Assessing Disparities in MVPD Stops: Summary Report

Alex Stephenson

Introduction and Project Context

As part of its efforts to improve police-resident engagement and continue a meaningful dialogue around racial justice and equity issues[1], the City of Mountain View and the Mountain View Police Department (MVPD) partnered with researchers at Stanford University and the University of Michigan to hire an outside research fellow to examine potential disparities and bias in policing. In collaboration with stakeholders, the research focuses on testing the existence of racially discriminatory behavior within MVPD. The report analyzes MVPD administrative data on all traffic stops conducted by the Police Department from 2014-2020. Traffic stops are one of the most common ways the public interacts with a police officer.[2] Nationwide, there are concerns that racial bias plays a role in officers' decision-making.[3] Most traffic stops are officer-initiated. Officers have the discretion to search both the driver and vehicle for drugs, weapons, or other contraband items if there is a "reasonable suspicion" of serious criminal activity.[4] While traffic stops are a subset of police activity, they are also the main activity for which data of interest on perceived race, gender, and age of community members are consistently reported over time in many jurisdictions, including Mountain View.

As the primary questions of interest are whether there are observed disparities and biases by race, it helps define what is meant by these two terms explicitly. Racial disparities refer

[1] Forestieri, K. 2020. "In the wake of protests, Mountain View residents demand police reform." *Mountain View Voice*
[2] Langton, L., and M. Durose. 2013. "Police Behavior During Traffic and Street Stops, 2011. *Bureau of Justice Statistics.*
[3] Pierson *et al.* 2020. "A large-scale analysis of racial disparities in police stops across the United States" *Nature Human Behavior* 4, 736–745
[4] Terry v. Ohio 392 U.S. 1 (1968)

to outcomes that differ by race or ethnicity. For example, if poverty levels are different between Black and White residents of Mountain View, there would be a racial disparity in poverty. Note that, by definition, a disparity may occur without assuming that the group characteristic is the reason for the observed difference. Racial bias refers to a difference in behavior attributable to another person's race or ethnicity. For example, consider a mortgage lender who provides a loan to a White family but does not provide a loan to a family of Nicaraguan descent. If the lender bases that decision entirely or partially on race, that behavior would constitute racial bias. Racial profiling behavior is a subset of racially biased behavior.

The research project evaluated disparities and questions of racially biased behavior at the patrol officer-motorist level. The goal is to assess whether police behavior during these encounters would have differed if the motorist had belonged to a different racial group, holding constant the circumstances of the traffic stop. As the primary unit of analysis, focusing on officer-motorist interactions provides leverage to discuss comparable situations in which the police observe minority and majority motorists. With appropriate data, and appropriate assumptions, I can estimate the counterfactual police behavior that would have occurred if the motorist in question had been replaced with a member of a different racial group.[5] While this approach provides for a valid counterfactual necessary to make causal inferences, it necessarily holds fixed pre-encounter decisions by both the police department and the community. Readers should keep this scope condition in mind. Before gathering traffic stop data, I interviewed command staff, data specialists, and patrol officers to understand how the Department approaches traffic stops and how MVPD's traffic stop policy is seen to improve traffic safety. These interviews also provided information on the Department's data collection processes. Beyond these interviews, the Department arranged for eight hours of ride-alongs with officers. The ride-alongs occurred during daylight hours and after dark, providing personal support for the proposition later tested empirically that darkness affects officers' ability to see the identity of motorists stopped at night. These qualitative findings were critical to understanding Department data

[5] Knox, D. Lowe, W. Mummolo, J. 2020. "Administrative Records Mask Racially Biased Policing." *American Political Science Review.*

and collection processes. Without doing this work, I would not have been able to feel confident in the nature of data collection by officers. I also would not have appreciated the circumstances and challenges officers face conducting their jobs.

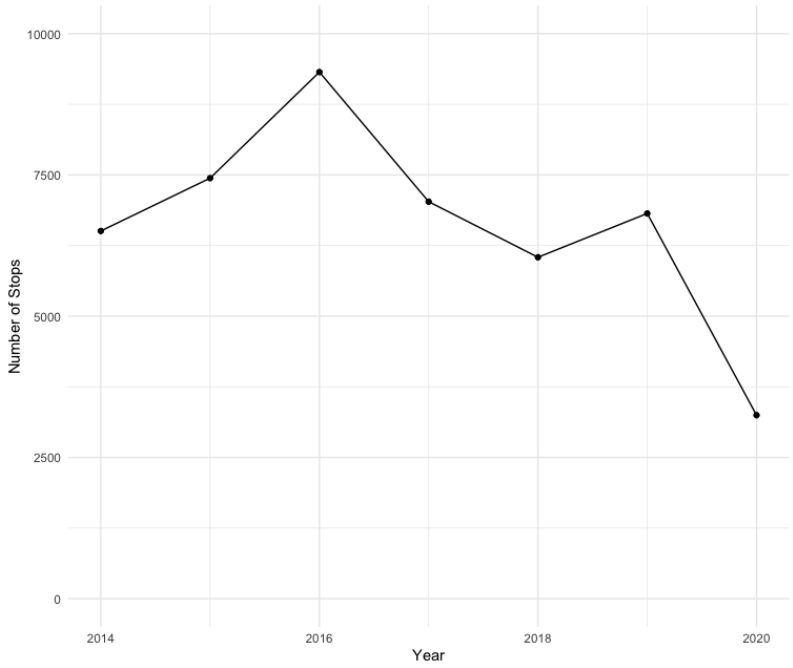
Project Goals

The project focuses on three overarching questions:

1. What is the nature of any observed disparities in traffic stops by the MVPD?

To assess potential sources of bias, I used records on all traffic stops made by the MVPD from 2014-2020. Figure 1 plots the total number of traffic stops per year in the dataset.

Figure 1: Total Traffic Stops Per Year conducted by MVPD



The dataset includes information on:

- When and where the traffic stop was collected,
- The reason for and outcome of the stop,
- The officer conducting the stop,
- The perceived race, gender, and age of the driver stopped and

- Information on whether an officer conducted a search and whether it was successful. By success, I mean that the search found illegal items, such as narcotics or weapons.

I conducted discrimination tests standard in the literature on discrimination.[6] I look at two different aspects of the officer-motorist encounter: the search decision and the stop decision. The tests related to the search decision are referred to as benchmark tests and outcome tests. The test that evaluates the stop decision is known as the veil of darkness test. Each one of these tests, if their assumptions are met, can determine that disproportionality exists between two groups. They cannot provide evidence as to *why* there is a disproportionate outcome. No matter how much data is collected on stops conducted by officers, data is not generated on encounters that have not occurred. A difference in benchmark rates for stops may simply reflect unobserved differences between drivers. Drawing causal inferences about the mechanism associated with disparities from observational data is a challenging problem.[7]

Benchmark tests compare the rates at which two groups are searched. Suppose group A is searched more often than group B. In this example, the difference in search rates may be the result of bias against group A. These figures are currently available in the Department's annual reports.

Outcome tests are based on the success rate, also known as the hit rate, of search decisions and posit that discrimination can be detected if searches of one group yield contraband less often than searches of another.[8] The intuition of this test is that even if a group is more

[6] Knox, D and J. Mummolo. 2020. "Toward a General Causal Framework for the Study of Racial Bias in Policing." *Journal of Political Institutions and Political Economy* 1(3); Neil, R., and C. Winship. 2019. "Methodological Challenges and Opportunities in Testing for Racial Discrimination in Policing." *Annual Review of Criminology.* 2; Ridgeway, G., and J. MacDonald. 2010. "Methods for Assessing Racially Biased Policing" *in Race, Ethnicity, and Policing: New and Essential Readings.* NYU Press. It is important to note that there is no single approach for identifying racial profiling or racial bias in policing. Each test conducted has methodological strengths and weaknesses. Additional discussion of these strengths and weaknesses are in the appendix.

[7] See Imbens, G. and D. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences.* Cambridge University Press for an extended discussion of deriving causal inferences from observational data

[8] Becker, G. 1957. *The Economics of Discrimination.* University of Chicago Press.

likely to carry contraband, in the absence of discrimination, that group should still be found at have contraband at the same rate as a searched baseline group.

To further assess potential bias in stop decisions, I use the "veil of darkness" test developed by Grogger and Ridgeway (2006).[9] This test uses the discontinuity created by daylight-saving time to compare the distribution of drivers stopped in the evenings immediately before daylight-saving time begins to the distribution after daylight-saving time begins when it is light at the same time of day during which it had previously been dark. If a share of the group of drivers is stopped less often when it is dark relative to when it is light, that suggests that the group is stopped more often during daylight hours because of their race.

Each of these tests attempts to approximate a type of experiment. Ideally, a researcher would like to do a randomized experiment to assess whether observed disparities constitute evidence of racial bias. Such an experiment is not feasible in Mountain View.[10] A benchmark test imagines a world in which racial groups' propensity for carrying contraband or being stopped is equal to the share of a population. This is a strong assumption about the world, and unlikely to ever hold empirically. I am most confident in

[9] Grogger, J and G. Ridgeway. 2006. "Testing for racial profiling in traffic stops from behind a veil of darkness. *Journal of the American Statistical Association*, 101, 878-887

[10] The ideal experiment to test the effect of race on police officer decisions would randomly assign white and non-white community members to drive through pre-existing MVPD beats and act in some prescribed way, such as purposely speeding or driving with a vehicle infraction such as a broken taillight. Under these conditions, we could observe the rate at which MVPD officers detain and search members of each racial group to obtain a valid estimate of racial bias among MVPD officers. While it is true in principle that the ideal randomized experiment solves the selection problem, such an experiment is not feasible in Mountain View for both ethical and experimental concerns. The ideal experiment requires confederates to violate laws in a proscribed way and to open themselves up to potential harm (Phillips 2021; Teele 2014). The process-related downstream effects of such an intervention can harm individuals and groups.

While the ethical concerns are enough in my view to not propose running a field experiment to test racial bias, they are not the only reason an experiment is not feasible. An experiment on traffic stops in Mountain View would likely violate a core assumption of causal inference known as the Stable Unit Treatment Value Assumption or SUTVA (Imbens and Rubin 2015). The first component in this assumption is commonly referred to as no interference between units. No interference is an assumption that the treatment applied to one unit does not affect the outcome for other units. Such an assumption is highly implausible to hold in a police department. In Mountain View, there are a limited number of police officers assigned to different shift schedules. An influx of similar vehicles with similar infractions is sure to be noticed by officers, who may adjust their behavior in response to the intervention. Standard adjustments in experiments to the problem of interference cluster units at the highest level for which no interference between treatment units is plausible. For the MVPD, that would be at the department level, which breaks the experiment because there is no longer a control group.

the results from the veil of darkness tests because it most closely resembles the experiment that I would like to conduct if such an experiment were possible. This experiment would randomly assign police officers to be unaware of motorists' race.

The limitation for every test is that I do not know much about the broader population of individuals on the road or the overall propensity for carrying contraband within Mountain View. Because the assumptions I make on the data generating process are only about the qualitative relationship between patrol officers' behavior and motorists, I can compute only qualitative tests rather than a quantitative measure about the specific extent of racial bias. Data limitations will also affect the precision of these qualitative insights. These are general limitations of any observational study of discrimination and deserve to be mentioned as a result.

2. Are any observed disparities indicative of potential bias in the MVPD?

Based on the tests specified, I did not find conclusive evidence of disparities in MVPD's stop behavior based on our analyses of the Department's data. I find no clear evidence from the veil of darkness tests that the MVPD stop decisions from 2014-2020 are discriminatory in the window for which the test provides relevant information.[11]

Table 1 provides a summary of the test conducted by sub-group. In all cases, the comparison is between a minority group and a majority group in Mountain View. The estimate describes the differences in the composition of stopped drivers before daylight and after dark after adjusting for clock time. If the point estimate is **negative *and* statistically significant**, it would suggest discrimination against a given group.

[11] This finding holds even when data from 2020 is excluded. I conducted the same veil of darkness test without data from 2020 due to worries that traffic patterns in Mountain View could be radically different because of the COVID-19 pandemic. The findings are also consistent with different functional form specifications including logit and probit models.

Table 1: Regression Estimates for Veil of Darkness Comparison Tests

Outcome	Estimate	Standard Error	p-value	Confidence Interval
Black/White	0.024	.029	.406	[-0.033, 0.081]
Hispanic/White	-0.008	.018	.634	[0.044, 0.027]
AAPI/White	-0.005	.029	.860	[-0.062, 0.052]

For the data range in Mountain View, I do not find any statistically significant result at the 95% level. In all cases, the confidence interval includes zero.[12] Given the wide range, I interpret the results of these tests to suggest that this test does not provide clear evidence of biased stop decisions by the MVPD.

We find mixed evidence of potential disparities in outcome tests depending on the year and relevant group comparison. Figure 2 shows the number of stops by year broken down by whether an officer conducted a search. As can be seen from the figure, searches are a rare phenomenon in the dataset of traffic stops. Only 4% of all traffic stops result in any kind of search. Because of the relative paucity of searches conducted, any individual search decision can be quite influential.[13]

[12] The variances of these estimates are large, suggesting that while there are several thousand observations within each subgroup (black/white: 1253 observations; hispanic/white: 3159 observations, AAPI/white: 1468 observations), these totals are too low to provide tight precision around the point estimate. These results may suggest that the data may not have the power to detect a precise estimate.

[13] In addition to the small number of searches in a year, some of the search types are mandated by Department policy, further reducing the number of searches within an officer's discretion. For example, an officer will conduct a search of an individual upon arrest. This search may or may not turn up contraband but is not a type of search that an officer can choose not to conduct.

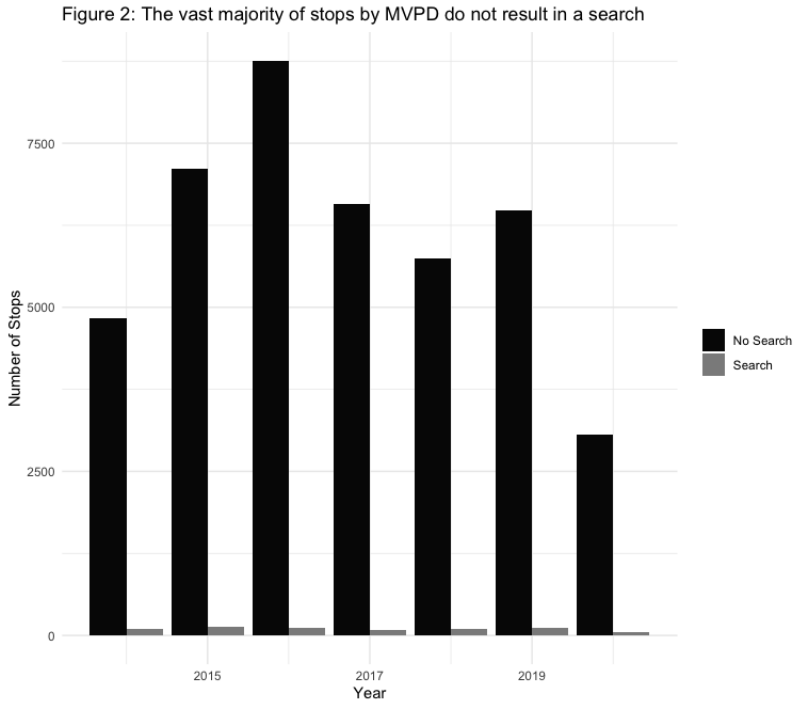
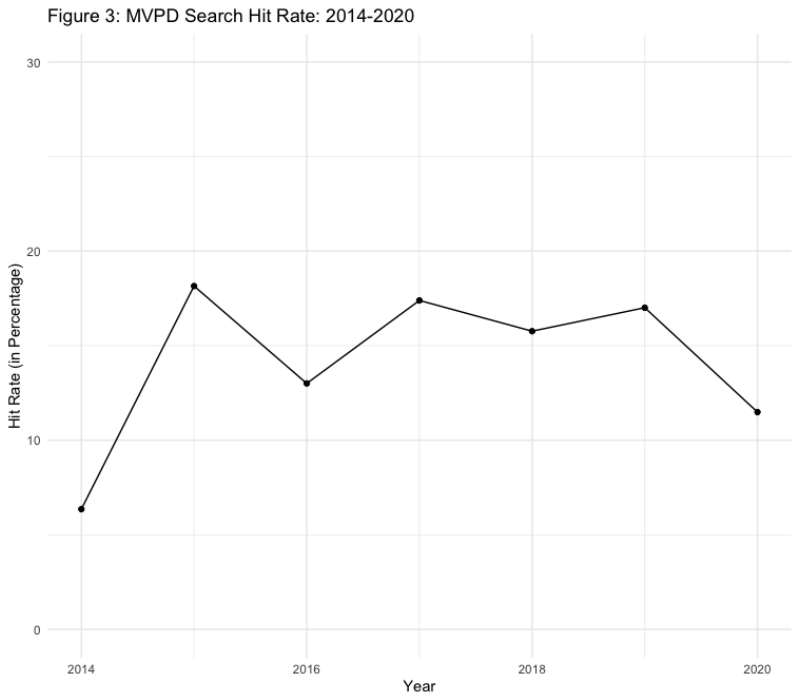


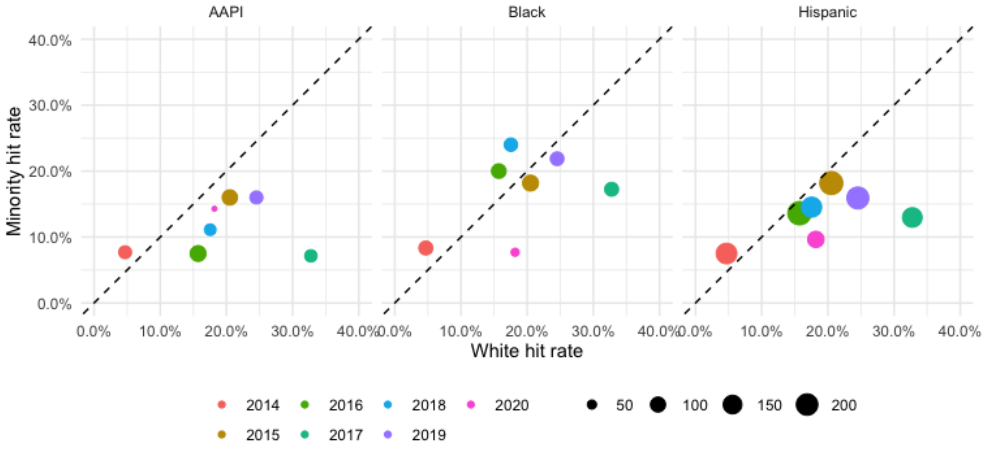
Figure 3 shows how the Department's overall hit rate, which is the rate of searches for which contraband was found, fluctuates year over year.



To interpret this graph, consider that a Hit Rate of 20% would imply that one out of every five searches conducted found contraband. There is no external benchmark of what would be considered a "good" hit rate for the outcome test. Instead, the test considers the differences between groups.

Figure 4 provides information on the hit rates for the same groups evaluated in the veil of darkness test. In each graph, there is a comparison between the hit rate of a minority motorist group and the hit rate for white motorists. Points above the dashed line represent years for which the minority hit rate is higher than the white hit rate. Points below the line indicate years for which the minority hit rate is lower than the white hit rate. Each point corresponds to a year in the dataset. The points are sized by the combined number of searches between groups.

Figure 4: Hit Rates by Year and Race of Driver



The results find some evidence to suggest that the hit rate is not consistent across comparison groups, but there is not a consistent pattern across groups. However, these results should be interpreted with extreme caution due to the low number of searches, suggesting that a few searches strongly drive the hit rate comparisons. Another concern is that the outcome test is known to suffer from a statistical problem called infra-marginality.[14] In brief, the hit rate is an average which is a function of an officer's search threshold (how likely they are to search a motorist) and a *signal distribution* (what the officer observes that may change whether they search a motorist). Since the hit rate depends on both factors, infra-marginality here means that it is possible to find lower hit rates and higher search rates for a group even in the absence of discrimination. Alternative methods have been proposed to deal with this challenge, but those methods require substantially more data than exists in Mountain View to produce valid estimates.[15]

[14] Simoiu, C. Corbett-Davies, S., and S. Goel. 2017. "The problem of infra-marginality in outcome tests for discrimination." *Annals of Applied Statistics* 11.
[15] See footnote 15.

3. What information should be collected going forward by the MVPD?

Most data collection limitations relate to a lack of information on contacts made outside of traffic stops. As mentioned, the Department did not consistently keep standardized information on stops of individuals other than traffic stops that did not result in a criminal citation or arrest. Data also was only consistently collected on the driver of the vehicle, as opposed to all passengers. The recommendations for improving the system are the same as the needed process changes to fully comply with the Racial Identity and Profiling Act (AB 953). Due to the MVPD's size, data collection for RIPA compliance must begin on January 1, 2022. The MVPD has a standing committee currently working to implement RIPA requirements. MVPD is one of over 400 policing agencies to start its implementation in 2022.[16] Any change in collection poses implementation challenges. RIPA significantly adds to the complexity of a given stop for an officer. The Department should develop a review process to ensure that officers have appropriate support to comply with new policies and procedures. Additional training for officers on new policies, as well as continued guidance and support for leadership, will be critical for successful compliance.[17] The additional changes from current practice are:

- Information on pedestrian stops. Currently, the Department does not collect extensive data on pedestrian stops like it does traffic stops. In addition, there are no systems that require officers to fill out detailed perceptual information on individual stops before an officer can be considered available for calls. This contrasts with departmental systems for traffic stops. At present, a patrol officer who has conducted a traffic stop cannot be dispatched to a subsequent call for service until they fill out the traffic step demographic details that this report relies upon. Beginning in 2022, MVPD will collect information regarding the stop, including reasons for stop and actions taken, as well as information regarding the officer's

[16] Racial Identity and Profiling Advisory Board. 2021. "Annual Report." p. 22
[17] Mummolo, J. 2017. "Modern Police Tactics, Police-Citizen Interactions,
and the Prospects for Reform" *Journal of Politics* 80(1) provides evidence from New York that affirms the common sense view that officers are highly responsive to managerial directives and that appropriate support from command staff is necessary for successful change in Department procedures.

perception of the person stopped, including perceived race or ethnicity, age, and perceived gender.

- Additional information on perceptions of individuals stopped. At present, the Department does not collect standardized information on perceived disability or perceived English fluency. Both will be mandatory data elements to be collected going forward. Such elements help understand whether officers respond to certain individuals differently than others. The Department should also look into using vendors that can allow for customized questionnaires in case priorities for data collection change or in response to community feedback. There is a trade-off between information and patrol officer time, especially in a community where the number of stops is comparatively small.

The City and MVPD should explore additional ways to make its data public and easily accessible for residents. Whenever possible, raw, de-identified data should be available to and accessible by the public, which helps build trust and improve perceptions of legitimacy (Center for Police Equity 2020). RIPA reporting requirements will ease this burden for the city. The Department should determine appropriate ways to provide necessary context about what data does and does not say. Providing a codebook is the best way to accomplish this.

In addition, public bodies such as the Public Safety Advisory Board (PSAB) should continue to evaluate Department data. The PSAB serves as a forum for public discussion of public safety matters and acts as a liaison between the community and the Department. In addition to the Department's current Annual Report, the PSAB could suggest different data collection and reporting, including some of the tests, run in this report. RIPA data has already been subject to some of these same tests at the statewide level.[18] In particular, outcome tests for contraband are tests that can be run on data collected yearly. Other tests, such as the veil of darkness, require more years of data collection to have enough observations to be run.

[18] Racial Identity and Profiling Advisory Board. 2021. “Annual Report.”

Future data collection and research efforts could also be made to understand community concerns and how the community uses police services beyond traffic stops. RIPA data collection may provide opportunities for the city and the community to map contacts between patrol officers and the community to understand what services the community regularly request. This could open space for research into how well that service is being provided.

A final research limitation is that no matter how much data is collected on stops conducted by officers, data is not generated on encounters that have not occurred. This problem is particularly challenging for benchmark tests. The population of interest is the population of motorists on the road, which is unlikely to be the same as the Census population of Mountain View. The City could choose to explore ways to survey drivers on the road at different days and times of the week and at different times of the year. Doing so would provide a better picture of the population of motorists, giving more credibility to benchmarking tests.[19] I do not make this a formal recommendation due to potential community concerns and likely cost, as such collection is often labor-intensive.

[19] Lange, J. E., Blackman, K. B., & Voas, R. B. (2005). Testing the racial profiling hypothesis for seemingly disparate traffic stops on the New Jersey Turnpike. Justice Quarterly, 22, 193–223