1. Introduction. Over the last 10 years, New York City residents have been detained and searched by the police millions of times in an effort to get weapons, drugs, and other contraband off the streets. Proponents of this stop-question-frisk policy (hereafter called stop-and-frisk) argue that by strictly enforcing weapon and drug possession laws, one indirectly reduces more serious crime, such as murder and armed robbery, in line with the “broken windows” theory of policing (Wilson and Kelling, 1982). Though it is difficult to rigorously assess this claim, wide adoption of stop-and-frisk by the New York City Police Department (NYPD) in the early 1990’s did coincide with a period of substantial decline in crime in the city. Opponents of stop-and-frisk, however, argue that regardless of whether the policy is effective, it violates two constitutional protections. First, they claim individuals
are stopped without legal basis, in violation of the Fourth Amendment. Indeed, in over 90% of cases, stopped suspects are released without any further action, suggesting that the vast majority of individuals stopped were not engaged in any criminal activity. Second, they claim the policy is not applied in a race-neutral manner, in violation of the Fourteenth Amendment. Notably, blacks and Hispanics make up 80% of individuals stopped even though they constitute only 50% of the New York City population.

Though substantial effort has gone into studying these issues, there unfortunately appears to be little consensus in the academic literature on the magnitude—or even on the existence—of racial discrimination in stop-and-frisk and related tactics, even among papers studying the same policy in the same city over the same timeframe. Moreover, nearly all such research has focused on claims of racial discrimination, with little statistical attention paid to possible Fourth Amendment violations. In an effort to cast further light on this ongoing statistical and policy debate, we analyzed three million stops conducted by New York City police officers between 2008 and 2012, one of the largest studies of stop-and-frisk to date. Of these three million stops, we focus our attention on the approximately 760,000 instances in which an individual was detained under suspicion of criminal possession of a weapon (CPW), since the success of these stops is readily determined by the presence or absence of a weapon.

We make three main contributions. First, to investigate possible Fourth Amendment violations, we assess the extent to which stops meet the standard of “reasonable suspicion” established in Terry v. Ohio (1968) and subsequently expanded on in several court rulings, including Illinois v. Wardlow (2000).

1 Reasonable suspicion exists when there are articulable facts or circumstances which would lead a reasonable person to suspect that a crime has been, is being, or will be committed—a standard of proof lower than probable cause but higher than a mere hunch. In particular, random searches, regardless of whether they are an effective deterrent, are generally prohibited under the Fourth Amendment. We take a statistical approach to infer the level of evidence used to justify a stop. Namely, we estimate the ex-ante likelihood—based only on information available to officers prior to the stop decision—that the individual has a weapon. We find that in 44% of the approximately 760,000 CPW stops we consider, there was at most a 1%

---

1 Stop-and-frisk has a complicated legal history, which we only briefly address in this paper. For a more comprehensive review, see Gelman, Fagan and Kiss (2007).
2 There are some exceptions to this general rule. For example, the U.S. Supreme Court has found sobriety checkpoints to be constitutional (Michigan Dept. of State Police v. Sitz, 1990).
chance of finding a weapon on the suspect. Whether these low hit rate stops in fact violate the Fourth Amendment is a complex legal question, and one that is largely outside the scope of this paper. Nevertheless, our results do suggest that individuals were often stopped with relatively little evidence of criminal activity, corroborating recent court rulings rebuking the NYPD for its stop-and-frisk tactics (Floyd v. City of New York, 2013a; Davis v. City of New York, 2013b; Ligon v. City of New York, 2013c).

Second, we find that blacks and Hispanics were disproportionally involved in low hit rate stops. We show that this pattern largely results from lower stop thresholds in high crime areas, particularly public housing, presumably reflecting more aggressive efforts to reduce crime in those locations. Since these areas are home to a disproportionate number of blacks and Hispanics, members of these groups were disproportionately impacted by stop standards that differed by location. After correcting for these highly localized policing tactics—as well as adjusting for several other factors—we find that stopped blacks and Hispanics were still somewhat less likely than similarly situated whites to possess a weapon, suggestive of at least a degree of racial discrimination. It appears, though, that racial disparities induced by bias are considerably smaller than those induced by location-specific stop thresholds.

Finally, we argue that because of the large number of low hit rate stops, it is possible to dramatically reduce the overall number of stops while largely preserving the number of successful stops. In particular, we show that one can recover 50% of weapons by conducting only the 6% of CPW stops with the highest ex-ante hit rate. Note that these ex-ante hit rates are based only on information observable to officers prior to the stop decision, and so it is at least in theory possible to implement such a strategy. Further, since low hit rate stops disproportionately involve blacks and Hispanics, optimizing for weapons recovery would simultaneously bring more racial balance to stop-and-frisk. To facilitate adoption of such strategies by police departments, we develop stop heuristics that approximate our full statistical model via a simple scoring rule. Specifically, we show that with a rule consisting of only three weighted stop criteria, one can recover the majority of weapons by conducting 7.5% of stops.

Given the significance and salience of stop-and-frisk, a number of statistical studies have assessed various aspects of the issue, particularly claims of racial bias, which we briefly review. In an early, comprehensive analysis, Gelman, Fagan and Kiss (2007) concluded that minorities were stopped more often than whites, both in comparison to their proportion in the local
population, and relative to local crime rates in those groups. A subsequent analysis also found evidence of racial disparities, but concluded that the magnitude of the effect was relatively small, and in particular estimated that only 15 out of 3,000 NYPD officers stopped an unusually high number of black and Hispanic suspects (Ridgeway, 2007; Ridgeway and MacDonald, 2009). Coviello and Persico (2013) fit an economic model of behavior to the stop-and-frisk data, and found no evidence of racial bias. Finally, investigating the ramifications of local events on policing, Legewie (2014) showed that in the days following fatal shootings of two NYPD officers by black suspects, there was an increase in the use of physical force against blacks—but not whites or Hispanics—during stops; moreover, such increase in force was not observed after the murder of two police officers by a white and an Hispanic suspect.

Several papers have also studied the closely related issue of racial discrimination in traffic stops, which is extensively detailed in Epp, Maynard-Moody and Haider-Markel (2014). In a novel design, Grogger and Ridgeway (2006) studied traffic stops in Oakland, and showed that the racial distribution of stopped individuals during the day, when a suspect’s race is readily apparent, matches the distribution at night, when the “veil of darkness” masks race, and thus concluded there was little bias in stop decisions. However, Ridgeway (2006) finds differences in post-stop outcomes by race; for example, black drivers are less likely than whites to have stops lasting less than ten minutes. Knowles, Persico and Todd (2001) distinguish between so-called statistical and taste-based discrimination (Arrow, 1973), and do not find evidence of racial prejudice against blacks in Maryland traffic stops. Examining traffic stops by the Boston Police Department, however, Antonovics and Knight (2009) show that officers are more likely to conduct a search if the race of the officer differs from the race of the driver, consistent with taste-based racial discrimination.

2. Data & Methods.

2.1. Data description. Our primary dataset consists of all 2.9 million stops conducted and recorded by the New York City Police Department (NYPD) between January 1, 2008 and December 31, 2012. Following a stop, officers are encouraged to complete a UF-250 stop-and-frisk form, recording various aspects of the stop, including demographic characteristics of the suspect, the time and location of the stop, the suspected crime, and the rationale for the stop (e.g., whether the suspect was wearing clothing common in the commission of a crime). After an individual is stopped, officers have discretion to frisk or search the suspect (which occurs in 56% and 9% of
Table 1
Summary of key information recorded on the UF-250 stop-and-frisk form.

<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>date</td>
<td>yyyy-mm-dd</td>
</tr>
<tr>
<td>time</td>
<td>hh:mm</td>
</tr>
<tr>
<td>location</td>
<td>GPS coordinates</td>
</tr>
<tr>
<td>precinct</td>
<td>1–123</td>
</tr>
<tr>
<td>location type</td>
<td>public housing, public transit, or neither</td>
</tr>
<tr>
<td>inside or outside</td>
<td>inside or outside</td>
</tr>
<tr>
<td>suspect’s sex</td>
<td>male or female</td>
</tr>
<tr>
<td>suspect’s race</td>
<td>white, black, Hispanic, Asian, or other</td>
</tr>
<tr>
<td>suspect’s build</td>
<td>heavy, medium, muscular, or thin</td>
</tr>
<tr>
<td>suspect’s age</td>
<td>integer (years)</td>
</tr>
<tr>
<td>suspect’s height</td>
<td>integer (inches)</td>
</tr>
<tr>
<td>suspect’s weight</td>
<td>integer (pounds)</td>
</tr>
<tr>
<td>observation period</td>
<td>integer (minutes)</td>
</tr>
<tr>
<td>officer in uniform</td>
<td>yes or no</td>
</tr>
<tr>
<td>radio run</td>
<td>yes or no</td>
</tr>
<tr>
<td>suspected crime</td>
<td>1 of 113 pre-specified categories (e.g., criminal possession of a weapon, and robbery)</td>
</tr>
<tr>
<td>primary stop circumstance(s)</td>
<td>suspicious object, fits description, casing, acting as lookout, suspicious clothing, drug transaction, furtive movements, actions of violent crime, suspicious bulge, and/or other</td>
</tr>
<tr>
<td>additional stop circumstance(s)</td>
<td>witness report, ongoing investigation, proximity to crime scene, evasive response, associating with criminals, changed direction, high crime area, time of day, sights and sounds of criminal activity, and/or other</td>
</tr>
<tr>
<td>suspect frisked</td>
<td>yes or no</td>
</tr>
<tr>
<td>suspected searched</td>
<td>yes or no</td>
</tr>
<tr>
<td>suspect arrested</td>
<td>yes or no</td>
</tr>
<tr>
<td>weapon found on suspect</td>
<td>yes or no</td>
</tr>
<tr>
<td>drugs found on weapon</td>
<td>yes or no</td>
</tr>
</tbody>
</table>

cases, respectively), and may decide to make an arrest (6% of instances) or issue a summons (6% of instances), all of which is recorded on the UF-250 form. Responses are subsequently standardized, compiled, and released annually to the public. A list of key information collected is summarized in Table 1.

While officers are in some instances mandated to complete a UF-250 form (e.g., if force was used), they are not always required to do so, and so a possibly large number of stops go undocumented. Also, there is evidence that officers follow “scripts of suspicion” when filling out forms to justify stops (Fagan and Geller, 2014). Further, it is not always even clear whether a police encounter formally constitutes a “stop”. (The legal test is whether
a reasonable person would not have felt free to terminate the encounter, though there is at times genuine ambiguity with this criterion.) Finally, since these forms are completed by hand, there are likely errors in recording and transcribing stop details. Our dataset is thus neither a complete nor fully accurate record of all conducted stops. Nevertheless, we note that in light of recent litigation (Daniels, et al. v. the City of New York, 2001), the NYPD now works to ensure UF-250 accuracy, including supervisor review. Moreover, these data (and related datasets) have been used in past academic work (Gelman, Fagan and Kiss, 2007; Ridgeway, 2007; Ridgeway and MacDonald, 2009), and in a variety of high-profile court cases, including Floyd v. City of New York (2013a). As such, we assume the data are generally suitable for our analysis, and we note the effect of possible problems with the data on our results where appropriate.

A common metric for evaluating stop-and-frisk is the so-called hit rate, the proportion of stops in which a suspect was arrested, a summons issued, or some other outcome occurred that suggests the guilt of a stopped individual. Hit rates are regularly used to assess the level of proof applied when stopping a suspect, with lower hit rates corresponding to less stringent standards (Ayres, 2002; Becker, 1993). In particular, lower arrest rates for stopped blacks relative to stopped whites are often interpreted as indicating that the threshold for stopping blacks is lower than for stopping whites, consistent with claims of racial discrimination. However, as noted in Gelman, Fagan and Kiss (2007), one could reasonably reach the opposite conclusion: relatively higher arrest rates of whites could indicate that officers are biased against whites in that they arrest them too often.

To circumvent these issues of interpretation, we first subset the data to include only the 760,502 stops between 2008 and 2012 for which the suspected crime was listed as criminal possession of a weapon (CPW), by far the most commonly occurring suspected crime in our dataset. We note that officers are required to articulate the suspected crime prior to conducting the stop, though this information, along with all other stop details, is recorded afterwards. Then, instead of considering whether or not the stopped individual was arrested, we look at whether the suspect was found to have a weapon, (which is also recorded on the UF-250 form). This approach has three advantages. First, relative to arresting a suspect, there is arguably less officer discretion involved in determining whether an individual has a weapon.3 Second, the presence or absence of a weapon directly indicates whether the stop was (ex-post) justified under the explicitly stated suspicion of CPW. In contrast, a stopped individual could be arrested for a variety of reasons.

---

3We indeed see that not all suspects found to have a weapon are arrested.
(e.g., drug possession) that are unrelated to the original purpose of the stop. Finally, by focusing on a single class of well-defined suspected crimes, we mitigate ecological fallacies due to different base hit rates for various crime categories. For example, since hit rates are generally higher when the suspected crime is drug-related, and since a relatively higher proportion of whites are involved in these drug stops, one could reach spurious conclusions by aggregating all stops.

2.2. Estimating stop-level hit rates. As discussed above, estimating race-specific hit rates helps both to assess whether stops meet the standard of reasonable suspicion and also to test for racial discrimination. Traditionally, hit rates are estimated by simply computing the overall percentage of stops among each race group that result in the outcome of interest (e.g., of finding a weapon on, or arresting the stopped suspect), possibly controlling for the distinct contexts (e.g., time of day) in which individuals of different races are stopped (Gelman, Fagan and Kiss, 2007; Ridgeway, 2007). In contrast to these aggregate, race-level hit rate statistics, our aim is to estimate stop-level hit rates. Specifically, our primary quantity of interest is the ex-ante likelihood that any given CPW stop results in finding a weapon on the stopped suspect. That is, at the moment an officer decides to stop an individual for suspicion of criminal possession of a weapon, we seek the probability—taking all information available to the officer at the time—that the suspect has a weapon. This methodological approach has two advantages. First, by computing the full hit rate distribution, we can estimate the fraction of CPW stops that fall below a given evidence threshold (e.g., where the likelihood of finding a weapon is less than 1%), which in turn helps to assess possible violations of Fourth Amendment protections against unreasonable search. Second, stop level probabilities can be efficiently aggregated to estimate hit rates for various small subgroups (e.g., stopped Hispanics in a given precinct), circumventing issues of data sparsity and allowing us to quantify the extent to which the threshold for stopping individuals differs across contexts.

To compute stop-level hit rates, we fit a logistic regression model on all 760,502 CPW stops between 2008 and 2012, where the left hand side is the probability of finding a weapon on the stopped suspect, and the right hand side includes several variables recorded on the officer’s UF-250 report that would have been available immediately before the stop. Specifically, we include indicator variables for the suspect’s demographics (sex, race, and build); whether the stop occurred on public transit, in public housing, or neither; whether the stop occurred inside or outside; the date and time of
the stop (month, day of week, and time of day, binned into disjoint four-hour blocks); one or more reasons for the stop (e.g., furtive movements and high crime area, as detailed in Table 1); whether the stop was the result of a radio run; and whether the officer was in uniform. We additionally include continuous variables for the year, suspect’s height, weight, and age, and the time for which the officer observed the suspect before stopping him or her (the latter four are all normalized to have mean 0 and variance 1). Importantly, we also include indicator variables for the precinct where the stop occurred, which helps to account for both local crime rates and enforcement standards that differ by precinct commander. Finally, we include in the model all pairwise interactions between these variables (including self-interactions). Thus the final form of the model is

\[
P(y_i = 1) = \text{logit}^{-1} \left( \sum_k \alpha_k x_{k,i} + \sum_{k \leq \ell} \beta_{k,\ell} x_{k,i} x_{\ell,i} \right),
\]

where \(y_i\) indicates whether the \(i\)-th stop resulted in finding a weapon on the suspect, \(x_{k,i}\) denotes features of the stop, and \(\alpha\) and \(\beta\) are the model coefficients.

Given the large number of variables and stop instances that we consider, we fit the model with stochastic gradient descent (c.f. Bottou (1998)) as implemented in the open-source package Vowpal Wabbit. To provide some insight into which features the model makes the most use of, we list in Table 2 the positive and negative coefficients with largest absolute value.

The strength of our conclusions rests on the accuracy of our model, and so we conducted several tests to assess model performance. First, we separated out the data into two sets, one comprised of the 631,334 CPW stops from 2008–2011, and another consisting of the 129,168 CPW stops in 2012. When the model in Eq. (1) is fit only on the 2008–2011 data, we find an AUC of 86%; and when that same fitted model is tested on the 2012 data, we have an AUC of 83%. Thus, not only is the out-of-sample performance high, but the similarity between in-sample and out-of-sample performance indicates the model is not overfitting to the data. We checked calibration by comparing the model-predicted probabilities to the empirical hit rates. Figure 7 in the Appendix confirms the model is well-calibrated along the entire range of predicted probabilities. Further, we conducted several predictive checks, comparing the model estimates to the empirical hit rates for various demographic groups, and find the model performs quite well (Figure 7, in

---

4We used default values for all algorithm parameters.
Table 2
The ten positive and negative model coefficients with largest absolute value.

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(location = public housing) × (suspicious object)</td>
<td>0.597</td>
</tr>
<tr>
<td>(location = public transit) × (precinct 73)</td>
<td>0.508</td>
</tr>
<tr>
<td>(location = public transit) × (radio run)</td>
<td>0.466</td>
</tr>
<tr>
<td>(location = neither transit nor housing) × (suspicious object)</td>
<td>0.417</td>
</tr>
<tr>
<td>(witness report) × (fits description)</td>
<td>0.395</td>
</tr>
<tr>
<td>(suspicious object) × (suspect race = black)</td>
<td>0.394</td>
</tr>
<tr>
<td>suspicious object</td>
<td>0.379</td>
</tr>
<tr>
<td>(suspicious object) × (suspect sex = male)</td>
<td>0.358</td>
</tr>
<tr>
<td>(Friday) × (suspicious clothing)</td>
<td>0.352</td>
</tr>
<tr>
<td>(witness report) × (precinct 50)</td>
<td>0.352</td>
</tr>
<tr>
<td>(precinct 114) × (acting as lookout)</td>
<td>-2.041</td>
</tr>
<tr>
<td>(precinct 46) × (time between 4 and 8 am)</td>
<td>-1.404</td>
</tr>
<tr>
<td>(precinct 48) × (acting as lookout)</td>
<td>-1.314</td>
</tr>
<tr>
<td>(precinct 101) × (November)</td>
<td>-1.287</td>
</tr>
<tr>
<td>(precinct 101) × (associating with criminals)</td>
<td>-1.272</td>
</tr>
<tr>
<td>(precinct 46) × (acting as lookout)</td>
<td>-1.255</td>
</tr>
<tr>
<td>(precinct 115) × (acting as lookout)</td>
<td>-1.245</td>
</tr>
<tr>
<td>(precinct 69) × (suspicious clothing)</td>
<td>-1.206</td>
</tr>
<tr>
<td>(precinct 88 × (suspicious clothing)</td>
<td>-1.184</td>
</tr>
<tr>
<td>(precinct 113) × (Sunday)</td>
<td>-1.164</td>
</tr>
</tbody>
</table>

the Appendix). The totality of evidence thus suggests our model produces reliable estimates of the ex-ante likelihood of CPW.

Lastly, we note that although a suspect’s height, weight, and age can only be approximated by the officer before the stop, the fitted model is largely robust to reasonable errors in these terms. In particular, if we assume officers estimate height, weight, and age with independent, mean zero errors that are normally distributed with standard deviations of 2 inches, 10 pounds, and 5 years, respectively, the mean absolute change in estimated ex-ante probability is only 0.2 percentage points.

3. Results.

3.1. Assessing reasonable suspicion. With the fitted model described by Eq. (1) in hand, we assign each CPW stop a model-inferred ex-ante probability of finding a weapon on the stopped suspect. Figure 1(a) shows the distribution of these ex-ante probabilities for the approximately 760,000 CPW stops between 2008 and 2012. As indicated by the dotted vertical line, the overall likelihood of finding a weapon is just 3%. Moreover, 44% of the stops had less than a 1% chance of turning up a weapon, and 20% of stops had
Fig 1. Distribution of the ex-ante probability of finding a weapon on a suspect stopped for suspicion of criminal possession of a weapon (CPW). Panel (a) shows the distribution over all such stops between 2008 and 2012, with the vertical line indicating the overall likelihood of finding a weapon on a stopped suspect. 44% of all CPW stops have less than a 1% ex-ante chance of turning up a weapon. Panel (b) disaggregates this distribution by suspect race, where the vertical lines show likelihood of finding a weapon on black, Hispanic, and white suspects. Stopped blacks and Hispanics are much less likely to have a weapon than stopped whites.

As a point of comparison, in the landmark New York City stop-and-frisk court case, Floyd v. City of New York (2013a), reasonable suspicion was assessed by hand-classifying each possible stated justification for the stop (as indicated on the UF-250 form) as reasonable or not. For example, whereas
“furtive movements” in the absence of any other indicator of criminality was deemed insufficient justification, “furtive movements” together with “high crime area” was deemed acceptable. Though that analysis resulted in 5% of all stops (including non-CPW stops) classified as unreasonable, the presiding judge in the case believed the classification overly conservative, and suggested the true number of stops lacking reasonable suspicion was likely considerably higher (Floyd v. City of New York, 2013a, p. 41). In contrast, while our purely statistical approach admittedly does not explicitly consider legal precedent, it does offer a straightforward, fast, and largely objective method for directly relating reasonable suspicion to criminality.

Figure 1(b) shows stop-level hit rate distributions broken down by the race of the stopped suspect (black, Hispanic, or white), with the vertical lines indicating overall hit rates for each race group. In particular, consistent with past results (Gelman, Fagan and Kiss, 2007), the overall hit rates for blacks and Hispanics (2.5% and 4%, respectively) are considerably lower than for whites (13%). In other words, when blacks and Hispanics are stopped, it is typically the result of a lower degree of evidence than when white suspects are stopped. Moreover, while 50% of blacks stopped under suspicion of CPW have less than a 1% chance of in fact possessing a weapon, the corresponding fraction for Hispanics is 36%, and is just 13% for stopped whites. Thus, if we equate reasonable suspicion with a particular probability threshold (say 1%), a far greater fraction of stops of blacks and Hispanics are unwarranted than are stops of whites.

3.2. Heterogeneity in hit rate by location. It is perhaps tempting to conclude that the lower hit rates of blacks (2.5%) and Hispanics (4%) relative to whites (13%) is indicative of racial discrimination. However, as Ridgeway (2007) points out, whites and minorities are typically stopped in different contexts, and so differing hit rates may not be the result of racial bias. Indeed, as we discuss below, stop-and-frisk is an extremely localized tactic, heavily concentrated in high-crime, predominantly black and Hispanic areas, and so lower tolerance for suspicious activity (and hence lower hit rates) in these areas could account for the racial disparity.

Figure 2(a) shows the distribution of CPW stops colored by the race of the stopped suspect (a random sample of 10,000 stops is plotted), illustrating how geographically specific the use of stop-and-frisk is. For comparison, Figure 2(b) plots each recorded homicide in New York City from 2006–2011, 2427 in total, as compiled by the New York Times. The distribution

---

6See http://projects.nytimes.com/crime/homicides/map for further details. Only data up until 2011 were available.
Fig 2. Panel (a) shows the geographic distribution of CPW stops between 2008 and 2012, colored by the suspect’s race. For comparison, Panel (b) shows the distribution of murders in New York City between 2006 and 2011, which indicates that stop-and-frisk is primarily employed in high-crime areas. Finally, Panel (c) shows the racial distribution of the general population based on 2010 block-level U.S. Census data, highlighting that these high-crime, high stop-and-frisk areas are disproportionately black and Hispanic.
of homicides is remarkably well-aligned with the distribution of stops, indicating that the NYPD concentrated its use of stop-and-frisk on high-crime areas. Finally, Figure 2(c) shows the distribution of the general New York City population, by race, based on 2010 block-level U.S. Census data. To generate the plot, 10,000 individuals were sampled from Census records and placed on the map at the middle of their Census block, the smallest geographic unit for which information is publicly available.

The maps in Figure 2 highlight three points. First, there is an almost one-to-one correspondence between areas with heavy use of stop-and-frisk (Figure 2(a)) and areas with high incidence of violent crime (Figure 2(b)). While this is a natural and possibly effective policing strategy, a consequence of the tactic is that individuals who live in high-crime areas, but who are not themselves engaged in criminal activity, bear the costs associated with being stopped. This observation is particularly relevant since the vast majority of stopped suspects are not found guilty of any wrongdoing. Second, these high-crime areas are overwhelmingly black and Hispanic. Accordingly, the cost of stop-and-frisk is largely shouldered by these minorities who live in high-crime areas. Third, by comparing Figures 2(a) and 2(c), we see that the racial composition of stopped individuals is similar to the racial composition of the neighborhoods in which stop-and-frisk is heavily employed. Thus, the striking racial composition of stopped CPW suspects (61% are black, 30% are Hispanic, and 4% are white) appears at least qualitatively attributable to selective use of stop-and-frisk in minority-heavy areas, illustrating the importance of understanding the localized nature of the policy.

Adding quantitative detail to these qualitative results, we estimate hit rates for each of the 77 precincts in New York City, and further distinguish between stops occurring in public housing, on public transit, or in other locations (primarily pedestrian stops). Figure 3(a) shows the results, plotting the hit rate of white versus black suspects for each location, with the size of the points indicating the number of stops. To generate the estimates, the ex-ante probabilities from Eq. (1) are averaged over stops in each geographic area; for areas with a large number of stops, the model agrees with the simple, empirical hit rate, but the model-estimated statistics lead to more stable estimates for the sparser regions.

As indicated by the plot, there is substantial variation in average hit rate across locations, ranging from less than 1% in some public housing units to more than 30% for transit stops in certain precincts. Moreover, within region, though the hit rates of white and black suspects are not identical,
they are much more similar than the city-wide averages (indicated by the dashed horizontal lines). Specifically, the average within-area ratio of white hit rate to black hit rate, weighted by the number of stops in each area is 2.2. By comparison, the city-wide ratio is 5.1. Thus, a significant fraction of the racial disparity in hit rates can be explained by policing tactics that vary considerably by area.

Figure 3(b) further illustrates this point, plotting for each area its overall hit rate (irrespective of race) by the percentage of stopped suspects who are white (among suspects who are either white or black). We again see that predominately black areas are associated with low hit rates, while predominately white areas have high hit rates, further demonstrating the importance of location for understanding the adverse race effects of stop-and-frisk.

---

8 This weighted average is $\frac{\sum_{i=1}^{n} w_i r_i}{\sum_{i} w_i}$, where $r_i = \text{white hit rate/black hit rate in area } i$, $w_i$ is the number of stops in area $i$, and $i$ ranges over all precinct/location-type combinations with at least one black and one white stop.
While much of the racial disparity in hit rates is explained by geography, we note that this finding is orthogonal to the question of whether stops meet the standard of reasonable suspicion, as discussed in Section 3.1. Indeed, it appears that stops in several public housing complexes have quite low average hit rates (less than 1%), calling into question that the bar for reasonable suspicion has been met. Corroborating our statistical findings, recent stop-and-frisk lawsuits have revealed that NYPD training materials explicitly instructed officers to question people in New York City Housing Authority buildings “without reasonable suspicion of trespass, and to arrest for trespass those who fail to leave or affirmatively establish their right” to be present in a building (Davis v. City of New York, 2013b).

3.3. Testing for racial discrimination. Although much of the racial disparity in hit rates disappears once we account for the location of a stop, hit rates for whites are still consistently higher than for blacks across geographic area, leaving open the possibility that racial bias is still at play. Location, however, is not the only possible confounding factor. For example, the demographic composition of the local populations could shift with the time of day, aligning with patrol schedules to affect race-specific hit rates. Alternatively, the distribution of age may vary across race, with certain age groups—and consequently race groups—more often the target of low hit rate stops.

To adjust for these alternative, non-race-based explanations for the difference in hit rates between whites and minorities, we use the logistic regression model described above in Eq. (1) to generate counterfactual hit rate estimates. Namely, for each of the 463,109 CPW stops of blacks between 2008 and 2012, we use the model to generate the ex-ante likelihood of finding a weapon on the stopped suspect assuming the suspect were white, but preserving all other aspects of the stop. We note that because the model includes a number of interaction terms, this counterfactual estimate does not simply differ from the original by a constant factor, but rather depends on the precise combination of features describing the stop.

The result of this exercise is displayed in Figure 4, where we plot the hit rate for stopped blacks against the counterfactual hit rate for similarly situated whites, grouped by location. The plot shows that by adjusting for

---

9One may be tempted to conclude that having stop thresholds that vary by location in and of itself indicates the reasonable suspicion standard is violated. Why, one might argue, should one precinct’s bar for reasonable suspicion differ from an adjacent precinct’s? The law, however, requires only that a minimum standard of proof be met, and a precinct may choose not to stop all individuals above that legal threshold for a variety of legitimate reasons, including constrained resources and alternative police priorities.
the various differences in context between stops of whites and blacks, we do indeed shrink the hit rate gap. We also find, however, that the gap does not completely disappear, with the overall counterfactual white hit rate 64% larger than the black hit rate (4.1% compared to 2.5%), indicated by the dashed lines. Further, the higher white hit rate is not simply due to a few anomalous areas, but holds consistently across nearly every location we consider. It thus appears that relative to similarly situated whites, black suspects are indeed stopped with less ex-ante evidence of a crime, corroborating claims of racial discrimination.

Such racial discrimination could in principle arise from two qualitatively distinct mechanisms, statistical or taste-based (Arrow, 1973; Ewens, Tomlin and Wang, 2014; Persico, 2009). With statistical discrimination, officers may genuinely believe that blacks are more likely to carry weapons than the data suggest, perhaps due to faulty heuristics or limited opportunity to estimate event probabilities. In contrast, with taste-based discrimination, officers may accurately estimate ex-ante hit rates, but apply a lower standard of proof when stopping blacks than whites. Our analysis cannot distinguish between these two possible underlying causes.
While we see that racial disparities in stop-and-frisk are in part driven by discrimination, variation in local stop thresholds still appear to be the primary driver. In particular, consider the 100,594 blacks who were stopped in housing projects between 2008–2012, a subgroup with overall hit rate of 0.9%. Now, if we suppose those individuals were white—but otherwise identical—we estimate a counterfactual hit rate of 1.6%, where the increase is indicative of racial discrimination. If, however, we suppose those 100,594 black individuals were stopped on the street, as opposed to in housing, with all other traits—including race and precinct—kept identical, we estimate a counterfactual hit rate of 1.8%, higher than both the actual hit rate (0.9%) and counterfactual white hit rate (1.6%).

3.4. Improving stop efficiency. As we have seen, individuals are regularly stopped under suspicion of CPW in contexts where it is unlikely that they in fact possess a weapon. This observation begs the question, how can we design a better policy? Towards this end, we first re-train the statistical model in Eq. (1) on only the 631,334 CPW stops between 2008 and 2011. Then, for each of the 129,168 CPW stops in 2012, we use this fitted model to generate the ex-ante probability that a stopped suspect has a weapon. Finally, we rank stops in descending order by their model-estimated likelihood of turning up a weapon, with the stops deemed most likely to result in finding a weapon accordingly appearing at the top of the list. This ranking is thus based on out-of-sample predictions, and moreover, only uses data available at the moment right before an officer decides to stop an individual.

For each 2012 CPW stop, we know whether or not a weapon was ultimately found on the suspect. We can thus evaluate how many weapons we would have recovered had only the top $p$-percent of stops been conducted. Figure 5(a) shows this curve, where we normalize the number of recovered weapons on the $y$-axis by the total number of weapons recovered in all CPW stops conducted that year. Remarkably, we find that only 6% of stops are needed to recover the majority of weapons, and only 56% are necessary to turn up 90% of the weapons. Since so few CPW stops have any significant chance of turning up a weapon, we can eliminate the majority of stops and still identify almost as many individuals who are carrying weapons.

Since a disproportionate number of the lowest hit rate stops involve blacks and Hispanics, eliminating these stops alters the racial composition of stopped suspects, as shown in Figure 5(b). For example, whereas blacks make up 62% of all CPW stops in 2012, they comprise 43% of the 10% of stops most likely to result in finding a weapon. A more racially-balanced pool of stopped suspects could temper public reaction to the policy, includ-
Fig 5. Panel (a) plots the estimated percentage of weapons recovered as a function of the number of stops conducted, where stops are ordered by their model-predicted likelihood of turning up a weapon, from highest to lowest. (Note that the x-axis is on a log scale.) In particular, the best 10% of stops result in 58% of weapons recovered, and the best 50% result in 88% of weapons recovered. Panel (b) shows how the racial composition of stopped suspects varies with the number of stops, where stops are again ordered from most to least likely to result in turning up a weapon. Since low likelihood stops disproportionately involve black suspects, reducing the number of stops results in lowering the overall proportion of stopped suspects who are black.

ing resentment and distrust of the police (Lerman and Weaver, 2014). One can thus view this to be an added benefit of improving stop efficiency.

Since late 2013, the use of stop-and-frisk in New York City has been severely curtailed, both because of several court rulings critical of stop-and-frisk as well as a newly-elected Mayor, Bill de Blasio, who openly opposes the tactic. Specifically, during the last four months of 2013, there were 3,985 CPW stops, compared to 33,683 for the same period in 2012, a reduction of 88%. Are officers systematically conducting only the “best” stops (i.e., those stops most likely to result in finding a weapon on the suspect)? Indeed we find the CPW hit rate for the end of 2013 is substantially higher, 11%, as compared to the same period in 2012, 3%. However, had the officers conducted the 3,985 stops ranked highest by our model, we would expect a hit rate of 17%. It thus seems that while the NYPD is indeed focusing on higher hit rate stops, there is still considerable room for improvement by rigorously optimizing the policy.
Table 3

Values for the five non-zero coefficients for stop circumstances in the reduced model, and the corresponding score for the heuristic model. In deciding whether to make a stop, officers add the relevant heuristic scores and check whether the sum exceeds an area-specific threshold.

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Value</th>
<th>Heuristic score</th>
</tr>
</thead>
<tbody>
<tr>
<td>suspicious object</td>
<td>2.59</td>
<td>3</td>
</tr>
<tr>
<td>sights and sounds of criminal activity</td>
<td>0.67</td>
<td>1</td>
</tr>
<tr>
<td>suspicious bulge</td>
<td>0.58</td>
<td>1</td>
</tr>
<tr>
<td>witness report</td>
<td>0.18</td>
<td></td>
</tr>
<tr>
<td>ongoing investigation</td>
<td>0.08</td>
<td></td>
</tr>
</tbody>
</table>

3.5. Heuristic stop strategies. Though the strategy of conducting only the ex-ante most efficient stops is conceptually simple, it is admittedly not straightforward to implement in practice. Officers cannot simply evaluate a complex statistical model in their heads when deciding whether or not to stop a suspect. Further, it seems unlikely that police departments would adopt an opaque machine learning model to inform stop decisions. To address these difficulties, we draw on a large body of work which has found that simple, transparent, and interpretable heuristics often work as well as complex statistical models (Czerlinski, Gigerenzer and Goldstein, 1999; Gigerenzer and Goldstein, 1996; Lovie and Lovie, 1986; Ustun and Rudin, 2014).

To start, as in Eq. (1), we model the likelihood of recovering a weapon in a CPW stop via logistic regression. This time, however, we use only the 18 stop circumstances officers already consider (listed in Table 1, excluding the two “other” categories), indicator variables for each of the 77 precincts, and indicator variables for the three location types (public housing, transit, and “neither”); we do not include interactions. Also, since each of the stop reasons is indicative of criminality, we constrain the corresponding 18 coefficients to be non-negative. We thus fit the reduced model,

\[
P(y_i = 1) = \text{logit}^{-1} \left( \sum_{j=1}^{18} \alpha_j a_{j,i} + \sum_{k=1}^{77} \beta_k b_{k,i} + \sum_{\ell=1}^{3} \gamma_\ell c_{\ell,i} \right)
\]

with the constraint \( \alpha_j \geq 0 \), where \( a, b, \) and \( c \) are indicator variables for stop reason, precinct, and location type, respectively.\(^{10}\) Only 5 of the 18 stop circumstances were found to have positive weight (the remaining were identically zero): (1) suspicious object; (2) sights and sounds of criminal

\(^{10}\)Model coefficients were computed with the penalized package in R (Goeman, 2010).
activity; (3) suspicious bulge; (4) witness report; and (5) ongoing investigation. Notably, all five circumstances are directly tied to criminal activity, and the more subjective conditions (e.g., “furtive movements”) drop out of the model.

For any CPW stop, Eq. (2) estimates the probability the stop results in recovery of a weapon. It thus suggests the following family of stop rules. For any threshold $p > 0$, stop an individual if: (1) the individual would have been stopped under the usual stop-and-frisk practice; and (2) the probability of recovering a weapon, as estimated under the reduced model, is at least $p$. The first condition is critical since the model is trained only on stops that in fact occurred, and so it may not generalize to the population at large. One can thus think of this as a two-step procedure, where an officer first relies on his or her training and intuition to determine whom to possibly stop, and then checks whether the model-estimated probability exceeds a pre-specified threshold.

The reduced model in Eq. (2) is more transparent and interpretable than the complete statistical model in Eq. (1), but it is still cumbersome to evaluate on the fly. We simplify the expression in two steps. First, to implement the stopping procedure described above, we need not compute the actual probability of recovering a weapon, but can instead compute a stop score.
that is monotonically related to the probability. We consequently ignore the logistic transformation, and simply check whether the sum of the relevant coefficients exceeds a given threshold. Second, we round the five coefficients for the stop circumstances to the nearest integer (listed in Table 3); we leave the precinct and location-type coefficients unaltered.\footnote{We tried several different rescaling and rounding schemes and obtained similar results.} This procedure results in only three non-zero coefficients for the stop reasons: suspicious object (value = 3), sights and sounds of criminal activity (value = 1), and suspicious bulge (value = 1). Letting $\tilde{\alpha}_j$ denote the rounded coefficients, and reindexing $\tilde{\alpha}_j$ so that the first three values correspond to the non-zero values, the score $S_i$ for the $i$-th stop is,

$$S_i = 3 \sum_{j=1}^{3} \tilde{\alpha}_j a_{j,i} + \sum_{k=1}^{77} \beta_k b_{k,i} + \sum_{\ell=1}^{3} \gamma_{\ell} c_{\ell,i}. \quad (3)$$

Now, suppose we have selected a stop threshold $T$, then the stop condition $S_i \geq T$ is equivalent to

$$\sum_{j=1}^{3} \tilde{\alpha}_j a_{j,i} \geq T - \sum_{k=1}^{77} \beta_k b_{k,i} - \sum_{\ell=1}^{3} \gamma_{\ell} c_{\ell,i}. \quad (4)$$

The left-most sum is a function of the three stop reasons, and $T_r$ is an area specific threshold that depends only on the precinct and location-type. Thus, to quickly and rigorously assess the likelihood a potential stop will lead to the recovery of a weapon, officers simply need to add at most three small integers, and check whether the sum exceeds a fixed threshold $T_r$ for the area they are patrolling. Since officers commonly patrol only a single area during a shift, this procedure is particular straightforward to carry out in practice.

To implement this scheme, one still needs to select a stop threshold $T$, which in turn determines area-specific thresholds $T_r$. The higher the threshold, the fewer people stopped, but also the fewer weapons recovered. Figure 6(a) plots this tradeoff. For various thresholds $T$, we compare the percent of individuals stopped under the heuristic model to the percent of weapons recovered, indicated by the open circles. For comparison, we also plot the tradeoff under the complete model given by Eq. (1), and the reduced model given by Eq. (2). The figure shows that performance of the heuristic model is virtually indistinguishable from the reduced model. Moreover, while the heuristic model is not quite as effective as the complete model, it still performs surprisingly well. For example, with the heuristic model, one recovers...
50% of weapons by making just 7.5% of stops; in comparison, 6% of stops are required under the complete model. We note that if such stop rules were ultimately adopted, the model would likely require periodic updating since changes in officers’ behavior could affect model performance.

We conclude by examining the area-specific thresholds $T_r$. Figure 6(b) shows the area thresholds for a policy that recovers 50% of weapons, where higher thresholds are indicated by lighter colors.\textsuperscript{12} A comparison with Figure 2 reveals that known high-crime areas also have relatively high thresholds. That is, in high crime areas, the policy requires a higher number of indicators of criminal activity to justify a stop. This counterintuitive observation stems from the disproportionately large number of low efficiency CPW stops in high-crime neighborhoods. It could be the case that indicators of criminal behavior (e.g., “suspicious object”) may not be as predictive in high-crime neighborhoods as in low-crime areas. Alternatively, it could be that officers patrolling high-crime areas simply have a lower threshold for considering an object “suspicious” than officers in safer areas. Finally, we note that many areas, particularly in lower Manhattan, have stop thresholds of zero. According to our stop rule, officers in such areas would thus stop suspects according to their usual procedures, without additionally checking whether the stop score exceeded a threshold.

4. Discussion. By estimating the ex-ante efficiency of stops, we were able to investigate claims that stop-and-frisk violated two constitutional protections: first, that individuals were detained without legal basis, in violation of the Fourth Amendment; and second, that the tactic was not applied in a race-neutral manner, in violation of the Fourteenth Amendment. Regarding the former claim, we find that in a substantial fraction of instances where a suspect was stopped for suspicion of carrying a weapon, it was in fact ex-ante very unlikely a weapon would be found on the individual. In particular, in 44% of such cases, the likelihood of finding a weapon was less than 1%. Though it is an open question beyond the scope of this paper to determine what constitutes reasonable suspicion in this context, our result raises concerns that the legal standard is often not met. Regarding the latter claim, we show that while the adverse effects of stop-and-frisk on blacks and Hispanics are largely attributable to heavy use of the tactic in high-crime, predominately minority areas, there still appears to be an element of racial bias. It is unclear whether this bias derives from racial prejudice or spurious

\textsuperscript{12}For ease of visualization, for each precinct we plot the average threshold over the different location types in the precinct, where the terms in the average are weighted by the number of stops in each location type.
statistical reasoning by officers. Regardless of the underlying cause, however, blacks and Hispanics are subject to stops conducted on the basis of less suspicion than similarly situated whites. Finally, we show that by reducing the number of low hit rate stops—which disproportionately affect minorities due to both highly localized tactics and racial bias—one can still recover most weapons while bringing more racial balance to stop-and-frisk.

In our primary analysis, we considered only instances in which an individual was stopped for suspicion of criminal possession of a weapon, the single most frequently recorded suspected crime, constituting one-fourth of stops. Our results, though, are not just restricted to weapons possession. In particular, for stops where the suspected crime is drug related—including criminal possession and sale of marijuana and other controlled substances, comprising 10% of all stops—11% of stops have less than a 1% ex-ante likelihood of contraband being found on the detained individual, and 52% have less than a 5% chance. We likewise find that the disparate effect of such drug-related stops on blacks and Hispanics is due to a combination of highly local policing strategies and racial discrimination.

A possible objection to our approach is that even for CPW stops, recovering weapons is not the only—or perhaps not even the primary—goal of the police. Officers, for example, may simply consider stops a way to advertise their presence in the neighborhood, or a means to collect intelligence on criminal activity in the area, regardless of how many weapons are directly recovered. In fact, stops conducted for these alternative motives could quite plausibly deter individuals from carrying weapons and might lead to information helpful in solving cases, both of which presumably would lower the incidence of violent crime over time. In the instances we consider, however, the explicitly stated reason for a stop is suspicion of criminal possession of a weapon, not one of various other reasons that may or may not withstand legal or public scrutiny, and so it seems most natural to consider whether individuals were in fact likely to be carrying weapons. Moreover, as we have previously noted, simply because a strategy may be effective does not make it legal. For instance, searching a suspect’s home before a warrant is issued may be an effective way to collect evidence, but is nonetheless illegal except under exigent circumstances.

A related worry is that “criminal possession of a weapon” is a catchall category for a variety of criminal offenses, and so by focusing on whether a weapon was found, we underestimate the value of a stop. Addressing this issue, we observe that our results are qualitatively similar if we instead use arrests as the outcome variable, mitigating cause for concern.

Looking forward, our results show that though stop-and-frisk does suffer
from potentially serious problems, the tactic is not beyond repair. By fo-
cusing on the relatively small number of high hit rate situations—situations
that can be reliably identified via statistical analysis—one may be able to
retain many of the benefits of stop-and-frisk for crime prevention while miti-
gating potential constitutional violations. This observation has the potential
to improve not only New York City’s stop-and-frisk program, but could also
aid similar policies throughout the country.

References.
from the Boston Police Department. The Review of Economics and Statistics 91 163-
177.
Ayres, I. (2002). Outcome tests of racial disparities in police practices. Justice research
and Policy 4 131–142.
of political economy 385–409.
neural networks 17.
Czerlinski, J., Gigerenzer, G. and Goldstein, D. G. (1999). How good are simple
heuristics?
A Large Sample Field Experiment. The Review of Economics and Statistics 96 119-134.
department’s “stop-and-frisk” policy in the context of claims of racial bias. Journal of
the American Statistical Association 102.
Gigerenzer, G. and Goldstein, D. G. (1996). Reasoning the fast and frugal way:
models of bounded rationality. Psychological review 103 650.
Biometrical Journal 52 70–84.
Trigger Periods of Increased Discrimination. Under review.
Lerman, A. E. and Weaver, V. (2014). Staying out of Sight: Concentrated Policing and
Local Political Action. Annals of the American Academy of Political and Social Science
651 6-21.
Lovie, A. D. and Lovie, P. (1986). The flat maximum effect and linear scoring models
DAVIS v. CITY OF NEW YORK (2013b). No. 10 Civ. 0699.
APPENDIX A: MODEL CHECKS

Fig 7. In panel (a), stops are binned by model-predicted hit rate to the nearest percent. A point, sized by the number of stops, is plotted for each bin, comparing the model-predicted hit rate to the actual (empirical) hit rate. In panel (b) stops are binned by precinct, race, and gender. For each bin we plot a point, sized by the number of stops and shaded according to race, comparing the model-predicted hit rate to the empirical hit rate. In panels (a) and (b) the plotted points lie close to the dashed 45 degree line, indicating that our model predicts well over the entire range of hit rates and for interactions between important features. Finally, Panel (c) shows that for various values of the features age, race, and gender, there is little difference between the model-predicted hit rate and the empirical hit rate.
APPENDIX B: AN ALTERNATIVE STOP HEURISTIC

In our primary analysis, we constructed stop heuristics that relied on area-specific thresholds. However, for social, political or legal reasons, one might prefer a policy that applies a uniform threshold across the city. To construct such a policy, we follow the procedure outlined in Section 3.5, but omit the precinct and location-type covariates in Eq (2). Figure 8 plots the performance of this model. While there is again little difference between the reduced and heuristic models, both fare considerably worse than when location information is included.

![Graph showing the estimated percentage of weapons recovered as a function of the number of stops conducted (note that the x-axis is on a log scale) for various models.]

**Fig 8.** The estimated percentage of weapons recovered as a function of the number of stops conducted (note that the x-axis is on a log scale) for various models.

**Sharad Goel**  
Management Science and Engineering  
Stanford University  
475 Via Ortega  
Stanford, CA 94305  
USA  
E-mail: scgoel@stanford.edu

**Justin M. Rao**  
Microsoft Research  
641 Avenue of the Americas  
7th floor  
New York, NY 11249  
USA  
E-mail: justin.rao@microsoft.com

**Ravi Shroff**  
Center for Urban Science and Progress  
New York University  
1 MetroTech Center, 19th Floor  
Brooklyn, NY 11201  
USA  
E-mail: ravi.shroff@nyu.edu