation, I'm going to need to give you a lot of services or you're probably going to fail," said Edward Latessa, a University of Cincinnati professor who is the author of a risk assessment tool that is used in Ohio and several other states.

But being judged ineligible for alternative treatment—particularly during a sentencing hearing—can translate into incarceration. Defendants rarely have an opportunity to challenge their assessments. The results are usually shared with the defendant's attorney, but the calculations that transformed the underlying data into a score are rarely revealed.

"Risk assessments should be impermissible unless both parties get to see all the data that go into them," said Christopher Slobogin, director of the criminal justice program at Vanderbilt Law School. "It should be an open, full-court adversarial proceeding."

Proponents of risk scores argue they can be used to reduce the rate of incarceration. In 2002, Virginia became one of the first states to begin using a risk assessment tool in the sentencing of nonviolent felony offenders statewide. In 2014, Virginia judges using the tool sent nearly half of those defendants to alternatives to prison, according to a state sentencing commission report. Since 2005, the state's prison population growth has slowed to 5 percent from a rate of 31 percent the previous decade.

In some jurisdictions, such as Napa County, California, the probation department uses risk assessments to suggest to the judge an appropriate probation or treatment plan for individuals being sentenced. Napa County Superior Court Judge Mark Boessenecker said he finds the recommendations helpful. "We have a dearth of good treatment programs, so filling a slot in a program with someone who doesn't need it is foolish," he said.

However, Boessenecker, who trains other judges around the state in evidence-based sentencing, cautions his colleagues that the score doesn't necessarily reveal whether a person is dangerous or if they should go to prison.

"A guy who has molested a small child every day for a year could still come out as a low risk because he probably has a job," Boessenecker said. "Meanwhile, a drunk guy will look high risk because he's homeless. These risk factors don't tell you whether the guy ought to go to prison or not; the risk factors tell you more about what the probation conditions ought to be."

Sometimes, the scores make little sense even to defendants.

James Rivelli, a 54-year old Hollywood, Florida, man, was arrested two years ago for shoplifting seven boxes of Crest Whitestrips from a CVS drugstore. Despite a criminal record that included aggravated assault, multiple thefts and felony drug trafficking, the Northpointe algorithm classified him as being at a low risk of reoffending (Figure 6.1.6).

"I am surprised it is so low," Rivelli said when told by a reporter he had been rated a 3 out of a possible 10. "I spent five years in state prison in Massachusetts. But I guess they don't count that here in Broward County." In fact, criminal records from across the nation are supposed to be included in risk assessments.

Less than a year later, he was charged with two felony counts for shoplifting about $1,000 worth of tools from Home Depot. He said his crimes were fueled by drug addiction and that he is now sober.

NORTHPOINTE WAS FOUNDED in 1989 by Tim Brennan, then a professor of statistics at the University of Colorado, and Dave Wells, who was running a corrections program in Traverse City, Michigan.

Wells had built a prisoner classification system for his jail. "It was a beautiful piece of work," Brennan said in an interview conducted before ProPublica had completed its analysis. Brennan and Wells shared a love for what Brennan called "quantitative taxonomy"—the measurement of personality traits such as intelligence, extroversion and introversion. The two decided to build a risk assessment score for the corrections industry.

Brennan wanted to improve on a leading risk assessment score, the LSI, or Level of Service Inventory, which had been developed in Canada. "I found a fair amount of weakness in the LSI," Brennan said. He wanted a tool that addressed the major theories about the causes of crime.

Brennan and Wells named their product the Correctional Offender Management Profiling for Alternative Sanctions, or COMPAS. It assesses not just risk but also nearly two dozen so-called "criminogenic needs" that relate to the major theories of criminality, including "criminal personality," "social isolation," "substance abuse" and "residence/stability." Defendants are ranked low, medium or high risk in each category.

As often happens with risk assessment tools, many jurisdictions have adopted Northpointe's software before rigorously testing whether it works. New York State, for instance, started using the tool to assess people on probation in a pilot project in 2001 and rolled it out to the rest of the state's probation departments—except New York City—by 2010. The state didn't publish a comprehensive statistical evaluation of the tool until 2012. The study of more than 16,000 probationers found the tool was 71 percent accurate, but it did not evaluate racial differences.

A spokeswoman for the New York state division of criminal justice services said the study did not examine race because it only sought to test whether the tool had been properly calibrated to fit New York's probation population. She also said judges in nearly all New York counties are given defendants' Northpointe assessments during sentencing.

In 2009, Brennan and two colleagues published a validation study that found that Northpointe's risk of recidivism score had an accuracy rate of 68 percent in a sample of 2,328 people. Their study

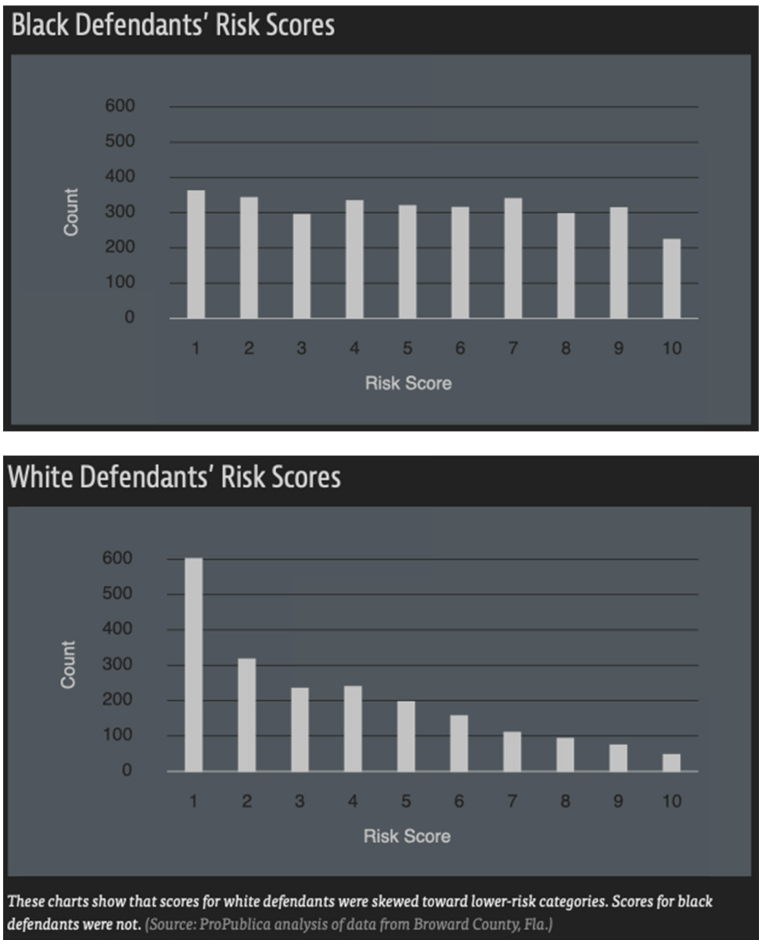


Figure 6.1.3 Distribution of White and Black defendants' risk scores.

also found that the score was slightly less predictive for black men than white men—67 percent versus 69 percent. It did not examine racial disparities beyond that, including whether some groups were more likely to be wrongly labeled higher risk (Figure 6.1.3).

Brennan said it is difficult to construct a score that doesn't include items that can be correlated with race—such as poverty, joblessness and social marginalization. "If those are omitted from your risk assessment, accuracy goes down," he said.

In 2011, Brennan and Wells sold Northpointe to Toronto-based conglomerate Constellation Software for an undisclosed sum.

Wisconsin has been among the most eager and expansive users of Northpointe's risk assessment tool in sentencing decisions. In 2012, the Wisconsin Department of Corrections launched the use of the software throughout the state. It is used at each step in the prison system, from sentencing to parole (Figure 6.1.4).

In a 2012 presentation, corrections official Jared Hoy described the system as a "giant correctional pinball machine" in which correctional officers could use the scores at every "decision point."

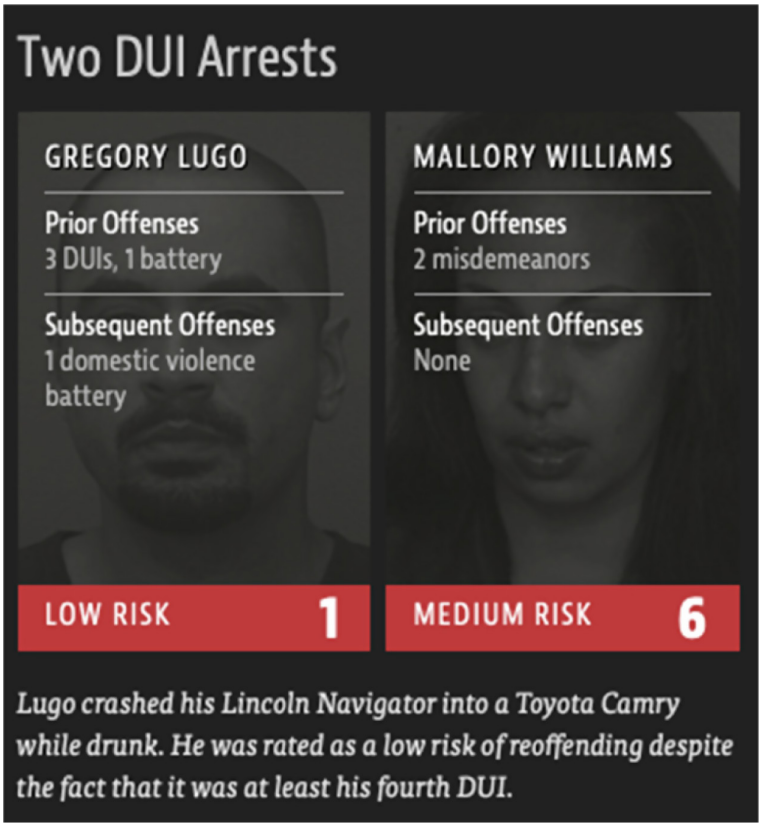


Figure 6.1.4 Comparison of risk scores for two DUI arrests.

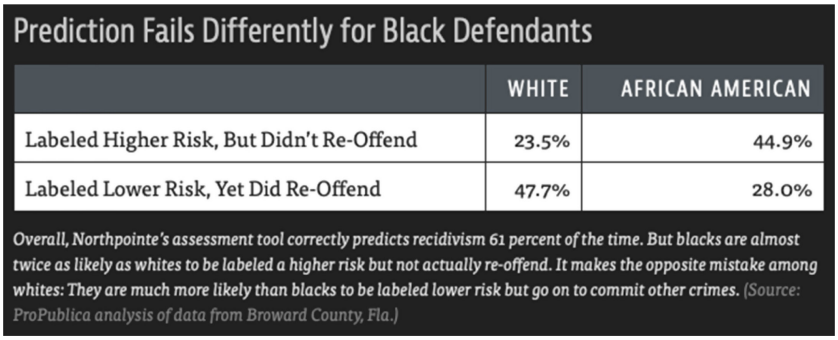
Wisconsin has not yet completed a statistical validation study of the tool and has not said when one might be released. State corrections officials declined repeated requests to comment for this article.

Some Wisconsin counties use other risk assessment tools at arrest to determine if a defendant is too risky for pretrial release. Once a defendant is convicted of a felony anywhere in the state, the Department of Corrections attaches Northpointe's assessment to the confidential presentence report given to judges, according to Hoy's presentation.

In theory, judges are not supposed to give longer sentences to defendants with higher risk scores. Rather, they are supposed to use the tests primarily to determine which defendants are eligible for probation or treatment programs.

But judges have cited scores in their sentencing decisions. In August 2013, Judge Scott Horne in La Crosse County, Wisconsin, declared that defendant Eric Loomis had been "identified, through the COMPAS assessment, as an individual who is at high risk to the community." The judge then imposed a sentence of eight years and six months in prison.

Loomis, who was charged with driving a stolen vehicle and fleeing from police, is challenging the use of the score at sentencing as a violation of his due process rights. The state has defended Horne's use of the score with the argument that judges can consider the score in addition to other factors. It has also stopped including scores in presentencing reports until the state Supreme Court decides the case.



Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

Figure 6.1.5 Percent of mistakes for White and Black defendants.

"The risk score alone should not determine the sentence of an offender," Wisconsin Assistant Attorney General Christine Remington said last month during state Supreme Court arguments in the Loomis case. "We don't want courts to say, this person in front of me is a 10 on COMPAS as far as risk, and therefore I'm going to give him the maximum sentence."

That is almost exactly what happened to Zilly, the 48-year-old construction worker sent to prison for stealing a push lawnmower and some tools he intended to sell for parts. Zilly has long struggled with a meth habit. In 2012, he had been working toward recovery with the help of a Christian pastor when he relapsed and committed the thefts.

After Zilly was scored as a high risk for violent recidivism and sent to prison, a public defender appealed the sentence and called the score's creator, Brennan, as a witness.

Brennan testified that he didn't design his software to be used in sentencing. "I wanted to stay away from the courts," Brennan said, explaining that his focus was on reducing crime rather than punishment. "But as time went on I started realizing that so many decisions are made, you know, in the courts. So I gradually softened on whether this could be used in the courts or not."

Still, Brennan testified, "I don't like the idea myself of COMPAS being the sole evidence that a decision would be based upon."

After Brennan's testimony, Judge Babler reduced Zilly's sentence, from two years in prison to 18 months. "Had I not had the COMPAS, I believe it would likely be that I would have given one year, six months," the judge said at an appeals hearing on Nov. 14, 2013.

Zilly said the score didn't take into account all the changes he was making in his life—his conversion to Christianity, his struggle to quit using drugs and his efforts to be more available for his son. "Not that I'm innocent, but I just believe people do change."

FLORIDA'S BROWARD COUNTY, where Brisha Borden stole the Huffy bike and was scored as high risk, does not use risk assessments in sentencing. "We don't think the [risk assessment] factors have any bearing on a sentence," said David Scharf, executive director of community programs for the Broward County Sheriff's Office in Fort Lauderdale.

Broward County has, however, adopted the score in pretrial hearings, in the hope of addressing jail overcrowding. A court-appointed monitor has overseen Broward County's jails since 1994 as a result of the settlement of a lawsuit brought by inmates in the 1970s. Even now, years later, the Broward County jail system is often more than 85 percent full, Scharf said.

In 2008, the sheriff's office decided that instead of building another jail, it would begin using Northpointe's risk scores to help identify which defendants were low risk enough to be released on bail pending trial. Since then, nearly everyone arrested in Broward has been scored soon after

Figure 6.1.6 Comparison of risk scores for two shoplifting arrests.

being booked. (People charged with murder and other capital crimes are not scored because they are not eligible for pretrial release.)

The scores are provided to the judges who decide which defendants can be released from jail. "My feeling is that if they don't need them to be in jail, let's get them out of there," Scharf said.

Scharf said the county chose Northpointe's software over other tools because it was easy to use and produced "simple yet effective charts and graphs for judicial review." He said the system costs about $22,000 a year.

In 2010, researchers at Florida State University examined the use of Northpointe's system in Broward County over a 12-month period and concluded that its predictive accuracy was "equivalent" in assessing defendants of different races. Like others, they did not examine whether different races were classified differently as low or high risk.

Scharf said the county would review ProPublica's findings. "We'll really look at them up close," he said.

Broward County Judge John Hurley, who oversees most of the pretrial release hearings, said the scores were helpful when he was a new judge, but now that he has experience he prefers to rely on his own judgment. "I haven't relied on COMPAS in a couple years," he said.

Hurley said he relies on factors including a person's prior criminal record, the type of crime committed, ties to the community, and their history of failing to appear at court proceedings.

ProPublica's analysis reveals that higher Northpointe scores are slightly correlated with longer pretrial incarceration in Broward County. But there are many reasons that could be true other

than judges being swayed by the scores—people with higher risk scores may also be poorer and have difficulty paying bond, for example.

Most crimes are presented to the judge with a recommended bond amount, but he or she can adjust the amount. Hurley said he often releases first-time or low-level offenders without any bond at all.

However, in the case of Borden and her friend Sade Jones, the teenage girls who stole a kid's bike and scooter, Hurley raised the bond amount for each girl from the recommended $0 to $1,000 each.

Hurley said he has no recollection of the case and cannot recall if the scores influenced his decision.

The girls spent two nights in jail before being released on bond.

"We literally sat there and cried" the whole time they were in jail, Jones recalled. The girls were kept in the same cell. Otherwise, Jones said, "I would have gone crazy." Borden declined repeated requests to comment for this article.

Jones, who had never been arrested before, was rated a medium risk. She completed probation and got the felony burglary charge reduced to misdemeanor trespassing, but she has still struggled to find work.

"I went to McDonald's and a dollar store, and they all said no because of my background," she said. "It's all kind of difficult and unnecessary."

Note

1. https://www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE.html