American Economic Journal: Economic Policy 2024, 16(4): 382–414
https://doi.org/10.1257/pol.20220620

Algorithmic Risk Assessment in the Hands of Humans[†]

By Megan T. Stevenson and Jennifer L. Doleac*

We evaluate the impacts of adopting algorithmic risk assessments in sentencing. We find that judges changed sentencing practices in response to the risk assessment, but that discretion played a large role in mediating its impact. Judges deviated from the recommendations associated with the algorithm in systematic ways, suggestive of alternative objectives. As a result, risk assessment did not lead to detectable gains in terms of public safety or reduced incarceration rates. Using simulations, we show that strict adherence to the sentencing recommendations associated with the algorithm would have had benefits (less incarceration) but also some costs (increased sentences for youth). (*JEL* D81, D91, H76, K41, K42)

Algorithmic predictions of future offending, known as risk assessments, are proliferating in criminal justice. They are used to help inform decision-making at almost every stage of criminal proceedings: setting bail, sentencing, determining probation supervision levels, placement within the prison system, and parole. Developed by statistically analyzing court and police records to identify which factors best predict recidivism, risk assessment tools are widely believed to be more accurate than humans (Grove and Meehl 1996; Lin et al. 2020). Risk assessment's proponents argue that we can lower incarceration rates and/or crime by making incarceration decisions on the basis of machine predictions (M) instead of human intuition (H) (Berk, Sorenson, and Barnes 2016; Jung et al. 2017; Kleinberg et al. 2018).

Research in this area has largely made the simplifying assumption that machines would take the place of humans (M versus H), but in the real world, risk assessments are used almost exclusively as a supplement to human discretion (M + H). Judges are free to follow or ignore the recommendations associated with the risk assessment. They may override the risk assessment when they think it's wrong, and they may override it when they are pursuing alternative goals. Thus, the relevant question is not man versus machine, but rather, how do the two interact? The impact of risk assessment

* Stevenson: University of Virginia (email: mstevenson@law.virginia.edu); Doleac: Arnold Ventures (email: jdoleac@arnoldventures.org). Matthew Notowidigdo was coeditor for this article. Stevenson is lead author. Both authors contributed to background research, idea development, conceptual framing, and writing. We are grateful to Kelsey Pukelis and Arjun Ravi for excellent research assistance. We also thank Alex Albright, Patrick Bayer, Ted Enamorado, Meredith Farrar-Owens, Brandon Garrett, Sandra Mayson, John Monahan, Michael Mueller-Smith, Derek Neal, Aurelie Ouss, Felix Owusu, Ben Schoenfield, Sonja Starr, the staff at the Virginia Criminal Sentencing Commission (VCSC), and numerous seminar participants for their helpful feedback. We received IRB approval from the University of Virginia (UVA IRB-SBS 4121).

[†] Go to https://doi.org/10.1257/pol.20220620 to visit the article page for additional materials and author disclosure statement(s) or to comment in the online discussion forum.

depends not only on the properties of the algorithm but on how the algorithm enters into the objective function of its user. This can lead to unexpected results.

This study is about what happens when risk assessments are placed in the hands of humans. In the early 2000s, Virginia courts began using risk assessments at sentencing for defendants convicted of nonviolent offenses. They were incorporated into the sentence guidelines in order to divert a large share of low-risk nonviolent offenders from jail or prison. Other offenders were not affected by the policy change, and judges retained final authority over sentencing. We evaluate the impacts of risk assessment adoption to provide new information about how these tools are used, who they impact most, and to what extent their use results in a reduction in incarceration and/or crime. We find that risk assessment use did not bring any detectable benefit in terms of public safety or reduced incarceration. We argue that this is likely due to conflicting objectives in sentencing. Judges used their discretion to mitigate some of the adverse consequences of risk assessment adoption at the expense of reducing potential gains.

We begin by evaluating whether risk assessment affected sentencing at all, as well as whether use changed over time. Our first strategy takes advantage of the fact that those who score right below the low-risk cutoff get a diversion recommendation, while those who score right above it do not. Using regression discontinuity, we find sharp changes in sentencing around the low-risk cutoff: being marginally above the cutoff leads to a 6 percentage point increase in the likelihood of incarceration and a sentence that is approximately 24 percent longer. Our second strategy is an event-study design that documents a sharp increase in the correlation between sentencing and the predicted risk score that corresponds with the adoption of risk assessment. Finally, we use both of these methods to show that judges used the risk assessment most in the first few years after it was adopted. After that, they appear to have mostly discontinued use.

We next evaluate how risk assessment's use affected sentencing relative to the status quo (i.e., judicial decision-making without risk assessment). We use a difference-in-difference research design that compares risk-assessment-eligible/risk-assessment-ineligible defendants before/after risk assessment was adopted. We benchmark our estimates of the impacts of risk assessment in the hands of humans against simulations of sentencing by algorithm, or risk assessment use without discretion. Following Virginia policy, our simulated sentences entail automatic diversion for low-risk defendants after risk assessment was adopted.

Sentencing by algorithm would have resulted in a sharp decrease in both the probability of incarceration and the sentence length. In practice, however, judges diverted far fewer individuals than were recommended by the algorithm. The net effect on both the incarceration rate and sentence length was near zero, and confidence intervals allow us to reject even small decreases. In other words, prison and jail terms appear to have been reallocated from defendants rated low risk to those rated high risk so that overall incarceration for nonviolent offenders remained the same.

We then look at discretion in use. Judges failed to divert low-risk defendants 63 percent of the time and diverted 25 percent of high-risk defendants against recommendation. This suggests that judges respond to the risk assessment selectively. We run a series of tests to see which demographic factors predict deviating upward or downward in sentencing. We find that, conditional on the risk score, judges are substantially more lenient with young defendants than older defendants and substantially harsher on Black defendants than non-Black defendants. These disparities are not explained by actual differences in recidivism risk across race or age.

Next, we explore how risk assessment affected race and age disparities. As previously, we benchmark our estimates of risk assessment in the hands of humans against simulated impacts of sentencing by algorithm. Using a triple-differences design, we find little evidence that risk assesment increased racial disparities, either in real life or in simulations. Standard errors preclude ruling out a moderate-sized increase or decrease.

In contrast, sentencing by algorithm would have led to a sharp 24 percent increase in relative sentence length for young defendants. However, judicial discretion mitigated the adverse impacts on this demographic group. While risk assessment use did increase age disparities, the magnitude of the increase (12 percent) was half as much as it would have been in the absence of discretion.

Finally, we evaluate risk assessment's impact on public safety. Using the same difference-in-difference design, we find no evidence that risk assessment use reduced recidivism. We can reject any reduction greater than 0.02 percentage points in the likelihood of a new felony conviction within 3 years over a mean of 19 percentage points. We also conduct simulations of recidivism effects under sentencing by algorithm with the caveat that these require strong assumptions about counterfactual crime rates. Our simulations suggest that sentencing by algorithm would have led to a 1.6 percentage point increase in recidivism. Virginia's risk assessment policy was designed to increase release rates, thereby mechanically increasing the potential for accruing new charges and convictions.

In theory, reshuffling jail beds to increase incarceration for those with high-risk scores and reduce it for those with low-risk scores should have led to reduced recidivism. While we cannot provide a definitive answer as to why that did not happen, we suggest a partial explanation. Sentencing is a multivalent decision, and incapacitating those at highest risk of reoffending is only one objective. Other objectives may occasionally conflict. For instance, young age is one of the most important predictors of recidivism and adds more points to the risk score than any other factor. But if judges are reluctant to give long sentences to young people, this makes one of the most important inputs to the risk score effectively off-limits. This, combined with other instances in which judges may have deviated from the risk score to follow alternative goals, could undermine the potential gains from risk assessment use. A risk assessment that is neutered by conflicting goals may not result in substantial improvements to public safety.

Our paper provides one of the first evaluations of the role of discretion in real-world risk assessment use. Most prior studies either focus solely on the algorithm or compare algorithms to humans. The prior literature has demonstrated risk assessment's accuracy (e.g., Berk, Sorenson, and Barnes 2016), conducted horse races of M versus H (e.g., Dressel and Farid 2018; Lin et al. 2020), simulated the impact of replacing humans with algorithms (e.g., Kleinberg et al. 2018), provided evidence of racial bias in risk assessment (e.g., Arnold, Dobbie, and Hull 2021), suggested corrections for algorithmic bias (e.g., Yang and Dobbie 2020; Johndrow and Lum 2019), and contested definitions of algorithmic fairness (e.g., Lipton, Chouldechova, and McAuley

2017; Kasy and Abebe 2021). Most studies that focus on discretion have taken place in the lab (e.g., Skeem, Scurich, and Monahan 2020). Others have looked at real-world use but lack a comparison to decision-making without algorithms (e.g., Garrett and Monahan 2020). The small literature evaluating the impacts of switching from H to M + H has mostly focused on outcomes, leaving the role of discretion somewhat black box (Berk et al. 2002; Berk 2017; Stevenson 2019; Sloan, Naufal, and Caspers 2023; Cowgill 2019; Imai et al. 2020). Albright (2023) is an exception, showing that the increase in racial disparities observed after the adoption of pretrial risk assessment was mostly due to discretion in use.

Our paper also provides useful information for policymakers considering risk assessment adoption. First, it suggests tempering expectations about the benefits of risk assessment in practice. Most criminal justice decisions take into account a variety of concerns beyond the risk of reoffending.[1] Some of these goals, like leniency for the young, conflict with the objective of incarcerating those at highest risk of reoffending. Conflicting objectives will limit potential gains and make the impacts hard to predict. Second, it demonstrates that risk assessment is an unreliable method of nudging judges toward specific policy goals (Cowgill and Stevenson 2020). As in Virginia, risk assessment is frequently adopted with the goal of reducing jail or prison populations. Policymakers hope to achieve that goal by pairing the risk score with recommendations for release. If judges are not equally committed to lowering incarceration, the recommendations may have little bite. Third, our results highlight the limits of algorithm-based solutions to concerns about algorithmic bias. Here, we show that judges sentence Black defendants much more harshly than White defendants with the same risk score. While a bias-adjusted risk assessment may reduce some of the disparities in the score, it is unlikely to affect disparities in use.

Finally, our work contributes to a broader literature about the use of predictive algorithms. A number of simulation-based studies have forecast substantial gains from adopting predictive tools in a variety of settings (Chetty, Friedman, and Rockoff 2014; Berk, Sorenson, and Barnes 2016; Goel, Rao, and Shroff 2016; Chalfin et al. 2016; Jung et al. 2017; Kleinberg et al. 2018). In many of these settings, however, machines are unlikely to fully replace the human. When there is a human decision-maker tasked with using the algorithm, that person's objective function is central to understanding the impacts of algorithm adoption (Hoffman, Kahn, and Li 2017). An early literature has begun to explore this in settings such as medicine and beyond (Abaluck et al. 2020; Gruber et al. 2020). We hope that the themes of M + H interaction explored here will provide a useful framework to think through these issues in other settings.

The paper proceeds as follows: Section I provides further background on risk assessment algorithms and the Virginia policy context; it also describes our data and presents some descriptive statistics. Section II demonstrates that judges do use the risk assessment, even though use decays over time. Section III examines the role of discretion in risk assessment's impacts. This section shows how risk assessment adoption affected overall sentencing and race/age disparities and benchmarks this against simulations of sentencing by algorithm. This section also shows

[1] Even when law specifies that a decision should hinge only on risk of misconduct, as in many bail statutes, judges almost certainly take other factors into account in practice.

which demographic features predict deviations from the risk score. Section IV looks at risk assessment's impact on recidivism. Section V concludes.

I. Background

A. *Algorithmic Risk Assessment in Criminal Justice*

Many decisions made in criminal justice take predictions about the risk of committing crime at least partially into account. In some settings, such as sentencing, risk of future crime is one factor among a wide range of legally permissible considerations. In others, such as bail, the legally permissible rationales are often limited to the risk of crime, nonappearance, or tampering with the case. Such legal formalities, however, are hard to enforce and, thus, may have little constraining power. In practice, judges usually have substantial latitude to make discretionary decisions on the basis of whatever factors they feel are relevant.

Risk assessment tools aim to improve the accuracy of crime predictions and are now used broadly in criminal justice when determining bail and pretrial custody, probation supervision levels, offender placement within the prison system, release on parole, and sentencing.[2] Risk assessment tools are used at sentencing in 27 states; at least 6 additional states use risk assessment at sentencing in some counties. (See online Appendix A1 for a full list.)

The construction and implementation of risk assessment follows a fairly typical pattern across different jurisdictions and even across different settings (e.g., sentencing, bail, etc.). Here we describe the standard approach, which Virginia follows as well. Risk assessment tools are built by statistically analyzing court and/or police records to evaluate how various inputs correlate to measures of future offending, such as rearrest or reconviction. The most common inputs include age, gender, and criminal history (prior convictions, arrests, periods of incarceration, etc.). Almost all risk assessment tools currently used today take the form of a weighted checklist, where inputs are assigned point values based on their statistical correlation with future offending, as measured by rearrest or reconviction. The risk score is the sum of the points. This type of simple linear model is not able to exploit nonlinear or interactive relationships in the same way that more sophisticated machine learning algorithms do. However, evidence suggests that simple tools, with integer weights and only a handful of inputs, can rival the accuracy of complex prediction models in a wide variety of domains (Jung et al. 2017).[3] As discussed below, we find this to be true in our setting as well. Beyond accuracy, an integer-weighted model is preferred due to ease of implementation: the administrator can simply add up the points by hand.

[2] Predictive algorithms are also used in policing, but the tools and applications differ somewhat from the court context discussed in this paper.

[3] Dressel and Farid (2018) shows that a simple classifier that includes only two inputs—age and the number of prior convictions—performs as well as a nonlinear support vector machine with a wider range of inputs and as well as the commonly used Correctional Offender Management Profiling for Alternative Sanctions Risk Assessment (COMPAS) algorithm. Angelino et al. (2018) and Jung et al. (2017) show similar results in other criminal justice contexts.

Judges are almost never provided with statistical probabilities. Rather, they are informed of the defendant's risk classification. Defendants who score above/below certain cutoffs in the risk score will receive classifications such as low, moderate or high risk, or a ranking within a limited range (e.g., 1 through 6). Usually, jurisdictions will have a policy that states which intervention (e.g., probation or incarceration) is recommended for defendants with different risk classifications. For example, jurisdictions might use a grid method, where the severity of the offense and the risk classification determine the recommended sentence or bail amount. However, judges or other criminal justice decision-makers almost always retain discretion to deviate from the recommended actions.

B. *Felony Sentencing in Virginia*

Virginia is one of the earliest adopters of criminal justice risk assessments. Beyond that, the state is fairly typical in terms of sentencing practices, incarceration rates, and racial composition.[4] Virginia uses a voluntary sentence guidelines regime in which judges are recommended, but not required, to sentence within a particular range.[5] This system has been in place since the 1980s. In 1994, Virginia adopted a major truth-in-sentencing reform act that abolished parole, mandated that offenders serve at least 85 percent of their sentence, and increased sentences for violent offenders. In order to free up state prison beds for violent offenders who were expected to serve longer terms as a result of the reform, Virginia also set the goal of diverting 25 percent of nonviolent offenders from jail or prison. Risk assessment was the proposed method of achieving this. It was to be designed and implemented by the newly founded Virginia Criminal Sentencing Commission (VCSC).

While most of the truth-in-sentencing reforms were implemented immediately after the bill passed, risk assessment came later. The VCSC designed a new risk tool for nonviolent offenders in the late 1990s and piloted it in six judicial circuits. They launched a revised risk assessment statewide on July 1, 2002. This tool was in use until 2013, when several further revisions were made. Our analysis is focused around the time that risk assessment was adopted statewide and uses data from judicial circuits that did not participate in the pilot.[6]

Virginia's risk assessment was developed by running logistic regression on a randomly selected sample of nonviolent offenders who either had received a noncarceral sentence or had recently been released from jail or prison. It was designed to predict which defendants would be reconvicted of a felony within three years of their return

[4] In 2003—the year that risk assessment was adopted in sentencing—Virginia had 472 incarcerated prisoners per 100,000 people, compared to a national average of 482 per 100,000 (Harrison and Beck 2004). Non-Hispanic Whites were 68 percent of the population, compared to a US average of 67 percent. However, Virginia had a slightly larger Black population and slightly smaller Hispanic population (19 percent and 6 percent, respectively) compared to the rest of the country (12 percent and 13 percent) (usafacts.org). Virginia is one of 17 states that use sentence guidelines and one of 18 states that have abolished parole (Mitchell 2017).

[5] The information in this section was derived from interviews in 2019 with the director of the VCSC, Meredith Farrar-Owens, as well as Ostrom et al. (1999), VCSC (2001), and Ostrom et al. (2002).

[6] We have also evaluated the impacts of risk assessment in the pilot circuits and find results that are qualitatively similar to those shown here. We do not make use of the pilot period in a panel difference-in-difference design because the instrument implemented then was meaningfully different from the one later implemented statewide.

to the community. Its inputs include indicators for whether the charge of conviction was drug, larceny, or fraud; whether there were any additional offenses; gender, age, employment, and marital status; recent arrests or confinement; prior felony convictions or adjudications; and prior adult incarcerations. The logistic regression coefficients were converted into weights and summed to produce a risk assessment score between 0 and 95.

Those who score below a 36 in Virginia's risk assessment are classified as low risk.[7] This cutoff score was selected based on the legal mandate to identify the 25 percent lowest risk defendants for diversion from jail or prison. For defendants whose guidelines-recommended sentence is prison (more than 12 months), diversion means probation or a shorter jail sentence. For defendants whose guidelines-recommended sentence is jail (less than or equal to 12 months), diversion means probation or some other noncarceral sentence.

The risk scores, as well as the sentence guidelines, are calculated using a set of worksheets that are filled out by a probation officer or a prosecutor either during the plea-bargaining process or after conviction. The nonviolent risk assessment is only administered on eligible cases: a defendant must be convicted of a drug, larceny, or fraud charge and cannot have a prior or current conviction for a violent crime. Since the tool is designed for diversion, only those who were recommended for jail or prison by the sentence guidelines are eligible. The final two eligibility requirements are that defendants must not have sold large quantities of cocaine or been convicted of an offense that includes a mandatory term of incarceration.

After conviction, the risk assessment and sentence guidelines worksheets are provided to the judge for sentencing.[8] The cover page of the worksheets prominently displays the sentence range that is recommended by the guidelines. The midpoint of the range is referred to as the guidelines-recommended sentence. A sentence within this range is considered guidelines compliant. Nonviolent offenders have a checkbox with the label "recommended for alternative punishment" right underneath the guidelines-recommended sentence. If the defendant scores below the low-risk cutoff, this box will be checked, and the set of guidelines-compliant sentences is expanded to include release without incarceration. If a judge chooses a sentence that is not guidelines compliant, they must provide a written justification. A judge who is interested in the exact risk score, as opposed to merely the risk classification and its associated sentence recommendation, can find it by flipping through the pages of the worksheets.

After sentencing, the final sentence is written on a separate worksheet. All sheets are mailed to the VCSC, which digitizes the information and maintains a database on all felony sentences in Virginia. The VCSC also maintains detailed records of every change to the sentence guidelines and every change to sentencing policy going back at least to 1995. This meticulous recordkeeping helps ensure that there are no important policy changes concurrent to the adoption of risk assessment that would confound our analysis. The only important policy change around the

[7] This cutoff was increased by 3 points on July 1, 2004. Unfortunately, this change affected too few defendants to be able to evaluate its impacts.

[8] The risk assessment worksheet and the sentencing cover page are provided at the end of the online appendix.

time of risk assessment adoption pertains to those convicted of sex offenses.[9] Sex offenders constitute only 3 percent of the sample and are dropped from the analysis. Beyond this, we find only trivial changes to sentencing practices during the time period of our main analysis.[10]

In this paper, we frequently discuss how judges use the risk assessment tool. We use this language for purposes of concision and because judges are formally responsible for determining the sentence. However, other criminal justice actors are also influential. About 85 percent of felony convictions in Virginia come from guilty pleas, which are often accompanied by sentencing agreements between the prosecution and the defense.[11] The risk assessment information is available during plea negotiations, and both prosecutors and public defenders report using it frequently (Metz et al. 2020).[12] This could be because they independently value the predicted likelihood of reoffending in determining the appropriate sentence. It could also be because plea negotiations occur within the shadow of the judge. Namely, since negotiated sentences need to be approved by the judge, plea negotiations are influenced by expectations about what the judge is likely to approve.

Judges are appointed by majority vote of the Virginia General Assembly.[13] They are up for reappointment every eight years. The reappointment procedure includes an interview with the legislative committee and then a vote in the general assembly. Compliance with sentencing guidelines is monitored by the VCSC, was reported (at the state level) in an annual report, and became formally included in the reappointment evaluation in 2009. Compliance rates have hovered around 80 percent since fiscal year 2000, with deviations evenly split between aggravation and mitigation.[14]

C. *Data and Descriptive Statistics*

The data used in this paper come from two sources. The primary data source is the VCSC (Virginia Criminal Sentencing Commission 2018), which collects and maintains records on all felony sentences in the state of Virginia since 1995. These data were acquired through a public records request and contain information relevant to the case, such as charges, sentences, guidelines-recommended sentences, and risk score. In addition, it contains dozens of variables that are used to calculate the risk score and the guidelines-recommended sentence, such as those pertaining to the current offense, the criminal record, and the personal characteristics of the defendant. Unfortunately, neither

[9] Virginia adopted a different risk assessment for sex offenders in fiscal year 2002.

[10] The only change concurrent with the adoption of the nonviolent risk assessment is a policy shift that makes it easier for prosecutors and probation officers to access juvenile records. However, this change applies to all cases, and we have no reason to believe that this will impact risk-assessment-eligible defendants more than anyone else.

[11] Approximately 10 percent of convictions come from a bench trial, in which the judge adjudicates guilt or innocence as well as determining the sentence (VCSC 2003). Jury trials constitute roughly 2 percent of convictions. The jury is responsible for the sentence in these cases, although the judge is authorized to reduce the sentence if he or she so chooses. The remainder are Alford pleas, in which the defendant doesn't contest the adjudication but doesn't admit guilt either.

[12] Both prosecutors and defense attorneys, however, report some distrust of the tool and describe using it in conjunction with their own experience-based assessment (Metz et al. 2020).

[13] The head prosecutors, known in Virginia as commonwealth's attorneys, are directly elected to serve four-year terms. They set policy, manage operations in each individual jurisdiction, and appoint line prosecutors to carry out the day-to-day work of prosecution.

[14] See the annual reports put out by the VCSC: http://www.vcsc.virginia.gov/reports.html.

the risk score nor the exact variables used to calculate the risk score are available for defendants who did not receive one. Therefore, we do not have risk score information for eligible defendants sentenced before risk assessment was adopted, nor for defendants whose charge/criminal history makes them ineligible for risk assessment. VCSC data also do not contain the race or gender of the defendants.

Race, gender, and an alternative recidivism measure (new charges) were obtained by matching the VCSC data with bulk court records scraped from a public online site (Virginia Circuit Courts 2018). For most counties, court records from as far back as 2000 can be found online.[15] Alexandria and Fairfax counties are an exception; we were unable to match VCSC data back to the original court records for cases in these counties. The match was conducted using the fastLink package in R, which conducts probabilistic matching across multiple variables (Enamorado, Fifield, and Imai 2019). The variables used for the match include the first name, last name, offense date, birth month, and county of arrest. With a probabilistic match threshold of 0.92, 96 percent of post-1999 cases in the VCSC data (excluding Alexandria and Fairfax) found a match in the court records.

As mentioned previously, we have two primary sample restrictions: we drop those convicted of sex offenses (3 percent of the sample) and we drop judicial circuits that had already piloted a different risk assessment (23 percent). When we use the phrase "all cases" in this paper, we are referring to all cases after pilot circuits and sex offenders have been dropped.

Descriptive statistics for defendants sentenced in the two years preceding risk assessment adoption (fiscal years 2001 and 2002) can be found in Table 1. The first column shows data for risk-assessment-eligible offenders. This includes all defendants who were convicted of a drug, larceny, or fraud offense; whose guidelines-recommended sentence was jail or prison; and who did not have a current or prior violent conviction.[16] The second column shows descriptive statistics for all other offenders: the primary control group for the difference-in-difference analysis. The third column shows the full sample.

Roughly one-third of the sample is convicted of a risk-assessment-eligible offense. Drug convictions make up the largest share of these cases, followed by larceny, and then fraud. A substantial portion of the ineligible group is also convicted of these three offenses. Other common offenses in this category include traffic, assault, burglary, and robbery. The probability of incarceration is slightly higher for the eligible group, likely because one of the eligibility criteria is a guidelines-recommended sentence of jail or prison. However, the eligible group lacks the long right sentencing tail seen in violent offenses. The average sentence for the ineligible group is almost twice as long as that of the eligible group.

Our primary measure of recidivism is an indicator for having been convicted of a new felony offense in the state of Virginia, committed within three years of the

[15] Many thanks to Ben Schoefield for the public service of scraping the data and making them available to researchers.

[16] This group is expected to be slightly overinclusive, since we can't identify offenders who would have been ineligible due to a mandatory minimum sentence or large quantities of cocaine unless they were sentenced during the post-risk-assessment period. However, in cases for which data are available, these ineligibility criteria rule out only 2 percent and 1 percent of cases, respectively.

Table 1—Summary Statistics from Fiscal Years 2001 and 2002

	Eligible	Ineligible	All
Variables from VCSC data:			
	Mean	Mean	Mean
Age	31.29	31.51	31.43
Under 23	0.26	0.29	0.28
Recent felony conviction	0.13	0.08	0.10
Drug	0.52	0.23	0.33
Larceny	0.29	0.16	0.21
Fraud	0.19	0.10	0.13
Burglary	0.00	0.11	0.07
Robbery	0.00	0.06	0.04
Assault	0.00	0.10	0.06
Traffic	0.00	0.17	0.11
Guidelines-recommended sentence	13.18	25.28	20.84
Pr(incarcerated)	0.79	0.59	0.66
Months incarcerated	10.18	21.83	17.56
New felony conviction—three years	0.17	0.12	0.14
Observations	11,533	19,880	31,413
Variables from court data:			
	Mean	Mean	Mean
Black	0.57	0.47	0.51
Female	0.24	0.19	0.21
New charge—three years	0.44	0.33	0.37

Notes: This table provides summary statistics from fiscal years 2001 and 2002—the two years before risk assessment was adopted. All variables except for age, the guidelines-recommended sentence, and months of incarceration are dummy variables. Age is measured in years and both the guidelines-recommended sentence and the actual sentence are measured in months. The likelihood of receiving a new charge or conviction within three years is measured from the time of sentencing. From left to right, the columns show statistics for those who are risk assessment eligible, those who are not eligible for risk assessment, and the full sample.

original sentencing date. We focus on this measure because it is very similar to the target variable that Virginia's nonviolent risk assessment was trained to predict. The main difference is that our time counter for recidivism begins at sentencing, and Virginia's time counter for recidivism begins at release. In our recidivism measure, incarcerated defendants will be incapacitated (at least for crimes in which the victims are outside of prison walls) during a portion of the recidivism time window. This is intentional: we are curious about the extent to which risk assessment use lowers recidivism by incapacitating those at the highest risk of reoffending. According to this definition, 17 percent of risk-assessment-eligible people and 12 percent of ineligible people recidivate.

The bottom panel shows descriptive statistics for variables derived from the circuit court dataset. As above, these descriptive statistics are for cases sentenced in fiscal years 2001 and 2002. Of eligible defendants, 57 percent are Black; of ineligible defendants, 47 percent are Black. Almost all of the remaining defendants are labeled as White in our data. Women composed 24 percent of eligible offenders and 19 percent of ineligible. Our second recidivism measure—being charged with a new offense in circuit court within three years of sentencing—shows higher rates than the previous measure. This measure captures both new felonies and new misdemeanors allegedly committed while still under supervision for the original offense. Of risk-assessment-eligible offenders, 44 percent have a new charge within 3 years compared to 33 percent of ineligible offenders.

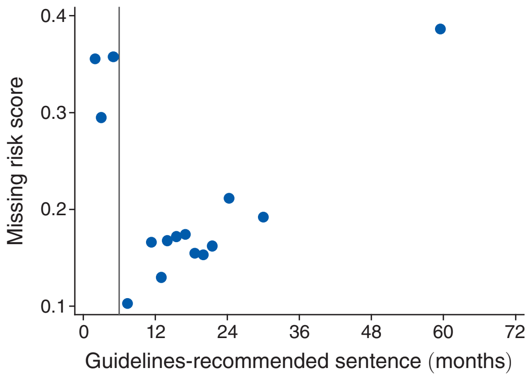

Figure 1. Relationship between Missing Risk Scores and the Guidelines-Recommended Sentence

Notes: This figure plots the likelihood of not receiving a risk assessment against the guidelines-recommended sentence for all risk-assessment-eligible defendants in fiscal years 2003 and 2004. The sample is divided into 15 equally sized bins; the per-bin average is shown on the vertical axis. Defendants to the left of the vertical line have a guidelines-recommended sentence of probation or jail; those to the right have the possibility of a prison recommendation.

One important thing to note is that the risk assessment was only conducted on 75 percent of eligible defendants. We don't consider the missing data a limitation, but rather, a first example of discretion in use. As shown in Figure 1, risk assessment information forms a V-shaped relationship with the guidelines-recommended sentence; it is most likely to be missing when the guidelines-recommended sentence is very short (under six months) or very long (more than three years). This suggests that factors associated with the guidelines-recommended sentence make risk assessment information more/less valuable. In the first scenario, criminal justice actors may already be in agreement about diverting the defendant, given the factors that yielded such a short guidelines-recommended sentence. In the second scenario, diversion may simply be unpalatable, given the factors that yielded such a long guidelines-recommended sentence. In either scenario, the risk score information would not be pivotal, since the decision was already made based on alternative considerations.

The only specifications in which we subset to defendants with a nonmissing risk score are those using regression discontinuity. The sample for those analyses therefore consists of people for whom the risk assessment was deemed worth calculating: disproportionately, those with a guidelines-recommended sentence between 6 and 30 months.

Finally, we show that Virginia's risk assessment has comparable accuracy to a more sophisticated instrument. As a benchmark, we build an alternative risk assessment that uses random forest to predict recidivism among a sample of more than 59,000 individuals. Our model uses all the input variables included in the actual risk score as well as dozens of additional input variables derived from sentencing worksheets and court records.[17] To maximize accuracy—and because we wanted

[17] This is discussed in more detail in online Appendix A2.

a race-weighted recidivism predictor to use as a control in one of our subsequent analyses—our self-built risk assessment includes prohibited factors such as race. The accuracy of Virginia's risk assessment, as measured by the area under the curve (an accuracy metric that is equivalent to the Wilcoxon rank-sum test [Mason and Graham 2002]), is statistically indistinguishable from ours.

II. Does the Algorithmic Tool Influence Sentencing Decisions?

We begin by evaluating whether judges changed sentencing practices after risk assessment was adopted. We use two methods: a regression discontinuity design to see if the low-risk label affects sentencing, and an event-study design to see if the correlation between sentencing and risk score increases after risk assessment adoption.

A. *Response to the Diversion Recommendation*

Since the low-risk classification is triggered by scoring below a cutoff in the risk score, we can use regression discontinuity to see whether the risk classification affects sentencing decisions. This research design shows the impact of the low-risk designation for those near the cutoff.

Our specification is shown in equation (1), where *riskscore* is the raw risk score, *high_risk* is an indicator for being above the cutoff, and *X_RD* includes the guidelines-recommended sentence, age, recent prior convictions, and offense. Observations i are at the case level. Our main regression discontinuity analysis uses all defendants who received a risk assessment during fiscal years 2003 and 2004—the first two years after risk assessment adoption. We focus on these years in part for consistency. Our event-study and difference-in-difference specifications use data from the two years before and two years after risk assessment adoption. But we also focus on these years because, as shown in Section IIC, they are the ones in which risk assessment appears most influential in sentencing.

$$(1) \quad Y_i = \alpha + riskscore_i \times \beta_1 + high\_risk_i \times \beta_2 + riskscore_i \times high\_risk_i \times \beta_3 + X\_RD_i \times \beta_4 + \epsilon_i .$$

The regression discontinuity design relies on the assumption that, with the exception of the risk classification, people who score right below the cutoff are very similar to people who score right above. This assumption is negated if the person who administers it strategically assigns points to ensure that someone is rated low or high risk. Such manipulation is often observable in an uneven bunching of scores right below or right above the cutoff or in discontinuous changes in personal characteristics such as age, criminal history, or race.

In online Appendix Figure A1, we show that there is no obvious bunching around the cutoffs that would indicate strategic manipulation of the score. However, the distribution of the risk scores is lumpy, which complicates formal distribution tests. The lumpiness is due to the fact that integer weights are used to calculate

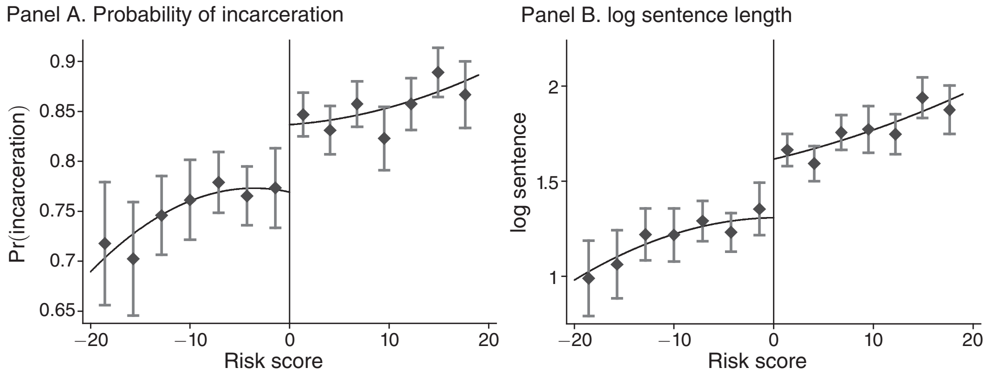

Figure 2. Do Judges Change Sentencing in Response to the Risk Assessment?

Notes: The vertical axes show the probability of incarceration and the log sentence (bottom coded at two weeks). The horizontal axes show the risk assessment, normalized so that defendants who score below zero receive a diversion recommendation. Each dot shows the mean for a bin of three risk scores, and the whiskers show the 95 percent confidence interval for that mean. The sample includes all individuals who received a risk assessment in 2003 and 2004.

the risk score; integers that are multiples of three are particularly common in the nonviolent risk score. Thus, we rely primarily on tests for discontinuities in defendant characteristics around the risk score cutoff to assuage concerns about strategic manipulation.

We test for discontinuous changes in predetermined characteristics using the optimal mean squared error bandwidth selection method proposed in Calonico, Cattaneo, and Titiunik (2014). Results are shown in panel A of Table 2 and in online Appendix Figures A1b–A1i. Out of the eight characteristics shown, only prior convictions is statistically significant.

We do find, however, that the cutoff delineating a change in risk classification corresponds to a discontinuous change in both the probability of incarceration and in the log sentence length. We use a log transform both to reduce the influence of extreme outliers and to allow coefficients to be interpreted approximately as a percent change in the outcome. To account for noncarceral sentences (sentence $= 0$), we bottom code the sentence length data at two weeks, which is the third percentile in the sentence distribution among eligible defendants sentenced to incarceration.[18] These results are shown graphically (Figure 2) and in formal estimates (panel B of Table 2), again using optimal bandwidth selection. Defendants above the nonviolent cutoff are 6 percentage points ($P < 0.05$) more likely to be incarcerated and receive sentences that are approximately 24 percent longer ($P < 0.05$). The inclusion of covariates (columns 2 and 4) doesn't affect the incarceration results but increases the magnitude and statistical significance of the results on sentence length. All together, these results demonstrate that judges do pay attention to the recommendations associated with the risk assessment, and adjust their sentences accordingly.

[18] Results are similar if instead we use an inverse hyperbolic sine transform or the nontransformed sentence length top coded at the ninety-ninth percentile, as shown in online Appendix Table A1.

Table 2—Regression Discontinuity around the Low-Risk Cutoff

	Female (1)	Black (2)	Age (3)	Priors (4)	Guidelines (5)	Drug (6)	Theft (7)	Fraud (8)
Panel A. Testing for covariate smoothness across the cutoff								
RD_Estimate	0.003	−0.026	−0.027	0.063	0.671	−0.009	0.016	−0.008
	(0.037)	(0.035)	(0.026)	(0.021)	(0.800)	(0.033)	(0.026)	(0.029)
Mean	0.276	0.533	0.255	0.138	13.232	0.497	0.293	0.210
Observations	8,810	8,808	9,348	9,404	9,404	9,404	9,404	9,404
Covariates	No	No	No	No	No	No	No	No

	Pr(incarceration)		log sentence	
	(1)	(2)	(3)	(4)
Panel B. Testing for sentencing changes across the cutoff				
RD_Estimate	0.066	0.060	0.197	0.241
	(0.026)	(0.031)	(0.110)	(0.112)
Mean	0.818	0.818	1.576	1.576
Observations	9,404	9,404	9,404	9,404
	No	Yes	No	Yes

Notes: This table uses regression discontinuity to test for changes around the low-risk cutoff. The risk score is the running variable. Panel A tests for discontinuous changes in defendant characteristics and panel B tests for discontinuous changes in the probability of incarceration and log sentence, bottom coded at two weeks. Means of the dependent variable are shown at the bottom. Optimal bandwidths are used. The sample includes all individuals who received a risk score in fiscal years 2003 and 2004.

B. *Event-Study Changes in Risk Responsiveness*

The gains that policy simulations predicted come from releasing low-risk defendants who otherwise would have been incarcerated or incarcerating high-risk defendants who otherwise would have gone free. One might expect that a judge who pays attention to the risk score for those near the cutoff might also pay attention to it for those closer to the tails of the risk distribution as well. But the regression discontinuity design does not provide any direct test of this. We therefore employ a secondary strategy: an event-study design to see if the correlation between risk and sentencing increases after risk assessment adoption.

Testing to see if sentencing changes for defendants at the tails of the risk score distribution would be straightforward if we had risk assessment information for those sentenced before risk assessment was adopted. Unfortunately, these data are not available. We do, however, have access to much of the same information that is used to calculate the risk score: age, gender (for most defendants), offense, and criminal history. It isn't perfect—the formatting and exact information differ somewhat. Also, we are missing marital and employment status. But, it allows us to generate an approximation of the risk score for those sentenced both before and after the reform.[19] We do so using a random forest model trained on all defendants with a risk score in fiscal years 2003–2006. We generate out-of-bag predictions (predictions

[19] Note: in this exercise, we are predicting Virginia's risk score, not building an alternative model predicting future offending.

in which that particular observation is not used to train the model) for the training sample, and out of sample predictions for all other eligible defendants sentenced between fiscal years 1999–2006. The correlation coefficient between the actual risk score and the predicted risk score is 0.80. While not a perfect prediction, the two are strongly correlated.[20]

We use an event-study design as an alternate method of evaluating how risk assessment use affected sentencing. The specification used can be found in equation (2), where $Predicted\_risk_i$ is the predicted risk score, normalized so that distance between the fifth and ninetieth percentile is equal to one;[21] $Post_i$ is an indicator for being sentenced during the years that the risk score was available; and X_i include the guidelines-recommended sentence, offense, age (in deciles), race, gender, recent prior charges, judge fixed effects, fixed effects for the month, year and weekday of sentencing, and indicators for five quintiles of the predicted risk score. Observations i are at the case level. The sample includes all risk-assessment-eligible defendants sentenced within two years before/after risk assessment adoption.[22]

$$Y_i = \alpha + Predicted\_risk_i \times Post_i \times \beta_0 + Predicted\_risk_i \times \beta_1 + Post_i \times \beta_2 + X_i \times \beta_3 + \epsilon_i. \tag{2}$$

Results are shown in Table 3. We see that even before risk assessment was adopted, judges gave more severe sentences to defendants with higher risk levels (second row coefficients). But this relationship became more pronounced after risk assessment became available (first row coefficients). In the years post-risk-assessment adoption, moving from the fifth to ninetieth percentile of the risk distribution led to an additional seven percentage point increase in the likelihood of incarceration and roughly 24 percent increase in the sentence length.[23] Note that these will be attenuated estimates due to measurement error in the predicted risk score.[24]

We also conduct a variation on the specification shown in equation (2) in which *Predicted_risk* refers to the quintiles of the predicted risk distribution, with the middle quintile omitted to serve as a reference point. The middle (third) quintile is also right below the cutoff in the risk score; the fourth quintile is right above it. Results are shown in Table 3, panel B. If risk assessment adoption only affected those right around the cutoff, one might see changes around the upper middle part

[20] As expected, there is no discontinuity in sentencing around the low-risk cutoff for the predicted risk score before risk assessment was adopted. However, the measurement error introduces enough smoothing that there is no longer a clearly visible discontinuity during the years that risk assessment was in use.

[21] Our goal is to show the impact of a high-risk score versus a low-risk score. Due to outliers at the tails, we think the fifth and ninetieth percentiles are more representative of the low and high ends than the first and ninety-ninth percentiles.

[22] The sample includes those who are risk assessment eligible in the postperiod, but for whom we do not have a risk assessment.

[23] Results are similar if instead we use an inverse hyperbolic sine transform or the nontransformed sentence length top coded at the ninety-ninth percentile, as shown in online Appendix A2.

[24] The random forest process generates some additional uncertainty that is not captured by the standard errors shown in Table 3. However, this uncertainty is small. We generate 100 different predicted risk scores and estimate equation (2) for each. The T statistic is above three for each specification.

Table 3—Event-Study Test for Increased Correlation between Sentence and Predicted Risk Score

	Pr(incarceration)		log sentence	
	(1)	(2)	(3)	(4)
Panel A. Event study with predicted risk entering linearly				
Predicted risk × post	0.061	0.069	0.163	0.239
	(0.023)	(0.021)	(0.080)	(0.064)
Predicted risk score	0.079	0.116	1.344	0.828
	(0.032)	(0.035)	(0.123)	(0.115)
Observations	24,040	24,040	24,040	24,040
R^2	0.016	0.093	0.041	0.357
Covariates	No	Yes	No	Yes
Panel B. Event study with predicted risk in quintiles				
First quintile × post	−0.031	−0.043	−0.029	−0.176
	(0.024)	(0.023)	(0.084)	(0.076)
Second quintile × post	0.002	−0.007	0.005	−0.036
	(0.019)	(0.019)	(0.063)	(0.052)
Fourth quintile × post	0.008	0.008	0.030	0.025
	(0.014)	(0.014)	(0.052)	(0.044)
Fifth quintile × post	0.027	0.025	0.099	0.088
	(0.015)	(0.014)	(0.055)	(0.044)
Observations	24,040	24,040	24,040	24,040
R^2	0.016	0.092	0.036	0.355
Covariates	No	Yes	No	Yes
Joint p-value	0.109	0.028	0.342	0.003

Notes: This table shows estimates of the extent to which the correlation between sentencing and defendant risk increased after risk assessment adoption. In panel A, we regress the probability of incarceration and the log sentence length (bottom coded at two weeks) on the predicted risk score interacted with a dummy for being post-risk-assessment adoption, controlling for the predicted risk score, predicted risk quintiles, and year dummies. In panel B, the predicted risk quintiles are interacted with a dummy for being post-risk-assessment adoption, controlling again for predicted risk, predicted risk quintiles, and year dummies. Standard errors are clustered at the judge level. The sample includes all risk-assessment-eligible defendants sentenced between 2001 and 2004. In panel B, the joint p-value is from an F test of the null hypothesis that the true value of each coefficient shown is equal to zero.

of the distribution but no changes out in the tails. Instead, we see large negative coefficients for the smallest quintile and positive coefficients for the largest quintile. The shift in sentencing appears to occur throughout the risk distribution and not just around the margins of the risk score cutoff.

The event-study research design assumes that any shift in the relationship between the predicted risk score and sentencing that occurs post-risk-assessment adoption can be attributed to risk assessment use. Such an assumption would be undermined if there were either preexisting trends or some sort of concurrent change that could confound the results. We can evaluate the existence of preexisting trends using a dynamic event-study graph. For this exercise, our sample includes eligible defendants convicted within four years before/after risk assessment adoption. The specification entails replacing *Post* with dummies for each year of sentencing and omitting the year prior to adoption to serve as a reference point. The coefficients for *Predicted_risk* interacted with each year are shown in Figure 3, panels A and B. We see no evidence of preexisting trends: the relationship between risk and sentencing

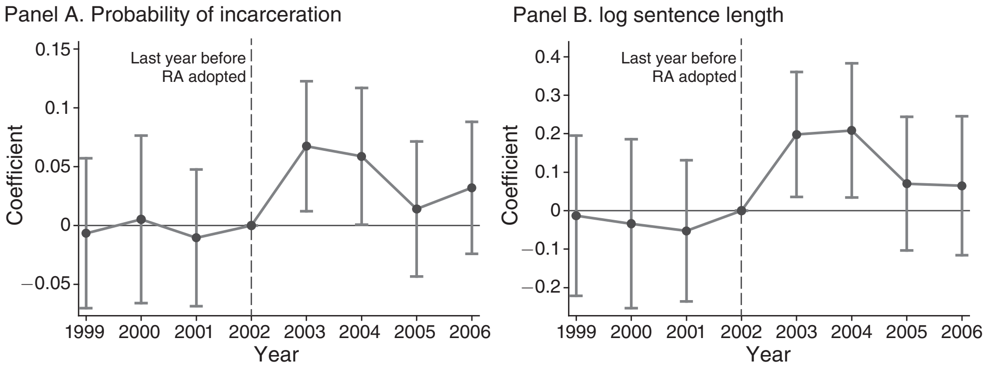

Figure 3. Do Judges Change Sentencing in Response to the Risk Assessment?

Notes: The figures show a dynamic event study in which sentencing outcomes are regressed on year dummies, the predicted risk score, and interactions thereof. The coefficient on the interactions are shown as dots on the graph; the whiskers show the 95 percent confidence intervals. The year before risk assessment adoption is omitted to serve as a reference. The sample includes all risk-assessment-eligible individuals sentenced in fiscal years 1999–2006.

is constant from 1999–2002. However, we see a sharp increase in the relationship between risk and sentencing beginning the year risk assessment was adopted.

Another concern is that risk assessment use brought about a change in policing or prosecution practices, thereby changing the composition of cases that are risk assessment eligible. An example of such a change would be a decrease in the willingness of prosecutors to charge someone with an eligible offense in order to avoid diversion. Such a change in practices could be visible in a change in defendant or case characteristics, or in the frequency of risk-assessment-eligible cases. By comparing specifications with and without the inclusion of controls, we've already provided some initial evidence on this issue. The estimates of β_0 that include controls are either the same or larger than estimates from specifications without controls, suggesting that any change in case composition would likely serve to understate the effect.

As additional evidence, panel A of online Appendix Table A3 shows tests for changes in defendant race, gender, age, recent prior convictions, guidelines-recommended sentence, predicted risk score, and offense category after risk assessment is adopted. This table shows estimates from a simple regression of the covariate on a dummy for being post risk assessment adoption. We also test for changes in case frequency by collapsing the data to the weekly level and regressing the weekly number of risk-assessment-eligible cases on a dummy for being sentenced post-risk assessment adoption. Out of the 11 covariate tests conducted, one is statistically significant at the 5 percent level. We conduct a graphical event-study analysis to evaluate whether a change in case characteristics was coincident with risk assessment adoption. We regress each of these covariates on biyearly dummies for fiscal years 2001–2004, with the year before risk assessment adoption omitted as a baseline. These graphs are shown in online Appendix Figure A2. There is no evidence of a change in characteristics coincident with the adoption of risk assessment.

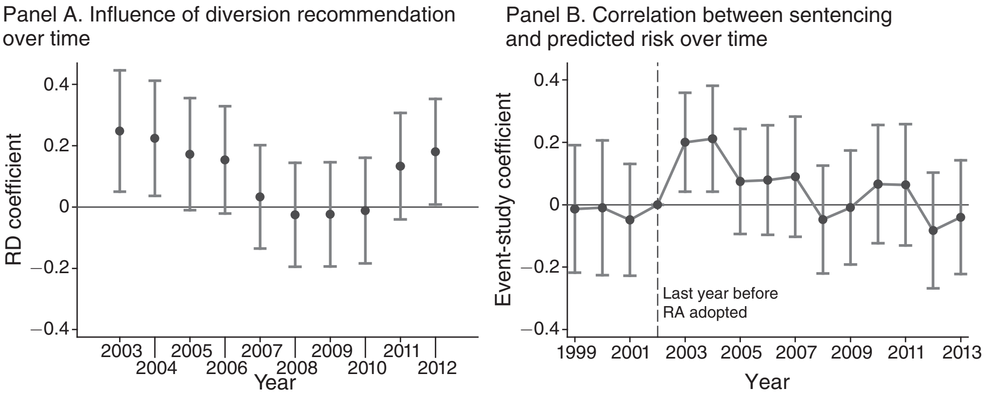

FIGURE 4. RISK ASSESSMENT USE OVER TIME

Notes: The left-hand figure is a coefficient plot designed to show how sentencing responds to the diversion recommendation in different time periods. Each dot represents an regression discontinuity estimate of the magnitude of sentencing discontinuity at the low-risk cutoff for all cases who received a risk score during a rolling two year period. The first year of each two year period is shown on the horizontal axis. The outcome is the log sentence length (bottom coded at two weeks). The right-hand figure shows a long-term event-study analysis of the correlation between sentencing and predicted risk. Each dot shows the coefficient on the predicted risk score interacted with year, derived from a regression with the log sentence length (bottom coded at two weeks) as an outcome and with controls for each quintile in predicted risk as well as year dummies and other covariates. The sample includes all risk-assessment-eligible defendants sentenced between 1999 and 2013. Our analyses terminate in 2013 since a new risk assessment was adopted after that year.

C. *Use over Time*

Finally, we test to see whether judges' use of the risk assessment changes over time. Our first metric for evaluating the intensity of use over time is the magnitude of the sentencing discontinuity around the high-risk cutoff. Specifically, we conduct regression discontinuity estimates as described in equation (1) on rolling two-year samples: first fiscal years 2003 and 2004, then fiscal years 2004 and 2005, and so forth.[25] Figure 4, panel A, shows results. Discontinuities are greatest in the first few years of risk assessment use. By fiscal year 2007, the discontinuity has completely vanished, although there is some evidence of an uptick in fiscal years 2012 and 2013, right before the new risk assessment was adopted.

The regression discontinuity design for evaluating use over time has both strengths and weaknesses, as discussed in the previous section. To provide a second piece of evidence, we extend our event-study design so that we can look at the correlation between sentencing and predicted risk over a longer time span. Figure 4, panel B is the same specification as Figure 3, panel B, but it includes the years up to 2013. The effect of the risk assessment appears strongest in the two years after adoption. The coefficient on predicted risk interacted with year is statistically insignificant and bounces around zero from fiscal years 2005–2013.

[25] There was also a slight drift upward in the fraction of risk-assessment-eligible cases where risk assessment wasn't conducted, ranging from 24–29 percent.

These two analyses suggest that use of risk assessment declined over time. Judges tried it but appear to have lost interest. There are several potential explanations for this. First, the algorithm may not have provided very useful information. Judges abandoned it because it did not provide value. Alternatively, this could be an example of algorithm aversion, or a reluctance by human decision-makers to use superior but imperfect algorithms (Burton, Stein, and Jensen 2020). Experimental evidence shows that people are quick to abandon algorithmic forecasts after seeing it make a mistake, even if the algorithm is, on average, more accurate than the human (Dietvorst, Simmons, and Massey 2015).

III. The Role of Discretion

We've shown that judges did use the risk scores—at least at the beginning. In this section, we evaluate how discretion altered risk assessment's impact. We first look at the net effect of risk assessment on sentencing, benchmarked against a simulation of risk assessment's impact in the absence of discretion. We then provide some descriptive information about how discretion is used: which demographic factors predict deviation from the risk score? Finally, we evaluate whether risk assessment had differential impact by race or age, again benchmarking against simulations of sentencing by algorithm.

A. *The Net Impact on Sentencing*

We begin by evaluating the net impact that risk assessment had on sentencing. Our primary method of analysis is a difference-in-difference design that compares outcomes in the two years before/after the adoption of risk assessment for defendants who were/were not eligible for risk assessment.

We benchmark the real-world impact of risk assessment against a simulation of how risk assessment would have affected outcomes if judges were not granted discretion to deviate from the sentencing recommendations associated with the algorithm. Following Virginia policy, we define sentencing by algorithm as automatic diversion for low-risk defendants. Specifically, we generate simulated sentences in which low-risk defendants are assigned a sentence of length zero (not incarcerated) if their guidelines-recommended sentence was jail. Low-risk defendants with a guidelines-recommended sentence of prison are assigned a six month jail sentence. Everyone else's sentence remains the same.

Note that Virginia does not have a formal policy for how a high-risk score should affect sentencing. Thus, sentencing by algorithm in the Virginia context is partly nondiscretionary and partly discretionary. In robustness tests described below, we use simulations that impose fully nondiscretionary definitions of sentencing by algorithm. We also explore alternative definitions of diversion. Results are largely similar.

Our specification is shown in equation (3), in which each observation i consists of a single case. *Post* refers to the period after risk assessment is adopted, *Eligible* refers to risk-assessment-eligible cases, and covariates (X) are as defined above. We cluster standard errors at the level of the judge, but results

Table 4—Risk Assessment's Net Impact on Sentencing: Simulated versus Actual

	Pr(incarceration)			log sentence		
	Simulated	Actual		Simulated	Actual	
	(1)	(2)	(3)	(4)	(5)	(6)
Eligible × post	−0.073	0.005	0.010	−0.210	0.021	0.041
	(0.008)	(0.010)	(0.008)	(0.024)	(0.035)	(0.023)
Observations	64,675	64,675	64,675	64,675	64,675	64,675
R^2	0.423	0.047	0.432	0.627	0.003	0.626
Mean DV	0.761	0.791	0.791	1.399	1.399	1.399
Covariates	Yes	No	Yes	Yes	No	Yes
p-value, actual vs simulated			0.000			0.000

Notes: This table presents difference-in-difference estimates in which outcomes are compared across eligible/ineligible cases before/after risk assessment is adopted. The outcomes are the probability of incarceration and the log sentence (bottom coded at two weeks), as well as simulations thereof. Simulated sentences (columns 1 and 4) entail automatic diversion for low-risk individuals after risk assessment was adopted. Standard errors are clustered at the judge level. The mean dependent variables for eligible cases during the pre-risk-assessment period are shown in the bottom row. The p-value pertains to the difference between the coefficients in actual (with covariates) and simulated using seemingly unrelated regression. The sample includes all defendants sentenced in fiscal years 2001–2004.

are robust to clustering at the judicial circuit level as well (see online Appendix Table A4).[26] We supplement the regression analysis with a dynamic graphical approach. This entails expanding the sample to fiscal years 1999–2006 and substituting *Post* with dummies for each fiscal year, with the year before adoption omitted as a reference.

$$(3) \quad Y_i = \alpha + Eligible_i \times Post_i \beta_0 + Eligible_i \times \beta_1 + Post_i \times \beta_2 + X_i \times \beta_3 + \epsilon_i.$$

Our primary sample includes all defendants sentenced in fiscal years 2001–2004.[27] Our outcome measures are the probability of incarceration and the log sentence, bottom coded at two weeks. For each, we show both the simulated sentences described above and the actual sentences, which judges determine with input from the algorithm. The estimates are shown in Table 4 and the graphical analysis is shown in Figure 5.

Sentencing by algorithm would have led to a sharp decrease in both the probability of incarceration and the sentence length. This is no surprise: the algorithm was implemented with the goal of reducing prison populations and the sentencing recommendations associated with the algorithm worked as a one-way ratchet to achieve this. In reality, however, risk assessment adoption had no detectable impact on the probability of incarceration. We can rule out more than a 1 percentage point decline at the 5 percent level. If anything, it may have led to a slight increase in the

[26] Judges vary in the extent to which they use the risk assessment tool, meaning that defendants seen by different judges have varying levels of treatment. Just as a medical trial that assigned varying intensities of treatment to different clusters of people (e.g., different hospitals) would cluster at the level of treatment assignment, we cluster at the level of judge assignment. With the exception of large counties in which there are multiple judges, this is effectively clustering at the county level.

[27] We do not drop those with missing risk scores as that could bias the sample.

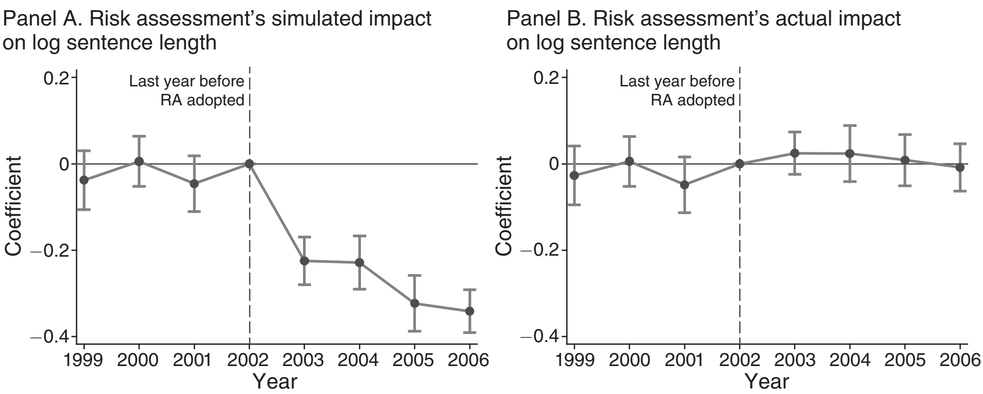

Figure 5. Risk Assessment's Impact on Sentencing: Simulated versus Actual

Notes: These figures show dynamic difference-in-difference estimates of the impact of risk assessment on sentencing. The outcomes are the the log sentence (bottom coded at two weeks) as well as simulations thereof. Simulated sentences entail automatic diversion for low-risk individuals after risk assessment was adopted. In each graph, the year prior to risk assessment adoption is omitted as the baseline. The sample includes all individuals sentenced between 1999 and 2006. Standard errors are clustered at the judge level; 95 percent confidence intervals are shown.

length of sentences: a roughly 4 percent relative increase, statistically significant at the 10 percent level.

A causal interpretation of the difference-in-difference estimates relies on the assumption that, absent risk assessment adoption, outcome trends for eligible and ineligible groups would remain parallel. This assumption cannot be directly tested, but indirect evidence supports it. The dynamic graphical analysis shows that trends were more or less parallel before the reform. The estimates remain stable to the inclusion of covariates, suggesting that if there was a change in the composition of eligible cases, it was mirrored by a change in the composition of ineligible cases. The previous subsection provided some initial evidence that prosecutors did not meaningfully change the practices by which they charge cases in response to risk assessment adoption. In panel B of online Appendix Table A3, we provide further evidence to support the identifying assumptions. Using the difference-in-difference specification shown in equation (3), but without the controls X_i, we test for differential changes in defendant race, gender, age, prior convictions, the likelihood of being recommended for prison, the guidelines-recommended sentence in months, and the weekly number of cases filed. We find no evidence that case composition or frequency changed differentially for eligible versus ineligible defendants after risk assessment adoption, supporting an argument that trends in outcomes would have remained parallel absent the reform.

We conduct a variety of tests to ensure that our estimates are robust. As discussed previously, our results are robust to clustering at the judicial circuit level (online Appendix Table A4). They are also robust to varying the sample time window. In our main specification, we used four years of data. Results are generally similar using two, six, and eight year samples as well (online Appendix Table A5). We then show that our results are robust to different ways of defining the control groups: drug, larceny, and fraud offenders who were risk assessment ineligible; all offenses except drug, larceny, and fraud offenders who were risk assessment ineligible; and all nonviolent offense categories (online Appendix Table A6). We also show robustness to various ways of defining simulated sentences: a 0 or 12 month sentence for low-risk individuals with a guidelines-recommended sentence of prison and replacing sentences for high-risk individuals with the average sentence judges gave to people with the same risk score and guidelines-recommended sentence in fiscal years 2003–2006 (online Appendix Table A7). Finally, we show robustness to alternative transformations of the sentence length: inverse hyperbolic sine and the sentence top coded at the ninety-ninth percentile of positive sentences (online Appendix Table A8).

Thus, it appears that risk assessment adoption did not have the impact on sentencing that it was designed to have. It was adopted specifically to lower incarceration rates by diverting low-risk offenders from jail or prison. Its incorporation into the sentence guidelines was unidirectional, nudging judges toward leniency for low-risk offenders while remaining silent about high-risk offenders. Section II showed us that judges did use the risk assessment tool. But apparently judges responded as much to the absence of the diversion recommendation as they did to the recommendation itself. Even though judges increased leniency for those at the lower end of the risk spectrum, they increased severity for those at the higher end. While there was a change in who got incarcerated, there was no net change in the incarceration rate or length of sentence.

B. *When Do Judges Deviate from the Risk Score?*

Even though the risk score is influential, judicial discretion still plays a large role. Judges diverted 25 percent of high-risk defendants against recommendations, and they failed to divert low-risk defendants 63 percent of the time.[28] Understanding the real world impacts of risk assessment adoption thus hinges on understanding when and why judges make these discretionary choices.

Understanding its impacts requires understanding when and why judges deviate from the recommendations associated with the algorithm.

In this subsection, we present descriptive information about how judicial discretion interacts with the risk score. Our strategy is to see when judges deviate upward or downward in sentencing for people with the same risk score but different race, gender, age, marital, or employment status. Our baseline specification entails regressing the log sentence length on each demographic factor, controlling for the exact risk score. Our second specification adds controls for a fully saturated set of fixed effects for the guidelines-recommended sentence, meaning that this specification accounts for all sentencing nudges given by the state to the judge. Our third specification adds judge fixed effects, which eliminates concerns about correlations between judge tendencies and demographic groups (e.g., if largely White jurisdictions have judges that are particularly strict). Our fourth

[28] Research by Garrett, Jakubow, and Monahan (2019) suggests that judges are more likely to divert low-risk defendants if some sort of alternative sanction, such as mental health treatment, is available in that county.

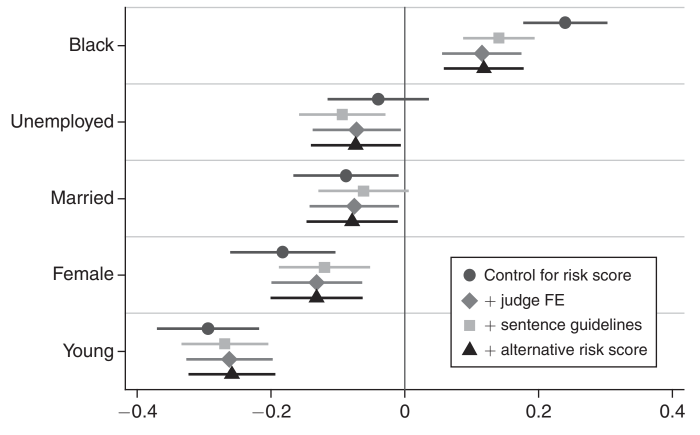

Figure 6. What Factors Predict Deviation from the Risk Score?

Notes: This coefficient plot shows estimates from regressions of log sentence length on demographic factors. The horizontal line shows the 95 percent confidence interval. Each specification, delineated by shade and marker symbol, includes all five demographic variables. The topmost specification controls only for the risk score. The second specification adds controls for judge fixed effects, the third adds the exact guidelines-recommended sentence, and the fourth adds controls for a race-weighted risk score. The sample includes all individuals who received a risk score in fiscal years 2003 and 2004, with the exception of those missing the demographic information listed.

specification is discussed in more detail below. A coefficient plot of results is shown in Figure 6.[29]

Several things stand out from this analysis. First, judges give young defendants sentences that are almost 30 percent shorter than those given to older defendants at the same risk level. (Young is defined as being under the age of 23.) Second, judges give Black defendants sentences that are 12–24 percent longer than White defendants with the same risk score. Finally, defendants who are female, unemployed, or married receive favorable treatment (4–18 percent shorter sentences). These patterns hold even when controls for the guidelines-recommended sentence and judge are included.

Of the five characteristics evaluated, race is the only one not explicitly included in the risk score. One possible explanation for the longer sentences given to Black defendants is that judges believe—accurately or not—that they pose a higher risk of reoffending than is conveyed by their risk score. To provide some suggestive evidence on this, our fourth specification includes controls for our self-built risk score, described in Section IC and in online Appendix A2.[30] This risk assessment includes race as a predictor in addition to a large number of other factors, including all those listed in Figure 6. If judges deviated upward for Black defendants because they actually had a higher risk of recidivism than others with the same risk score, controlling for the race-weighted recidivism risk should reduce or eliminate the upward deviation. Instead, we find that this control has virtually no effect on the magnitude of the coefficient.

[29] Results are shown in table format in online Appendix Table A9.
[30] We control for this alternative risk score using dummies for the ten deciles of estimated risk.

This analysis suggests that judges are deviating away from the risk score for reasons other than actual risk of recidivism. It could be because judges are overestimating the risk of recidivism for Black defendants. Or, it could be because other factors are influencing the decision: racial bias, political concerns, etc. Regardless of the reason, these results highlight the importance of judicial discretion when considering the racial impacts of an algorithm. An algorithm designed to reduce bias in the scores will not necessarily reduce bias in use.

Alternative objectives could also explain why judges systematically deviate downward in sentencing for young defendants. Age is one of the best predictors of criminal activity. Youth is associated with an additional 13 points on the risk score. For context, a defendant would receive fewer points for five prior incarcerations. If the goal at sentencing is to incarcerate those at highest risk of reoffending, young people would be incarcerated at a very high rate, but incapacitation is not the only goal at sentencing. Retributivists, who believe that punishment is based on desert, often argue that young people are less culpable because they are more impulsive and lack of self-control (Stevenson and Slobogin 2018). Furthermore, they are in a phase where they are particularly susceptible to peer influence. Incarceration could expose them to negative influence, thus increasing recidivism in the long run (Bayer, Hjalmarsson, and Pozen 2009; Aizer and Doyle Jr. 2015; Stevenson 2017).

Our hypotheses are exploratory, but the regression results demonstrate that deviation from the risk score is nonrandom. Judges are choosing to deviate upward/downward in sentencing because other factors are at play in their decision, with disparate impact on different demographic groups.

C. *Did Risk Assessment Use Affect Race/Age Disparities?*

We next evaluate whether risk assessment adoption affected defendants differently by race or age.[31] As previously, we benchmark the actual impact against the simulated impact of sentencing by algorithm.

Risk assessment's potential for disparate impact by race and age is important for several reasons. Race has been a central theme of recent debates about risk assessment, with many expressing concerns that its use will negatively impact Black defendants. Understanding the racial impact will help policymakers who are concerned about racial disparities in criminal justice. Risk assessment's impact on young people has received less attention, despite the fact that there are strong political movements to reduce young people's exposure to incarceration. This may be partially because people are unaware of how much weight age receives in the risk scores (Stevenson and Slobogin 2018). Risk assessment's impact on the young is also interesting as an example of conflicting objectives in algorithm adoption.

Virginia's risk assessment embeds both race and age disparities. Although race is not explicitly included in the algorithm, many of its input factors correlate with race. As a result, Black defendants are 10 percentage points less likely to be categorized as

[31] While we would be interested to see if risk assessment had differential effects by employment and marital status, we do not have this information from before risk assessment was adopted, and our sample of women is too small to evaluate differential impact by gender.

low risk than White defendants with the same guidelines-recommended sentence. Age is heavily weighted in the risk assessment and young defendants are 16 percentage points less likely to be categorized as low risk than older defendants with the same guidelines-recommended sentence.

Judicial discretion, however, pulls in opposite ways across race and age. As shown in the previous subsection, judicial discretion added on to the racial disparities in the algorithm, giving Black defendants longer sentences than White defendants with the same risk score. In contrast, judicial discretion attenuated the age disparities, giving young people shorter sentences than they gave older defendants at the same risk level.

However, risk assessment's impact on race and age disparities depends on more than disparities in the instrument and disparities in use. It also depends on how judges sentenced before risk assessment adoption. If judges were already sentencing in a racially disparate manner, for instance, risk assessment use could actually reduce these disparities. In the presence of statistical discrimination, changing the amount of information available can have counterintuitive effects. In the hiring context, scholars have found that removing information about the criminal record lowers hiring rates for Black applicants since it increases the likelihood that employers will use race as a proxy for criminal history (Doleac and Hansen 2020; Agan and Starr 2018). Since the risk assessment increases the amount of information available to the judge, it could lower statistical discrimination and prove net beneficial to Black defendants.

Theory provides no clear answers about risk assessment's expected impact on racial disparities. The little empirical research currently available in the pretrial context has documented an increase in race disparities after risk assessment was adopted, although this was largely due to regional variation in take-up (Stevenson 2019; Albright 2023). A vignette study conducted on a sample of practicing judges suggests that risk assessment use would increase socioeconomic disparities in sentencing, potentially by triggering negative stereotypes about poor people (Skeem, Scurich, and Monahan 2020). We are unaware of any prior empirical work on risk assessment's impact on age disparities.

We test whether risk assessment affects defendants differently by race or age using a triple-differences research design, as shown in equation (4). Depending on the specification, Z is an indicator for being Black or being under the age of 23. The rest of the variables are as described above. The sample includes all defendants sentenced in fiscal years 2001–2004, with the exception of those missing race information (for the racial disparities specification) or age information (for the age disparities specification).

$$(4) \quad Y_i = \alpha + Post \times Eligible_i \times Z_i \times \beta_0 + Post_i \times Eligible_i \times \beta_1 + Eligible_i \times Z_i \times \beta_2 + Post_i \times Z_i \times \beta_3 + Eligible_i \times \beta_4 + Post_i \times \beta_5 + Z_i \times \beta_6 + X_i \times \beta_7 + \epsilon_i .$$

Table 5—Risk Assessment's Impact by Race and Age: Simulated versus Actual

	Pr(incarceration)		log sentence	
	Simulated (1)	Actual (2)	Simulated (3)	Actual (4)
Panel A. Impact by race				
Eligible × post × Black	−0.002	−0.014	0.076	0.043
	(0.016)	(0.016)	(0.047)	(0.047)
Eligible × post	−0.068	0.019	−0.245	0.019
	(0.012)	(0.012)	(0.034)	(0.033)
Observations	60,186	60,186	60,186	60,186
R^2	0.425	0.435	0.629	0.627
Mean DV, Black	0.798	0.819	1.552	1.552
Covariates	Yes	Yes	Yes	Yes
p-value on difference		0.195		0.084
Panel B. Impact by age				
Eligible × post × young	0.076	0.042	0.238	0.118
	(0.018)	(0.017)	(0.052)	(0.052)
Eligible × post	−0.092	−0.001	−0.268	0.014
	(0.009)	(0.008)	(0.024)	(0.023)
Observations	63,853	63,853	63,853	63,853
R^2	0.423	0.432	0.627	0.625
Mean DV, young	0.705	0.731	1.264	1.264
Covariates	Yes	Yes	Yes	Yes
p-value on difference		0.000		0.000

Notes: This table shows triple difference estimates of risk assessment's impact by race and age. The outcomes are the probability of incarceration and the log sentence (bottom coded at two weeks), as well as simulations thereof. Simulated sentences (odd-numbered columns) entail automatic diversion for low-risk individuals after risk assessment was adopted. In addition to the standard set of covariates, each specification includes an indicator for being post-2003, for being risk-assessment-eligible, for being Black/young (respectively), as well as for interactions between all of the above. Standard errors are clustered at the judge level. The mean dependent variables for eligible cases during the pre-risk-assessment period are shown in the bottom row. The *p*-value pertains to the difference between the coefficients in actual and simulated using seemingly unrelated regression. The sample includes all individuals sentenced in fiscal years 2001–2004, with the exception of those for whom race/age information is missing (for the race/age specifications, respectively).

We also conduct a dynamic graphical analysis of the triple-differences specifications, with the sentence length as the outcome. In particular, we substitute *Post* in equation (4) with a set of lag/lead dummy variables in order to get a better sense of the timing of the changes. The age specification covers fiscal years 1999–2006; in the race specification we are only able to document trends from 2000 to 2006. Results are shown in Table 5 and Figure 7.

We find little evidence that sentencing by algorithm would have increased racial disparities. The point estimate on the log sentence is moderately sized, suggesting a 7.6 percent increase, but it's not statistically significant, nor is the coefficient on the probability of incarceration. We find no evidence that risk assessment increased racial disparities in practice, although the standard errors are too large to rule out moderate sized increases.

In contrast, sentencing by algorithm would have led to a large increase in age disparities: a 7.6 percentage point relative increase in the probability of incarceration, and

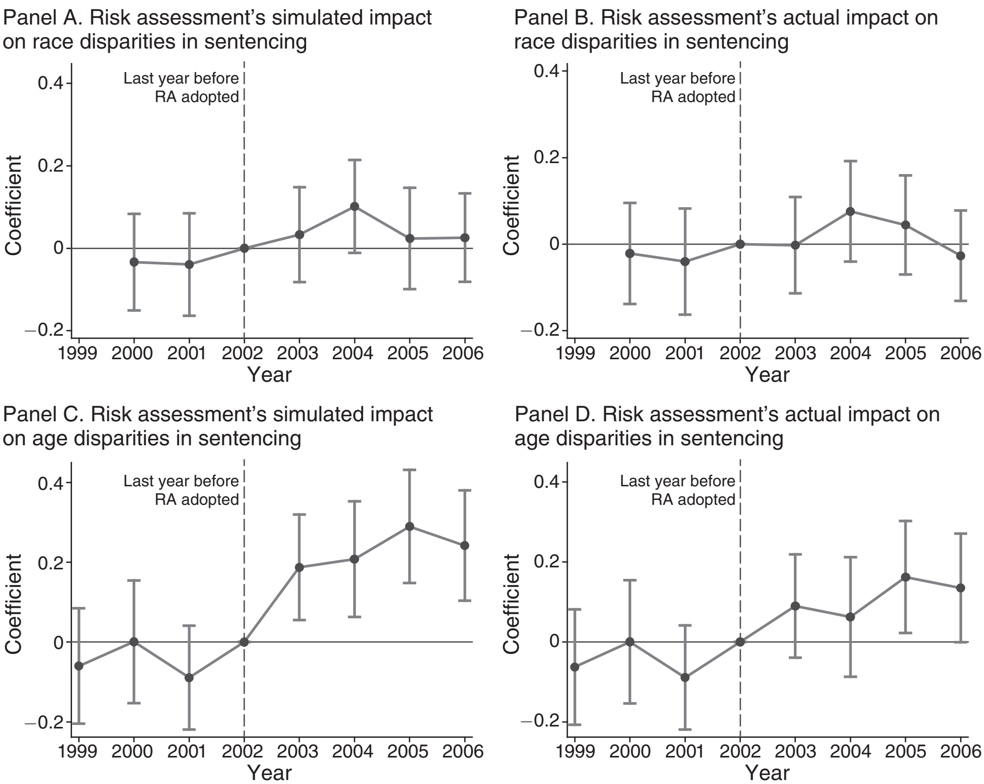

Figure 7. Risk Assessment's Impact on Race/Age Disparities: Simulated versus Actual

Notes: These figures show dynamic triple-differences estimates of the impact risk assessment has on race and age disparities in sentencing. The outcomes are the the log sentence (bottom coded at two weeks) as well as simulations thereof. Simulated sentences entail automatic diversion for low-risk individuals after risk assessment was adopted. In each graph, the year prior to risk assessment adoption is omitted as the baseline. The race disparities sample includes all individuals sentenced between 2000 and 2006, with the exception of those missing race information (race information is not available at all in 1999). The age disparities sample includes all individuals sentenced between 1999 and 2006, with the exception of those missing age information. Standard errors are clustered at the judge level; 95 percent confidence intervals are shown.

a 24 percent relative increase in the sentence length. Judicial discretion appears to have attenuated this somewhat. In actuality, risk assessment use led to a 4 percentage point relative increase in the likelihood of incarceration and a 12 percent increase in the sentence length. For those who see young age as a mitigating factor, such an increase in age disparities is an adverse consequence of risk assessment adoption.

We conduct robustness tests for the race/age disparity results similar to those shown for the main difference-in-difference analysis, shown in online Appendix Tables A10–A13. Both the age and race results are robust to varying the sample time window as well as varying the control groups. We also conduct an additional robustness tests for the age results, in which young is defined as under 30 (online Appendix Table A14). The results remain statistically significant and qualitatively similar. Likewise, we provide robustness tests for the simulation results similar to those discussed in Section IVA and in online Appendix Tables A15 and A16. The impact on race disparities depends on the simulation. The impact on age disparities is highly robust across simulations. Finally, in online Appendix Table A17, we show robustness to alternative transforms of the sentence length: inverse hyperbolic sine and the sentence length top coded at the ninety-ninth percentile of positive sentences for eligible defendants.

IV. Risk Assessment's Impact on Recidivism

Finally, we consider risk assessment's actual and simulated impact on recidivism. As discussed previously, our main recidivism measure is the likelihood of a new felony conviction for a crime committed within three years of sentencing. Generating simulated recidivism measures is a bit more complicated, as it requires assumptions about counterfactual crime rates. Absent a strong natural experiment, there is no great way to identify the likelihood that an incarcerated person would have committed crime if released. This is the same fundamental problem that complicates the development of risk assessments. Developers must generate estimates of the probability of committing crime if released using data on people who actually were released. The individuals used in the training data may differ in important ways from those who eventually receive a risk score.

Virginia's method for constructing the risk scores is to train a model to predict recidivism in the three years after release. The implicit assumption is that postrelease recidivism provides useful information about what recidivism would have been, had the person been immediately released. We follow a similar approach here. Namely, for individual i who would have gained an additional Y_i months of freedom under sentencing by algorithm, we estimate the additional recidivism risk during those months based on postrelease recidivism rates of individuals with the same risk score.

Of course, this method is imperfect. Among other things, released defendants are older and therefore have a lower statistical likelihood of crime than they did when they were younger. The hope is that it nonetheless provides a good enough approximation for use in simulations. At the very least, these simulations should be considered similar to the predictions a policymaker might have made if they had relied on the assumptions used in developing Virginia's risk scores.

Our specific method for building simulated recidivism variables is as follows:

(i) For each individual, we build a set of variables *recidivism_X* that are equal to 1 if that person has a new felony conviction for an offense committed within X months of release ($X \in 1,2,3, \ldots ,36$).

(ii) We regress each variable *recidivism_X* on dummies for risk score and fiscal year and generate a set of predictions $\widehat{recidivism\_X}$.

(iii) For each individual, we estimate the number of months of additional freedom Y_i they would gain under the sentencing-by-algorithm schema described above. We round to the nearest month for sentences that are not already delineated in round numbers (most are).

TABLE 6—RISK ASSESSMENT'S IMPACT ON RECIDIVISM: SIMULATED VERSUS ACTUAL

	Recidivism		
	Simulated	Actual	
	(1)	(2)	(3)
Eligible × post	0.016	0.007	0.008
	(0.005)	(0.006)	(0.005)
Observations	64,675	64,675	64,675
R^2	0.044	0.006	0.044
Mean DV	0.192	0.192	0.192
Covariates	Yes	No	Yes
p-value on difference			0.000

Notes: This table presents difference-in-difference estimates in which outcomes are compared across eligible/ineligible cases before/after risk assessment is adopted. The outcomes are both simulated and actual recidivism levels. Actual recidivism is defined as a new felony conviction for a crime committed within three years of sentencing. Simulated recidivism is an estimate of what recidivism rates would have been if judges had not been granted discretion to deviate from the sentencing recommendations associated with the risk score. The increased recidivism risk due to increased time on release is estimated using the postrelease behaviors of individuals with the same risk score. Standard errors are clustered at the judge level. The mean dependent variables for eligible cases during the prerisk assessment period are shown in the bottom row. The p-value pertains to the difference between the coefficients in actual (with covariates) and simulated using seemingly unrelated regression. The sample includes all defendants sentenced in fiscal years 2001–2004.

(vi) We generate a variable *simulated_recidivism* that is:

- The same as actual recidivism for high-risk individuals and ineligible individuals,
- Equal to 1 for low-risk individuals that recidivated in real life,
- Equal to $\widehat{recidivism\_X}$ for low-risk individuals who did not recidivate in real life, but would have had $Y_i = X$ additional months of freedom under sentencing by algorithm.

We then estimate simulated and actual impacts on recidivism using the same difference-in-difference specification shown in equation (3). Results are shown in Table 6 and Figure 8.

Our simulations suggest that sentencing by algorithm would have led to a 1.6 percentage point increase in the likelihood of a new felony conviction within three years, over a mean of 19 percent. The increase in recidivism is because Virginia's policy was a one-way ratchet toward lower sentences. With more people released there is a mechanical increase in the likelihood of new felony convictions, even if those people are relatively low risk.[32]

In real life, risk assessment had no detectable impact on overall incarceration rates. There was both an increase in releases for those with lower risk scores and a decrease in releases for those with higher risk scores. In theory, this should have led to lower recidivism rates, since those at higher risk of committing crime were incapacitated

[32] Crime is also rampant in jails and prisons, but this rarely results in new criminal charges.

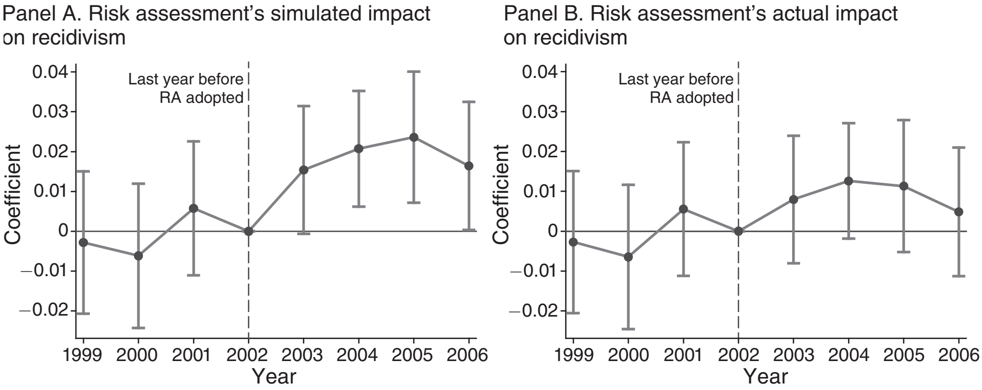

Figure 8. Risk Assessment's Impact on Recidivism: Simulated versus Actual

Notes: These figures show dynamic difference-in-difference estimates of the impact of risk assessment on recidivism. Actual recidivism is measured as the likelihood of a new felony conviction for a crime committed within three years of sentencing. Simulated recidivism measures allow for increased recidivism risk for those who would have had increased time on release under an automatic diversion system. Recidivism risk is estimated using postrelease behaviors of people with the same risk score. In each graph, the year prior to risk assessment adoption is omitted as the baseline. The sample includes all individuals sentenced between 1999 and 2006. Standard errors are clustered at the judge level; 95 percent confidence intervals are shown.

by incarceration. Unfortunately, there is no evidence that this happened. The point estimates are small and allow us to reject more than a 0.02 percentage point decrease in recidivism at the 5 percent level.[33]

We use the same set of robustness tests as described above for the difference-in-difference estimates of the impact on sentencing: varying the time window, varying the control group, and clustering at the judicial circuit level. These results are shown in online Appendix Tables A4–A6. We also show difference-in-difference estimates with alternative measures of recidivism (online Appendix Table A18): a new felony conviction or a new felony charge for an offense committed within six months, one year, three years, five years, and seven years postsentencing, as well as three years postrelease, and new charges for offenses committed within those same six time windows. In most specifications, the point estimates are small and statistically insignificant. If anything, the evidence points toward a small increase in recidivism, particularly in the postrelease period.

Why didn't risk assessment use lead to a detectable decrease in crime? We expect that this is partly because conflicting objectives reduced the potential gains from risk assessment. We have shown that i) risk assessments aren't even calculated for some cases, ii) judges frequently deviate from the sentencing recommendations associated with the risk assessment, and iii) deviation is strongly associated with defendant characteristics. Taken comprehensively, this suggests that judges are considering other factors than the risk of reoffending in the sentencing decision.

[33] Given the widespread prediction that risk assessment use lowers recidivism, we use a one-sided hypothesis test. A two-sided test would allow us to reject decreases greater than 0.2 percentage points.

Some of these factors may create conflicting goals. For instance, if the goal at sentencing is to incarcerate those at highest risk of reoffending, then young people should generally get longer sentences. However, if the goal at sentencing is to punish those who are most deserving of punishment, then many would argue that young people should get shorter sentences. After all, youth are impulsive, shortsighted, and susceptible to peer influence. The analysis shown in Section IVB indicates that judges consider young age to be at least partially a mitigating factor. If one of the most important predictors of recidivism is effectively off-limits, the potential to reduce crime by incarcerating those at highest risk is curtailed.

V. Conclusion

The conversation about risk assessment is often framed as man versus machine. But in practice, risk assessment is almost always just an input into a human decision. The decision-maker must decide when to follow the recommendation associated with the risk assessment, and when to deviate from it. In doing so, the decision-maker balances concerns about reoffending against other relevant objectives.

This paper provides information about the impacts of risk assessment in the hands of humans. We show that judges are influenced by the risk assessment, but that discretion in use remains important. They diverted far fewer defendants than were recommended by the algorithm and diverted others against recommendations. The willingness to follow the recommendation is strongly associated with demographic factors. Conditional on the risk score, judges sentenced White people more leniently than Black, and young people more leniently than older ones.

Risk assessment's on-the-ground impact was different from what many had anticipated. Despite the explicit goal of reducing prison populations, the net sentencing effect was null. It led to an increase in relative sentences for the young—although far less than would have been in the absence of discretion—and it had no detectable public safety benefits. We argue that this is at least partially due to conflicting objectives in sentencing: treating young age as a mitigator instead of an aggravator. Our results underline the importance of considering human incentives and cognitive processes when implementing an algorithm.

REFERENCES

Abaluck, Jason, Leila Agha, David C. Chan Jr., Daniel Singer, and Diana Zhu. 2020. "Fixing Misallocation with Guidelines: Awareness versus Adherence." NBER Working Paper 27467.

Agan, Amanda, and Sonja Starr. 2018. "Ban the Box, Criminal Records, and Racial Discrimination: A Field Experiment." *Quarterly Journal of Economics* 133 (1): 191–235.

Aizer, Anna, and Joseph J. Doyle Jr. 2015. "Juvenile Incarceration, Human Capital and Future Crime: Evidence from Randomly-Assigned Judges." *Quarterly Journal of Economics* 130 (2): 759–804.

Albright, Alex. 2023. "The Hidden Effects of Algorithmic Recommendations." Unpublished.

Angelino, Elaine, Nicholas Larus-Stone, Daniel Alabi, Margo Seltzer, and Cynthia Rudin. 2018 "Learning Certifiably Optimal Rule Lists for Categorical Data." arXiv:1704.01701.

Arnold, David, Will Dobbie, and Peter Hull. 2021. "Measuring Racial Discrimination in Algorithms." *AEA Papers and Proceedings* 111: 49–54.

Bayer, Patrick, Randi Hjalmarsson, and David Pozen. 2009. "Building Criminal Capital behind Bars: Peer Effects in Juvenile Corrections." *Quarterly Journal of Economics* 124 (1): 105–47.

Berk, Richard. 2017. "An Impact Assessment of Machine Learning Risk Forecasts on Parole Board Decisions and Recidivism." *Journal of Experimental Criminology* 13 (2): 193–216.

Berk, Richard A., Heather Ladd, Heidi Graziano, and Jong-Ho Baek. 2002. "A Randomized Experiment Testing Inmate Classification Systems Inmate Classification." *Criminology & Public Policy* 2 (2): 215–42.

Berk, Richard A., Susan B. Sorenson, and Geoffrey Barnes. 2016. "Forecasting Domestic Violence: A Machine Learning Approach to Help Inform Arraignment Decisions." *Journal of Empirical Legal Studies* 13 (1): 94–115.

Burton, Jason W., Mari-Klara Stein, and Tina Blegind Jensen. 2020. "A Systematic Review of Algorithm Aversion in Augmented Decision Making." *Journal of Behavioral Decision Making* 33 (2): 220–39.

Calonico, Sebastian, Matias D. Cattaneo, and Rocio Titiunik. 2014. "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs." *Econometrica* 82 (6): 2295–326.

Chalfin, Aaron, Oren Danieli, Andrew Hillis, Zubin Jelveh, Michael Luca, Jens Ludwig, and Sendhil Mullainathan. 2016. "Productivity and Selection of Human Capital with Machine Learning." *American Economic Review* 106 (5): 124–27.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014. "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood." *American Economic Review* 104 (9): 2633–79.

Cowgill, Bo. 2019. "The Impact of Algorithms on Judicial Discretion: Evidence from Regression Discontinuities." Unpublished.

Cowgill, Bo, and Megan T. Stevenson. 2020. "Algorithmic Social Engineering." *AEA Papers and Proceedings* 110: 96–100.

Dietvorst, Berkeley J., Joseph P. Simmons, and Cade Massey. 2015. "Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Fail." *Journal of Experimental Psychology* 144 (1): 114–26.

Doleac, Jennifer L., and Benjamin Hansen. 2020. "The Unintended Consequences of 'Ban the Box': Statistical Discrimination and Employment Outcomes When Criminal Histories Are Hidden." *Journal of Labor Economics* 38 (2): 321–74.

Dressel, Julia, and Hany Farid. 2018. "The Accuracy, Fairness, and Limits of Predicting Recidivism." *Science Advances* 4 (1): eaao5580.

Enamorado, Ted, Benjamin Fifield, and Kosuke Imai. 2019. "Using a Probabilistic Model to Assist Merging of Large-Scale Administrative Records." *American Political Science Review* 113 (2): 353–71.

Garrett, Brandon L., Alexander Jakubow, and John Monahan. 2019. "Judicial Reliance on Risk Assessment in Sentencing Drug and Property Offenders: A Test of the Treatment Resource Hypothesis." *Criminal Justice and Behavior* 46 (6): 799–810.

Garrett, Brandon L., and John Monahan. 2020. "Judging Risk." *California Law Review* 108: 439–93.

Goel, Sharad, Justin M. Rao, and Ravi Shroff. 2016. "Personalized Risk Assessments in the Criminal Justice System." *American Economic Review* 106 (5): 119–23.

Grove, William M., and Paul E. Meehl. 1996. "Comparative Efficiency of Informal (Subjective, Impressionistic) and Formal (Mechanical, Algorithmic) Prediction Procedures: The Clinical-Statistical Controversy." *Psychology, Public Policy, and Law* 2 (2): 293–323.

Gruber, Jonathan, Benjamin R. Handel, Samuel H. Kina, and Jonathan T. Kolstad. 2020 "Managing Intelligence: Skilled Experts and AI in Markets for Complex Products." NBER Working Paper 27038.

Harrison, Paige M., and Allen J. Beck. 2004. *Prisoners in 2003*. Washington, DC: Bureau of Justice Statistics.

Hoffman, Mitchell, Lisa B. Kahn, and Danielle Li. 2017. "Discretion in Hiring." *Quarterly Journal of Economics* 133 (2): 765–800.

Imai, Kosuke, Zhichao Jiang, James Greiner, Ryan Halen, and Sooahn Shin. 2020. "Experimental Evaluation of Algorithm-Assisted Human Decision-Making: Application to Pretrial Public Safety Assessment." arXiv:2012.02845.

Johndrow, James E., and Kristian Lum. 2019. "An Algorithm for Removing Sensitive Information: Application to Race-Independent Recidivism Prediction." *Annals of Applied Statistics* 13 (1): 189–220.

Jung, Jongbin, Connor Concannon, Ravi Shroff, Sharad Goel, and Daniel G. Goldstein. 2017. "Simple Rules for Complex Decisions." arXiv:1702:04690.

Kasy, Maximilian, and Rediet Abebe. 2021. "Fairness, Equality, and Power in Algorithmic Decision-Making." In *FAccT '21, Proceedings of the 2021, ACM Conference on Fairness, Accountability, and Transparency*, 576–86. New York: Association for Computing Machinery.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. "Human Decisions and Machine Predictions." *Quarterly Journal of Economics* 133 (1): 237–93.

Lin, Zhiyuan, Jongbin Jung, Sharad Goel, and Jennifer Skeem. 2020. "The Limits of Human Predictions of Recidivism." *Science Advances* 6 (7): eaaz0652.

Lipton, Zachary C., Alexandra Chouldechova, and Julian McAuley. 2017. "Does Mitigating ML's Impact Disparity Require Treatment Disparity?" arXiv:1711:07076.

Mason, S. J., and N. E. Graham. 2002. "Areas beneath the Relative Operating Characteristics (ROC) and Relative Operating Levels (ROL) Curves: Statistical Significance and Interpretation." *Quarterly Journal of the Royal Meteorological Society* 128 (584): 2145–66.

Metz, Anne, John Monahan, Luke Siebert, and Brandon Garrett. 2020. "Valid or Voodoo? A Qualitative Study of Attorney Perceptions of Risk Assessment in Sentencing and Plea Bargaining." *Journal of Community Psychology* 48 (6): 2053–68.

Mitchell, Kelly Lyn. 2017. *Sentencing Commissions and Guidelines by the Numbers: Cross-Jurisdictional Comparisons Made Easy by the Sentencing Guidelines Resource Center*. Minneapolis, MN: Robina Institute in Criminal Law and Criminal Justice.

Ostrom, Brian J., Fred Cheesman, Ann M. Jones, Meredith Peterson, and Neal B. Kauder. 1999. *Truth-in-Sentencing in Virginia*. Williamsburg, VA: National Center for State Courts.

Ostrom, Brian J., Mathew Kleiman, Fred Cheesman, Randall M. Hansen, and Neal B. Kauder. 2002. *Offender Risk Assessment in Virginia: A Three Stage Evaluation*. Williamsburg, VA: National Center for State Courts.

Skeem, Jennifer, Nicholas Scurich, and John Monahan. 2020. "Impact of Risk Assessment on Judges' Fairness in Sentencing Relatively Poor Defendants." *Law and Human Behavior* 44 (1): 51–59.

Sloan, CarlyWill, George Naufal, and Heather Caspers. 2023. "The Effect of Risk Assessment Scores on Judicial Behavior and Defendant Outcomes." *Journal of Human Resources*. https://doi.org/10.3368/jhr.0221-11470R3.

Stevenson, Megan. 2017. "Breaking Bad: Mechanisms of Social Influence and the Path to Criminality in Juvenile Jails." *Review of Economics and Statistics* 99 (5): 824–38.

Stevenson, Megan. 2019. "Assessing Risk Assessment in Action." *Minnesota Law Review* 103 (1): 303–84.

Stevenson, Megan, and Christopher Slobogin. 2018. "Algoirithmic Risk Assessment and the Double Edged Sword of Youth." *Washington University Law Review* 96 (3): 681–706.

Stevenson, Megan T., and Doleac, Jennifer L. 2024. *Data and Code for: "Algorithmic Risk Assessment in the Hands of Humans."* Nashville, TN: American Economic Association; distributed by Inter-university Consortium for Political and Social Research, Ann Arbor, MI. https://doi.org/10.3886/E193667V1.

Virginia Court Data. 2018. "Circuit Criminal Court Data." Virginia Court Data. https://virginiacourtdata.org/ (accessed July 1, 2018).

Virginia Criminal Sentencing Commission (VCSC). 2001. *Assessing Risk among Sex Offenders in Virginia.* Richmond, VA: VCSC.

Virginia Criminal Sentencing Commission (VCSC). 2003. *Virginia Criminal Sentencing Commission: Annual Report 2003*. Richmond, VA: VCSC.

Virginia Criminal Sentencing Commission (VCSC). 2018. "Sentencing Commission Data." VCSC. URL (accessed July 1, 2018).

Yang, Crystal, and Will Dobbie. 2020. "Equal Protection under Algorithms: A New Statistical and Legal Framework." *Michigan Law Review* 199 (2): 291–395.