RE: ACLU Matter vs. - REF# 1340012232
Analysis of Coded ISR Narratives
January-June 2016
For Input to
Hon. Arlander Keys' (Ret.)
First Period Report

REVISED FINAL Technical Report

Ralph B. Taylor & Lallen T. Johnson

DATE: March 17, 2017

Acknowledgments. The authors appreciate helpful input from Sharad Goel, Aziz Huq, Jens Ludwig and Justin McCrary on the design of the present analyses. All the material herein represents only the views of the authors and does not reflect the views or policies of any other organization including the City of Chicago, the Chicago Police Department, or ACLU-Illinois, or their experts. Any mistakes or misinterpretations herein are solely the authors.
Declaration of Conflicting Interests. The authors declare no potential conflicts of interest with respect to the research, authorship and/or dissemination of this work.
Funding. The authors disclose receipt of the following financial support for the research and authorship of this work: Authors were paid by the City of Chicago as part of the above referenced agreement to provide statistical input to the Hon. Arlander Keys (Ret.).

1 TABLE OF CONTENTS

2 INTRODUCTION TO REVISED VERSION

Comments by the Parties and their experts on the initial version of this report led to modifications that appear in this version. The major modifications include the following.

1. Clarifying the key question tested by each analysis.
2. Adding a descriptive table showing the average score on each outcome for each of the three ethnoracial groups examined using appropriately weighted data. These weighted descriptive results reflect the full set of ISRs capturing investigatory stops for the reporting period.
3. Clarifying the three levels of scrutiny that can be applied to ethnoracial differences on each outcome: gross impact, net impact, and statistically significant net impact of a race or ethnicity predictor.
4. Clarifying how geographic variation on each outcome was presented.
5. Discussing the partialling fallacy as a potential limitation when interpreting net impacts of race or ethnicity variables.

3 FOR THE NON-TECHNICAL READER: FAQ

This section asks and answers questions that the non-technical reader might have about this report. It simultaneously guides the non-technical reader to findings that might be of most interest to him or her. Even technical readers might benefit from scanning the questions and answers listed here.

3.1 PURPOSE

Q: What is the **purpose** of this report?

A: This report looks at two features of stops: the *legal basis* for the investigatory stop itself, and the *legal basis* for a pat down, if a pat down occurred. For each of these features, the key question is whether civilian race or ethnicity had an impact on that basis.

3.2 SCORING STOPS AND PROTECTIVE PAT DOWNS

Q: How was that **legal basis determined**?

A: The Consultant made that judgment. He has legal expertise making similar decisions as a federal judge. For each sampled investigatory stop report (ISR) he independently reviewed narratives written by the officer and other key fields in the report. He determined in each case whether the narratives adequately articulated facts supporting the idea that the police officer's decision to stop the subject of the ISR was based on "reasonable articulable suspicion" (RAS) to suspect that a crime had been, or was about to be committed (stop legal basis). If the Consultant determined that the police officer articulated RAS for the stop, he coded the stop as "good" and assigned a corresponding numeric code. If he determined that the officer failed to articulate RAS for the stop, it was coded as "bad" and a different numeric code was assigned. These two numeric codes became the outcome variable that statistical models sought to link to race and ethnicity.

The same process was used for assessing the protective pat down, when applicable. The Consultant reviewed narratives for an independent set of facts giving the police officer an additional "reasonable suspicion" that the subject possessed or had access to a weapon or firearm, creating danger to the officer or bystanders nearby (protective pat down legal basis).

Again, pat downs like stops could be coded as "good" or "bad," with corresponding numerical codes.

In short, the legal bases for stops and pat downs involved a coding process in which the Consultant independently reviewed the factual content, documented by the police officer, in the narrative remarks section of the submitted ISR, with due consideration given to the boxes checked by the officers in each ISR reviewed.

If the stop involved a search, the Consultant used a different legal standard to gauge whether that search was "good" or "bad." This report, however, will not analyze the legal basis of searches because the number of "bad" searches was extremely small.

3.3 Using samples

Q: Were ***all*** **investigatory stop records for the period January-June, 2016 examined** by the Consultant?

A: No. We asked the computer to draw a random sample of reports from each of three groups: stops involving a White non-Hispanic civilian, a Hispanic civilian, and a Black non-Hispanic civilian. We asked for the same number of sampled reports from each of the three groups.

Q: Is it important that the **samples** you drew were **random**?

A: Yes. Random samples have sampling error in them. They do not perfectly capture the full set of records for each of the three groups because we have just a subset of the records for each group. But because the sampling error is random it will not distort the picture we have of each group. It will give us an unbiased picture of each group.

Q: Can the results from these three samples, when put together, **reflect the full set of stop records** for these three groups altogether?

A: Yes, if we take two things into account. First, we must remember that when we sampled we sampled a bigger fraction of some group's records (White non-Hispanics), and a smaller fraction of other group's records (Black non-Hispanic). So when we put all three groups in the sample back together to reflect the full set of stops we count a group's records more if we took a smaller fraction from that group in the first place, and less if we took a bigger fraction. That way when we put all three groups together the proportional contribution of each group roughly matches what we find in the full set of records investigatory stop records. Second, we remember that the samples have error so we take that into account when inferring back to the full set of records.

3.4 Race and ethnicity impacts on stop basis

Q: What do you mean by race or ethnicity "had an **impact**"?

A: We, the authors, are thinking about "influence," or "impact" ***using a social science framework. That's because we're social scientists, not legal scholars***. In this report, we think about impact in three ways: a gross impact of race or ethnicity, a net impact of race or ethnicity, and whether that net impact is or is not statistically significant.

3.4.1 Gross impact

Q: What you mean by **gross impact**?

A: Gross impact refers to different groups having different average scores on an outcome. So a gross impact of race or ethnicity refers to these three groups having different average scores on either the judgements about legal basis of stops or of pat downs. No other factors beyond race/ethnicity and the outcome in question are considered. This is pure ***description***.

Q: **Where** in the technical, statistical reports are these gross impacts described?

A: If you are interested in the gross impact of race or ethnicity on the legal basis of the stop, you can find the relevant percentages showing both absolute differences [1] and relative differences [2] described as ratios in Table 11. These are for weighted data and refer to the full set of investigatory stops from which the examined records were sampled.

More specifically, to look at absolute differences we start with absolute percentages. For example, look at in the bottom third of Table 11 under the section labeled "66 bicycle/sidewalk excluded" under the subsection "percent", under the row "improper (zero)." You see the numbers 3.52, 4.83, 8.21. This means that 3.52 percent of white non-Hispanic stops, 4.83 percent of Hispanic stops, and 8.21 percent of black non-Hispanic stops were improperly premised. [3] The absolute differences in these percentage describe gross impact of ethnoracial category on this outcome, that is, the differences across the three groups.

For example you can say that stopped Black non-Hispanic civilians had the highest rate of “bad” stops (8.21percent) because their percentage is higher than the percentage for either of the other groups. You also could say that the percent of bad stops involving Black non-Hispanic civilians was (8.21-3.52) 4.69 percent higher than the percent of bad stops involving White non-Hispanic civilians.

To look at relative differences compare the ratio of two absolute percentages. You are asking: how many times higher or lower was the bad stop rate for Black as compared to White non-Hispanics?

---

1 An absolute difference is just a difference in the proportion or percentage of a group that has an attribute. Say you have two groups of 10 people each, group A and group B. Five out of 10 or 50 percent of those in group A have tuberculosis. Four out of 10 or 40 percent of those in group B have tuberculosis. The absolute difference is 50 – 40 = 10 percent. The gross impact of being in Group A or B on tuberculosis is 10 percent.

2 Relative differences are created when the percentage for one group is expressed relative to the percentage for another group. Go back to the two groups of 10, A with a 50 percent tuberculosis rate and B with a 40 percent tuberculosis. The relative difference in the disease rate between the two groups can be expressed in two ways. If you want to talk about the disease rate in group A relative to group B you would take the ratio 50:40; alternatively you could say the rate in A was 20 percent higher. If you want to talk about the disease rate in group B relative to group A you would take the ratio 40:50; alternatively you could say the disease rate in B was 20 percent lower.

3 The term “improperly premised” means, here, that the Consultant determined that this percentage of the coded legal narratives failed to satisfy the RAS standard.

The ratio 8.21:3.52 = 2.33. You can say the Black "bad" stop rate was 2.33 *times* the White "bad" stop rate. This is also the same as saying that the Black "bad" stop rate was 133 *percent more* than the white stop rate.[4] These relative differences are also shown in Figure 1.

Q: Where do I find the **gross impact** of race and ethnicity on **pat down basis**?
A: See Table 20.

3.4.2 Net impact and other factors

Q: What do you mean by **net impact** of ethnicity or race?
A: The net impact of the race or ethnicity variable refers to the size of the connection between that factor and the outcome *after taking into account other factors*. It is a part of the connection that is unrelated to these other factors.

For example, if you are interested in the impact of being Black non-Hispanic on being in a bad vs. good stop. You start by describing the gross impact, the percent of Hispanics (4.83) vs. the percent White non-Hispanics (3.52) vs. the percent Black non-Hispanics (8.21) involved in bad stops.

After taking into account other factors means that the influence of each of these other factors (see below) on the Black variable, and on the outcome variable, has been removed. The link now reflects *only* the portions of each of these two variables – Black non-Hispanic and good vs. bad stop, that are *unrelated* to these other factors. Thus, "net impacts" are a part of the observed influence of the predictor on the outcome, which is unrelated to the other "controlled for" factors.

Q: What **other factors** do you take into account?
A: The police district in which the stop took place, and the gender and age of the civilian as well. Reasons for including these specific other factors appear in section 8.2.

Q: What are the implications of removing these other factors to examine net race impacts?
A: There are two broad implications. On the one hand, it helps focus on *only* the influence (effect/impact) of the isolated factor of race or ethnicity on the outcome being studied. This is *recommended social science best practice* in situations like this. On the other hand, race connects to these other factors so by removing these other factors we might be removing a substantial part of the race influence on the outcome. The analyses conducted here took steps to address this latter concern. [5] But more importantly, at least for the race impact on stop basis, the size of the net impact is comparable to the size of the gross impact. This suggests the latter issue was not a concern.

Q: Suppose your model had **expanded** the set of **other factors** that you took into account? Could that have changed the results shown here?
A: Yes it could. Statistical results shown here (see below on "statistical significance") are

---

[4] When switching from one number times another to one number as a percent of another, we subtract 100 percent because if a number is 100 percent of another number it is the same number.

[5] More specifically, models were tested to see if the race variable depended on (interacted with) other possible demographic combinations (being young and black, being black and male, being young and black and male). Those models allowing race impacts on the outcome to depend on these other factors were not noticeably better. Further, geography, in the form of district differences, was examined and described.

specific to the predictors used in these models. Different models with different predictors could have resulted in a statistically significant race effect shown here in some models (Table 16 for example) become non-significant.

Q: How **big was the net impact of race** on stop basis?
A: It was between four and six percent. In other words, after controlling for these other factors, the percentage of properly premised stops for Black non-Hispanic civilians was about four to six percent lower when compared to the percentage for White non-Hispanic civilians stopped. This is shown by the line in each panel in Figure 4. These numbers are for investigatory (coded based on the presence or absence of RAS) stops only. (Generally, stops coded as probable cause stops were dropped. Some analyses did include a particular type of probable cause stop, bicycle sidewalk violations, to see how that affected results.)

3.5 GEOGRAPHY AND STOP BASIS

Q: You took police district into account. Does this mean you **threw away** all the **geographic variation** in the outcome?
A: Not at all. That geographic variation was just put into a separate compartment, and the geographic compartment was split into two sub-compartments.

Q: Can you explain these two geographic sub-compartments?
A: One geographic sub-compartment shows the influence of geography that *relates* to age, gender, race and ethnicity of the stopped civilians, that is, is the factors we used to model the outcome (stop properly or improperly premised. The other sub-compartment is the part of the geographic variation that is *unrelated* to the factors used in our models.

Q: Where do I find the gross impact of geography that was due to the civilian factors you mentioned?
A: If you are interested in the results with just investigatory stops, look at Figure 8. The length of each bar refers to the portion of stops in that district that were predicted to be improperly premised, given the impacts of age, gender, race and ethnicity, and lcoation on stop premise.

Q: In Figure 8, which district was predicted to be the best?
A: Districts 19 and 24 had fewer than three percent of their stops predicted to be improperly premised.

Q: In Figure 8, which district was predicted to be the worst?
A: District 3, where over ten percent of their stops were predicted to be improperly premised.

Q: Where do I find the gross impacts of geography that were not explained by the civilian factors you mentioned?
A: The geographic portion of the outcome not explained by model factors appears in Figure 9. Each district has a filled in circle. If the filled in circle for a district is **below** the **red line** it means that in that district, even after taking civilian age, race, ethnicity and gender into account, the **proportion of proper stops** in that district was **lower than overall**. If the filled in circle for a district is **above** the **red line** it means that in that district, even after taking civilian age, race,

ethnicity and gender into account, the **proportion of proper stops** in that district was **higher than overall**.

Q: Why does each filled in circle have lines coming out of it?
A: Those lines take sampling error into account. After we consider that error, our best guess is that in the full set of investigatory stop records the true mean score for that district on that district is somewhere between where the upper line ends and the lower line ends.

Q: Are any of these district differences in Figure 9 meaningful?
A: They may be. Look at the two left-most district means, which are for Districts 10 and 3. The top end of their lines do not cross the red zero line. This means that if we were to repeat this sampling and analysis 100 times with 100 independent samples, 95 times out of 100 these two districts would have lower-than-average fractions of properly premised stops.

Q: So are you saying there may be something going on in Districts 10 and 3, based on Figure 9, that is unrelated to the civilian factors you used, that is resulting lower fractions of good investigatory stops in those districts?
A: We are.

Q: Do you know what is responsible?
A: We do not. It could be something about the district organization itself, something about the mix of people encountered on the street walking or driving, something about the mix of land uses or public transit in these districts, or some other factor. We just don't know.

3.6 Statistical significance

Q: So you've explained a gross impact of race, and a net impact of race; what does a **statistically significant impact of race** mean?
A: It means that the net impact of race is not due to chance alone. Stated differently, when we infer from the sample finding, back to the full set of investigatory stops, and take sampling error into account, if a net race impact is statistically significant we are confident that the impact of race, after taking other factors into account, in the full set of records, is not zero.

Q: How confident are you?
A: We are confident that if we repeated these analyses with 100 independent random samples, and all sampling and analytic steps were the same, 95 times out of 100 our sample estimate of net race impact after taking sampling error into account would encompass only *non-zero* net impacts of race in the full set of records.

Q: So it sounds like you are thinking in three increasingly restrictive ways about impacts of race on the outcomes: any connection, any connection after taking other factors into account, and a connection after taking other factors into account that may be "true" in the full set of records.
A: Yes.

Q: If race has a statistically significant impact on stop basis, like it does here, does this mean that the race of the civilian encountered by the officer is *causing* the outcome?
A: In a social science framework, not necessarily. In social science, correlation does not always mean causation. Figuring out whether the impact might be causal, wholly or in part, requires additional social science steps not undertaken here.

3.7 PAT DOWN BASIS AND RACE

Q: Does civilian race link to whether the civilian experienced in improperly premised or "bad" pat down?

A: There is no statistically significant impact of civilian race on whether a bad pat down took place. But there seems to be a **noticeable gross** geographic connection between race and this outcome. See Figure 13. The predicted chances that the stop would involve a bad pat down are higher, about 3 percent rather than 2 percent, in districts where higher proportions of stopped civilians are Non-Hispanic Black. There are too few districts to allow for a meaningful test of a net connection at the district level.

3.8 LEGAL QUESTIONS

Q: How do these ways of thinking about race impacts connect to legal ideas about disparate race impact and disparate race treatment?

A: We don't know. Those are legal determinations. We leave that to those with legal training, such as the Consultant, who will consider our findings along with all other relevant features of the data and the broader context of these assessments.

3.9 BOTTOM LINE

Q: What are the most important take away lessons?

A: In the authors' view, there are three. *First*, the majority of investigative stops, somewhere around 90 percent, appear to be sufficiently premised or "good" stops. *Second*, stops of non-Hispanic Black civilians, compared to those of non-Hispanic White civilians, were less likely to be "good" stops. Thus, even though the fraction of "bad" stops is relatively small, there is racial patterning within that fraction. That is, there is a statistically significant difference by race on this outcome after controlling for other factors. *But bear in mind* that the significant net race effect depends on the type of model used and the set of stops included in that model. *Third*, "good" stops seem less likely in a couple of districts, Districts 3 and 10, for reasons that are not clear at this time.

4 EXECUTIVE SUMMARY

This report analyzes a sample of ISR data from the period January through June 2016. These records were coded to determine the legal sufficiency of the stop itself, the legal sufficiency of a pat down if it occurred, and the legal sufficiency of a search if it occurred. The sample of records coded included equal numbers of non-Hispanic Black stopped civilians, Hispanic stopped civilians, and white non-Hispanic stopped civilians. The following factors were used in a statistical model predicting whether or not the stop itself was properly premised on reasonable articulable suspicion: race, ethnicity, age, gender, and district context. Stop premising was considered three different ways: with bicycle sidewalk violations meeting the probable cause standard excluded; with those same violations included but classified as improperly premised because the stop was based on probable cause rather than reasonable articulable suspicion relevant to investigation; and finally, with those same violations included but classified as properly premised because there was a reason for the stop even though that reason was not investigatory. Except for the bicycle sidewalk violations as noted above, all other stops based on probable cause were excluded.

These models sought to learn whether race, unrelated to its links with other factors; and ethnicity, unrelated to its links with other factors, had significant **net** impacts on the outcomes in question. Those outcomes in question were: properly or improperly premised stop basis; properly or improperly premised pat downs; and properly or improperly premised searches. It turned out, however, that the third outcome presented such a small number of improper searches that models were not run.

Proper or improper stop premises. Finding a significant net impact of race depended on how the aforementioned bicycle sidewalk violations meeting the probable cause standard were treated. Results showed a statistically significant net impact of race ($p < .05$) on stop premise with bicycle sidewalk violations excluded, and when those violations were included but classified as properly premised. But if probable cause bicycle sidewalk violations were treated as improperly premised investigatory stops no significant net impact of race appeared. To see how sturdy the significant net race impact was, models with the significant race effect were repeated using a different type of analytic approach. The significant net impact of race, however, failed to replicate with this different type of analytic model.

In models where there was a significant net race impact, gross race impacts on the outcomes were examined as well. These examinations do not seek to isolate the impacts **solely** associated with race or ethnicity or gender. These showed that stops which had the highest average predicted probability of being improperly premised were stops involving Black males.

Turning to geographic variation in the models with significant race impacts, some districts had a significantly higher portion of improperly premised stops after taking model factors into account (Districts 10, 3).

Pat down premises. Another outcome of interest was the sufficiency of the reasons given for a pat down, if such a pat down occurred. At the level of individual records, a net effect of gender surfaced. This suggested that although women were less likely to be patted down at all, if women were patted down their chances were higher than men's chances of being in an improperly premised pat down. This finding should not be leaned on too heavily, however, since it was based only on seven improperly premised pat downs of women. Descriptively, at the district level, a gross relationship between race and pat down premise appeared. Districts that had higher average predicted probabilities of improperly premised pat downs also had higher fractions of stopped civilians who were non-Hispanic Black. This is a very small difference, but noticeable.

Search basis. After removing custodial searches, there were too few searches lacking probable cause to allow any analysis of multiple factors determining whether searches were properly premised.

Overall. Results suggest the following

- A significant net impact of race on stop premising surfaces in some models.
- But whether this impact is significant or not depends on how the subset of probable cause stops examined here, bicycle sidewalk violations, are classified in terms of stop premising, and the type of analysis used.

- Models showing a significant net race impact align with the descriptive pattern of gross race impacts. Descriptive patterns based on these models showed that stopped non-Hispanic Black civilians were predicted by the models to have the highest chances of being in an improperly premised stop. When predicted probabilities, based on age, gender, race, ethnicity and district context are considered rather than raw data, the group predicted to be most likely involved in an improperly premised stop were Non-Hispanic Black civilians, especially if they were male (Figure 3).
- There may be an ecological link between good or bad pat downs and race. Districts with higher fractions of stopped civilians who were Black Non-Hispanic were districts where a higher fraction of stops involved bad pat downs.
- There are three important points of context. First, this significant net race effect seen on stop premising in two out of the three models occurred in a context where roughly 90 percent of stops appeared properly premised. Some might think this makes the net race impact small. Second, others might think it a testament to the race link that it occurred *even though* such a small fraction of stops were improperly premised. Third, the net race impact failed to prove significant when alternate single-level rather than multilevel analytics were used.
- A significant net impact of gender on proper pat down premising surfaced, with women more likely to be involved in an improper pat down. But this finding is built on only seven improperly premised pat downs, and thus should be interpreted with extreme caution.

5 PURPOSE

This report examines a sample of investigatory stop reports (ISRs) generated by the Chicago Police Department during the period January-June 2016. In this sample, three different races/ethnicities are equally represented: Black Non-Hispanic, White Non-Hispanic, and Hispanic stopped civilians. The narrative fields in these stops have been coded to reflect the propriety of the legal premises of three actions: the stop taking place; a pat down, if it took place; and a search, if it took place. The first two outcomes are legitimate as investigatory procedures if based on reasonable articulable suspicions, as specified in the narrative fields of the ISRs completed by the officers. The third is legitimate if based on probable cause, as specified in the same way. So for each outcome, this report examines the rate at which each of these three outcomes was properly premised, or improperly premised, given legal considerations.

At the outset probable cause stops were excluded because they were not investigatory stops, and the focus of The Agreement is on police investigative stop protocols. Nevertheless, to explore the implications of bounding and classifying such probable cause stops, and how that bounding and classification might affect the outcome, as an illustration one specific type of stop was treated in different ways. Those stops involved officers notifying civilians over the age of 12 riding bicycles on sidewalks that doing so was a municipal violation. There were 10 stops involving bicycles on sidewalks that were investigative in nature, but there were 66 that were probable cause stops to notify civilians of their lawbreaking. So analyses were done three ways. Bicycle sidewalk violations could be included but classified as improperly premised because they were not investigatory in nature. Or, they could be included but classified as properly premised because the officers had a reason for making some type of stop. Or they could be excluded.

If models indicated significant net race impacts of race or ethnicity on the propriety of stop premises surfaced under one of these three bicycle coding arrangements, further explorations were conducted with that model. More specifically, geographical residual variation was explored to learn whether some districts, even after taking account of the determinants of stop premising had been taken into account, deviated significantly; and, predicted gross differences on the outcome, depending on ethnoracial and gender combinations, were described.

This report only addresses how race/ethnicity connect with the legal premises of the actions examined. This report will not address how race/ethnicity connect to the *occurrences* of stops, or pat downs, or searches. Those linkages receive attention in a separate report on post stop outcomes.

6 METHODOLOGY

6.1 SAMPLING STRATEGY

From the full set of ISRs for the period, three random samples were pulled: one for each of the three key racial/ethnic groups. Simple random sampling was used. Each group was sampled at a rate to provide 1,800 sampled records for each group for the period January-April. When data became available for May-June, additional records for each race/ethnicity were sampled, using the same sampling ratios as were used in January-March.

For each race/ethnicity, the random sample was further sampled, taking a random 50 percent sample of each. These 50 percent subsamples were then joined together so that records for all three races/ethnicities could be analyzed.

Given the samples drawn, none of the results here apply to any other racial/ethnic groups not examined (e.g., differences between Asian and White Non-Hispanic stopped civilians).

6.2 Implications for weighted and unweighted analyses

The equal size subsamples for each race/ethnicity maximize the statistical power of analyses examining race/ethnicity differences. Descriptive information is usually presented for unweighted data, with roughly equal numbers of stops in each of the three racial/ethnic groups. Statistical models are conducted with weighted data. Therefore, patterns of statistical significance from the models indicate whether an impact observed with the sample likely applies as well to the full population of records.

6.3 Coding

The 50 percent subsamples were coded. Coding categories appear in Appendix A.

6.4 A Priori power analyses

A priori statistical power analyses showed that with at least 1,800 records, a difference in proportions of five percent would have slightly better than 80 percent statistical power. This is considered an acceptable level of statistical power in many fields (Cohen, 1992). Statistical power analyses specific to the multivariate and mixed effects models conducted here were not estimated.

6.5 Outcomes

The three outcomes examined are:

- Was the stop properly premised on reasonable articulable suspicion (RAS) factors?
- Was the pat down, if it occurred, properly premised on reasonable articulable suspicion (RAS) factors?
- Was the search, if it occurred, properly premised on probable cause?

7 Descriptive statistics

Descriptive statistics for predictor variables and binary outcomes appear in Table 1. Descriptive statistics for the pat down outcome appear in Table 2. Search outcome descriptive statistics appear in Table 3. Specific outcome variables are explained later as they are introduced.

Table 1. Descriptive statistics: Predictors and binary outcome variables

7.1.1 Predictor variables		N	Min	Max	Mean	SD	Median
Black	dblack	3,376	0	1	0.709	0.454	1
Hispanic	dhisp	3,376	0	1	0.214	0.410	0
Male	dmale	3,376	0	1	0.869	0.337	1
Age	age2	3,376	7	100	28.823	13.276	24
Age (Centered)	c_age2	3,376	-22.550	70.450	-0.727	13.276	-5.550
District 1	dist01	3,376	0	1	0.016	0.125	0
District 2	dist02	3,376	0	1	0.042	0.201	0
District 3	dist03	3,376	0	1	0.071	0.257	0
District 4	dist04	3,376	0	1	0.076	0.265	0
District 5	dist05	3,376	0	1	0.034	0.182	0
District 6	dist06	3,376	0	1	0.039	0.193	0
District 7	dist07	3,376	0	1	0.076	0.265	0
District 8	dist08	3,376	0	1	0.064	0.245	0
District 9	dist09	3,376	0	1	0.078	0.269	0
District 10	dist10	3,376	0	1	0.076	0.266	0
District 11	dist11	3,376	0	1	0.110	0.312	0
District 12	dist12	3,376	0	1	0.037	0.189	0
District 14	dist14	3,376	0	1	0.014	0.119	0
District 15	dist15	3,376	0	1	0.059	0.235	0
District 16	dist16	3,376	0	1	0.023	0.151	0
District 17	dist17	3,376	0	1	0.016	0.127	0
District 18	dist18	3,376	0	1	0.019	0.136	0
District 19	dist19	3,376	0	1	0.023	0.151	0
District 20	dist20	3,376	0	1	0.019	0.136	0
District 22	dist22	3,376	0	1	0.017	0.129	0
District 24	dist24	3,376	0	1	0.042	0.201	0
District 25	dist25	3,376	0	1	0.048	0.214	0
Binary outcome variables							
Stop properly premised v. 1	stopsuff3	3,376	0	1	0.927	.261	1
Stop properly premised v. 2	stopsuff4	3,376	0	1	0.930	0.255	1
Additional information							
Pat down occurred	dpat	3,376	0	1	0.347	0.476	0
Search occurred	dsearch	3,376	0	1	0.150	0.357	0

Note. Stop premise v. 1 treats bicycle sidewalk probable cause violations (n=66) as improperly premised stops; v. 2 treats those as properly premised stops. Both versions treat stops lacking RAS factors as improperly premised. Unweighted data. Source: Jan.-Jun. 2016 legal narratives equal race sample.

Table 2 Descriptive statistics: Pat down premising (pat_ras3) with bicycle sidewalk violations included

Code	Category	N	Percent
1	Properly premised	955	28.29
2	Improperly premised	80	2.37
3	No pat down	2341	69.34
Total		3376	100

Note. Unweighted data. Source: Jan.-Jun. 2016 legal narratives equal race sample.

Table 3 Descriptive statistics: Search sufficiency basis

Code	Category	N	Percent
0	Properly premised	218	6.46
1	Improperly premised	16	0.47
2	Custodial	385	11.4
.	No search	2,743	81.25
.i	Insufficient information	14	0.41
Total		3,376	100

Note. Unweighted data. Source: Jan.-Jun. 2016 legal narratives equal race sample.

8 RESULTS: STOP PREMISING

This section considers the determinants of whether the stop was properly premised on reasonable articulable suspicion factors (RAS), or not. The approach starts just by examining the connection between the outcome and racial/ethnic combinations. These provide clues to gross race and ethnicity connections with the outcome without taking additional factors into account. Later models then introduce those additional factors so the *net* impact of race, ethnicity, and gender can be gauged.

Table 4 below shows the distribution of race/ethnicity for the coded ISRs. 17 inappropriately duplicated sampled ISRs have been removed.

Table 4 Numbers and percent in sample by race/ethnicity

White NH	1,416	33.45
Hispanic	1,394	32.93
Black NH	1,423	33.62
Total	4,233	100.00

Note. White NH = White Non-Hispanic; Black NH = Black Non-Hispanic. Source: Jan.-Jun. 2016 legal narratives equal race sample. All sampled and coded records included. Unweighted data.

The table shows the number of ISRs for each of the three racial/ethnic groups: White Non-Hispanic, Hispanic, and Black Non-Hispanic. Each group contributed, as planned with the sampling design, about a third of the coded records.

In these 4,233 records there were 857 stops (20.2 percent) classified as probable cause stops. These were dropped so the focus could be exclusively on investigatory stops, except as noted below.

Close examination was made of stopping civilians for sidewalk bicycle riding. 76 stops included both “BICYCLE” and “SIDEWALK” in the narrative fields. Almost all of these involved a person over 12 years of age riding a bicycle on a sidewalk. One involved private contractors removing a bicycle rack on the sidewalk. Of these 76, 66 were probable cause stops and 10 were investigatory stops. As explained above, analyses were repeated treating these bicycle sidewalk violations three different ways.

The distribution on race/ethnicity is shown in Table 5 with probable cause stops, save the 66 bicycle sidewalk violations, excluded.

Table 5 Numbers and percent in sample by race/ethnicity: Investigatory stops only

Racial/ethnic combination	N	Percent
White NH	1,134	33.59
Hispanic	1,142	33.83
Black NH	1,100	32.58
Total	3,376	100

Note. Unweighted data. White NH = White Non-Hispanic; Black NH = Black Non-Hispanic. Probable cause stops dropped. Bicycle on sidewalk stops included. Source: Jan.-Jun. 2016 legal narratives equal race sample.

Table 6 shows the distribution of cases to RAS codes with probable cause stops, save the 66 bicycle sidewalk violations, excluded.

In the sample, 92.6 percent (3,128/3,376) of the ISR stops were properly premised on reasonable articulable suspicions (code 0).

The next largest group of records were those 99 (2.9 percent) where the narratives captured insufficient information to make a determination about stop basis (code=7). Another four records had a different insufficiency code, .i, making the total number of records in this group 103 (3.1 percent).

The next largest group of 66 records (1.95 percent) were instances of police stopping individuals who were over the age of 12 but riding bicycles on public sidewalks in violation of municipal code.

The next largest group were records where no criminal activity appeared to be underway or planned ("afoot"); 57 records (1.7 percent) were classified improperly premised on these grounds.

All of the other codes, individually, were applied to less than one percent of the reviewed records. Among these, the only code applied to more than ten records was the judgement (code=11, 13 cases) that there was no basis whatsoever for a Terry investigatory stop.

Table 6. Assessment of stop premises: Investigatory stops only

	N	Percent
0. RAS sufficient	3,128	92.65
1. Bicycle on sidewalk	66	1.95
2. Time/distance too attenuated	2	0.06
4. Hunch not personal observation	5	0.15
7. Not enough facts	99	2.93
8. Fleeing or avoidant subject only	2	0.06
9. No criminal activity afoot	57	1.69
11. No basis for Terry or PC stop	13	0.39
.i (insufficient information)	4	0.12
Total	3,376	100

Note. Unweighted data. Bicycle on sidewalk stops included. Probable cause stops dropped.
Source: Jan.-Jun. 2016 legal narratives equal race sample.

According to the municipal code in Chicago, individuals over the age of 12 are not permitted to ride bicycles on sidewalks. These cases here include one instance where a bicycle rider hit a pedestrian (ISR 6174) and another instance (ISR 82496) where a person riding a bicycle on a public sidewalk was "approaching several unknown subjects and engaging in short conversations at a location that has been subjected to numerous civilian complaints regarding narcotic activity and multiple arrests pertaining to such."

If these bicycle sidewalk violation stops are classified as properly premised because they meet a higher probable cause standard, being a clear violation of municipal code, the percent of stops in the sample that were properly premised rises to 94.6 percent. See Table 7. If the bicycle sidewalk violations are classified as improperly premised because they are not properly premised as *investigative* stops, the percent of stops properly premised is, as shown in Table 6, 92.6 percent.

Table 7 Assessment of stop premises: Investigatory stops only, bicycle sidewalk violations included and coded as RAS sufficient

	N	Percent
0. RAS sufficient	3,194	94.61
2. Time/distance too attenuated	2	0.06
4. Hunch not personal observation	5	0.15
7. Not enough facts	99	2.93
8. Fleeing or avoidant subject only	2	0.06
9. No criminal activity afoot	57	1.69
11. No basis for Terry or PC stop	13	0.39
.i	4	0.12
Total	3,376	100

Note. Unweighted data. Bicycle on sidewalk stops included. Probable cause stops dropped. Source: Jan.-Jun. 2016 legal narratives equal race sample.

If the 66 bicycle on sidewalk violations that met the probable cause standard are removed altogether, 3,310 records remain. For these records, the distribution on stop premises appears in Table 8.

Table 8 Assessment of stop premises: Investigatory stops only, 66 bicycle sidewalk violations removed

	N	Percent
0. RAS sufficient	3,128	94.5
2. Time/distance too attenuated	2	0.06
4. Hunch not personal observation	5	0.15
7. Not enough facts	99	2.99
8. Fleeing or avoidant subject only	2	0.06
9. No criminal activity afoot	57	1.72
11. No basis for Terry or PC stop	13	0.39
.i	4	0.12
Total	3,310	100

Note. Unweighted data. Bicycle on sidewalk stops excluded. Probable cause stops dropped.
Source: Jan.-Jun. 2016 legal narratives equal race sample.

In order to present this outcome in a more condensed format, a summary stop premise variable (**stopsuff4**) was constructed with just two values, 0 and 1. A value of 1 means that the stop was properly premised (code = 0 in Table 6); the narrative revealed reasonable articulable suspicion. A value of 0 collapses all the other codes (Table 6, or Table 7 or Table 8 depending on the analysis) suggesting the stop was improperly premised. The 66 bicycle sidewalk violations rising to the probable cause standard are classified here as RAS sufficient.

So here

1 = stop had sufficient RAS (code 0 in Table 6)

0 = stop premised neither on reasonable suspicion nor on probable cause (codes 2 and higher in Table 6 or Table 7 or Table 8 depending on the analysis)

Another version of this variable (**stopsuff3**) was the same as the preceding variable, except that here the 66 bicycle sidewalk violations were classified as *improperly premised* because they were not investigatory stops.

The relationship between these two versions of the stop premise outcome variable appears in Table 9.

Table 9 Propriety of stop premises: relationship between two coding approaches to outcome

		Propriety of stop premises: bicycle sidewalk violations included and treated as proper (stopsuff4)		
		Improperly premised (0)	Properly premised (1)	Total
Propriety of stop premises: bicycle sidewalk violations included and treated as improper (stopsuff3)	Improperly premised (0)	182	66	248
	Properly premised (1)	0	3,128	3,128
	Total	182	3,194	3,376

Note. Unweighted data. Probable cause stops dropped. Source: Jan.-Jun. 2016 legal narratives equal race sample.

8.1 Race and Ethnicity: Descriptive Patterns of Gross Impacts

Table 10 shows the counts and proportions of properly premised stops, and improperly premised stops, by race/ethnicity combinations. These appear under three arrangements. In the top portion of the table bicycle sidewalk violations meeting the probable cause standard are included but classified as improperly premised because they are not investigatory stops. In the middle portion of the table those same bicycle sidewalk violations are included but are now treated as properly premised because the officers had a reason, albeit not an investigatory one, for making the stop. In the bottom portion of the table bicycle sidewalk violations meeting the probable cause standard (n=66) are removed from the calculations.

Please note that the figures in the total column, on the right hand side of the table, <u>apply only to the equally weighted sample</u>. We describe below the totals that apply to the entire set of stops, with each of the three groups weighted proportional to their contribution to the total set of stop records for these groups.

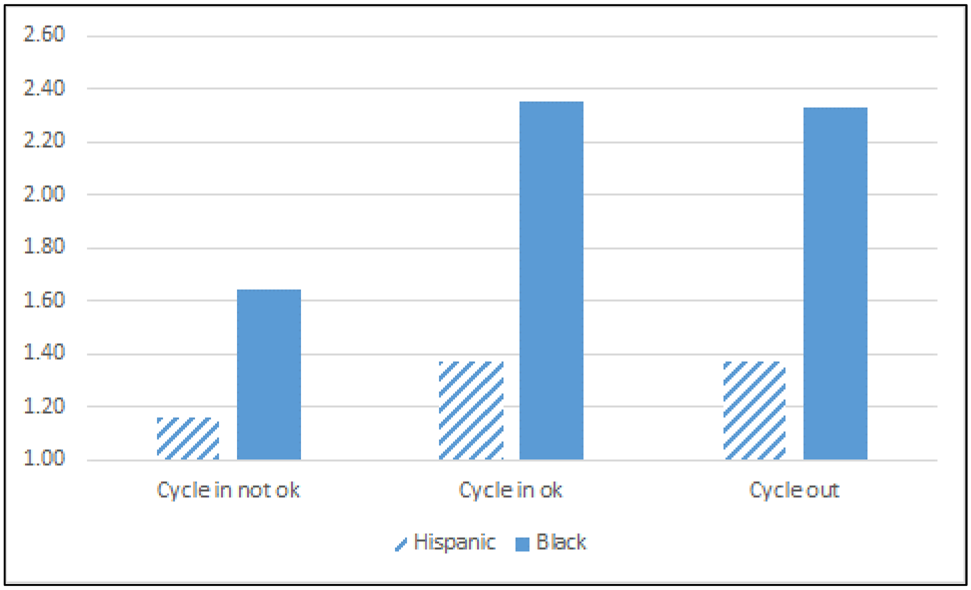
The differentials in improperly premised stop rates also are depicted graphically in Figure 1. Two points are clear

- **How the probable cause sidewalk violations are handled has a noticeable impact on the size of the disparities across ethnoracial groups.** The disparities relative to the white improperly premised rate are clearly lower if bicycle sidewalk violations get included as improperly premised, and higher if those same stops are included as properly premised, or excluded.

- **In addition, the disparities relative to the white stop properly premised rate is greater when Black Non-Hispanic civilians are contrasted with White Non-Hispanic civilians than it is when Hispanics are contrasted with Whites.**

The relevant detailed numbers appear in Table 10.

Figure 1 Improperly premised stop rate, relative to white improperly premised stop rate, for Hispanic and Black Non-Hispanic civilians, under three treatments of bicycle sidewalk violations.



Source: Jan.-Jun. 2016 legal narratives equal race sample.

Table 10 Comparing counts and proportions of improperly premised stops across three race/ethnic groups

66 Bicycle/sidewalk in but improper premises for investigatory purposes	White NH	Hispanic	Black NH	Total
		N		
Improper (0)	66	77	105	248
Proper (1)	1,068	1,065	995	3,128
Total	1,134	1,142	1,100	3,376
		Percent		
Improper (0)	5.82%	6.74%	9.55%	7.35%
Proper (1)	94.18%	93.26%	90.45%	92.65%
Total	100.0%	100.0%	100.0%	100.0%
Ratio Improperly premised percent relative to White percent Improperly premised		1.16	1.64	

66 Bicycle/sidewalk in & properly premised

Stop premises	White NH	Hispanic	Black NH	Total
		N		
Improper (0)	39	54	89	182
Proper (1)	1,095	1,088	1,011	3,194
Total	1,134	1,142	1,100	3,376
		Percent		
Improper (0)	3.4%	4.7%	8.1%	5.4%
Proper (1)	96.6%	95.3%	91.9%	94.6%
Total	100.0%	100.0%	100.0%	100.0%
Ratio Improperly premised percent relative to White percent Improperly premised		1.37	2.35	

66 Bicycle/sidewalk excluded

Stop premises	White NH	Hispanic	Black NH	Total
		N		
Improper (0)	39	54	89	182
Proper (1)	1,068	1,065	995	3,128
Total	1,107	1,119	1,084	3,310
		Percent		
Improper (0)	3.5%	4.8%	8.2%	5.5%
Proper (1)	96.5%	95.2%	91.8%	94.5%
Total	100.0%	100.0%	100.0%	100.0%
Ratio Improperly premised percent relative to White percent Improperly premised		1.37	2.33	

Note: Do NOT interpret total column as representative of entire population of non-probable cause stops. See Table 11. Source: Jan.-Jun. 2016 legal narratives equal race sample.

Of course the results in Table 10 do not reflect the full population of non-probable cause (investigatory) stops. This is because the relative representation of each of the three ethnoracial groups in the sample does not match the proportions of each of those groups in the full set of investigatory stops. In order to apply results with the sample to the full population of investigatory stops, the cases in the sample need to be appropriately weighted so that each group is proportionally represented.

In effect this means "counting" each Black non-Hispanic record in the sample as more than one record because this group's relative representation in the sample, just a third, is far smaller than its representation in the full set of investigatory stops, which is about 71 percent. So each Black non-Hispanic record counts for 2.15 records in the full set of investigatory records.

The reverse is the situation for White non-Hispanics. In the sample they are a third, but they make up about eight percent in the full set of records. So each White non-Hispanic record is counted as equivalent to about a quarter of a record (weight=.23) in the full set of investigatory records.

Hispanics, like White non-Hispanics, are over counted in the equal race sample. Making up a third of the equal race sample, they are only about 21 percent of the full set of investigatory stops. So each Hispanic record is counted as equivalent to about two thirds (weight=.62) of a record in the full set of investigatory stops.

Once these weights are "turned on," we can estimate the fraction of properly and improperly premised stops in the full set of investigatory stops.

One additional point merits mention before getting to the bottom line. When extrapolating from the sample back to the full set of records, uncertainty must be added in because of sampling error. Sampling error is captured with the upper and lower bounds of the confidence interval. These intervals inform us that, according to sampling theory, if the entire sampling procedure were to be independently replicated 100 times with comparable data, 95 times out of 100 we estimate that the "real" mean or proportion for the full set of records would lie within that interval.

Looking at the last two columns of Table 11 reveals the following. Although the numbers vary depending on which of the three scenarios are investigated,

- **The estimated proportion of properly premised non-probable cause stops ranges from 90 percent to 94 percent.**
- **The estimated proportion of properly premised investigatory stops ranges from 6 to 10 percent.**
- **Although the estimated proportion of improperly premised stops varies somewhat depending on the inclusion and coding scenario used, the confidence intervals overlap meaning the estimates are essentially equivalent.**
- **The same holds for the estimated proportion of properly premised stops.**

Here are the details.

If bicycle/sidewalk violations are included but viewed as improper for investigative purposes, the percent of good, properly premised stops using weighted data is **91.3 percent.** Taking sampling error into account our best guess is that the true percent, when considering all investigatory stops, is between **90.4 and 92.3 percent.**

The percent for bad or improperly premised stops is **8.7 percent**, and our best estimate is between **7.7 and 9.6 percent**.

In the middle portion of the table, if bicycle/sidewalk violations are included and treated as properly premised stops, the percentage of properly premised or good stops using weighted sample data is **93 percent.** Taking sampling error into account our best guess for the "true" percentage of good stops for all investigatory stops is between **92.1 and 93.9 percent.**

The proportion of bad or improperly premised stops is **7 percent** and the best estimate taking sampling error into account is between **6.1 and 7.9 percent.**

In the bottom portion of the table with bicycle/sidewalk violations excluded, the weighted sample percentage is **92.9 percent** and the best guess for the "true" percent of good or properly premised stops in the full set of investigatory records is between **92 and 93.8 percent.**

For bad or improperly premised stops the weighted mean **is 7.1 percent.** The best estimate for percent bad or improperly premised stops after taking sampling error into account is between **6.3 and 8 percent.**

Table 11 Comparing counts and proportions of improperly premised stops across three race/ethnic groups: Weighted sample

Stop premises	WHITE NH	HISPANIC	BLACK NH	TOTAL	95% LCL	95% UCL
66 Bicycle/sidewalk in but improper premises for investigatory purposes						
N						
Improper (0)	15.24	48.61	228.44	292.29		
Proper (1)	246.65	672.37	2164.70	3083.71		
Total	261.89	720.98	2393.13	3376.00		
Percent						
Improper (0)	5.82	6.74	9.55	8.66	7.71	9.61
Proper (1)	94.18	93.26	90.45	91.34	90.39	92.29
Total	100	100	100	100		
Ratio Improperly premised percent relative to White percent Improperly premised						
		1.16	1.64			
66 Bicycle/sidewalk in & properly premised						
N						
Improper (0)	9.01	34.09	193.63	236.72		
Proper (1)	252.88	686.89	2199.51	3139.28		
Total	261.89	720.98	2393.13	3376.00		
Percent						
Improper (0)	3.44	4.73	8.09	7.01	6.15	7.87
Proper (1)	96.56	95.27	91.91	92.99	92.13	93.85
Total	100	100	100	100		
Ratio Improperly premised percent relative to White percent Improperly premised						
		1.38	2.35			
66 Bicycle/sidewalk excluded						
N						
Improper (0)	8.98	33.98	193.02	235.98		
Proper (1)	245.87	670.25	2157.89	3074.02		
Total	254.85	704.24	2350.91	3310.00		
Percent						
Improper (0)	3.52	4.83	8.21	7.13	6.25	8.01
Proper (1)	96.48	95.17	91.79	92.87	91.99	93.75
Total	100	100	100	100		
Ratio Improperly premised percent relative to White percent Improperly premised						
		1.37	2.33			

Note. Results based on weighted data, so the proportion of records for each of the three ethnoracial categories matches their proportions in the sample with probable cause stops removed. Last two columns show the upper and lower bounds of the 95 percent confidence interval. Numbers of cases are not integers because these are weighted counts. Source: Jan-Jun 2016 legal narratives equal race sample.

8.2 WHICH ADDITIONAL FACTORS BEYOND RACE AND ETHNICITY AND GENDER SHOULD BE TAKEN INTO ACCOUNT?

1. District context. Each district presents its own complex of crimes, disorder problems, and populations using the streets in the district. Within one police department, partly in response to the above, district cultures can vary (Klinger, 1997).
2. Gender requires attention. (a) The agreement directs attention to gender, and females do get stopped. In the sampled investigatory stops, the bulk of stopped civilians (2,840 or 84 percent) are male, but the sample includes 536 women (16 percent). (b) Further, gender is linked to race/ethnicity. [6] Whereas 21 percent of White Non-Hispanic investigatory stops were of women, the corresponding percentage was 14.8 for Hispanic investigatory stops and 11.7 for Black Non-Hispanic investigatory stops. To get at the net race/ethnicity connection with stop premise disadvantage, gender must be taken into account. (c) Further, gender has an overall relationship with stop premise sufficiency.[7] Whereas 7.8 percent of investigatory stops of males were improperly premised, the corresponding percentage for females was 5.0 percent. (d) Finally, gender may be relevant to the outcome in a particular combination. Given intersectionality theory (Fader & Traylor, 2015), one might expect Black Non-Hispanic women to be at particular risk of an improperly premised stop.
3. Given results in other jurisdictions finding younger Black males more at risk of discretionary searches (Rosenfeld, Rojek, & Decker, 2012), one might anticipate that younger Black males are more at risk of improperly premised investigatory stops.
4. Civilian age is relevant. Given the age-crime curve (Laub & Sampson, 2003), officers might pay closer attention to younger civilians on the street. Alternatively, older people might stand out as more suspicious at certain places and times.

8.3 MODELING APPROACH

Mixed effects logit models (melogit) conducted in Stata 15 control for district context (Rabe-Hesketh & Skrondal, 2012). Given recent concerns about mixed models with small numbers of level 2 units (Bryan & Jenkins, 2016; Schmidt-Catran & Fairbrother, 2016), the final model will be repeated with a single level logit model, controlling for district using fixed effects with District 1 (The Loop) as the reference string.

8.4 DECIDING WHICH SPECIFIC MODEL IS THE "BEST" MODEL

Considerable discussion among scholars and activists raises the possibility that particular combinations of demographic factors prove influential for the outcomes under consideration in this report. For example on the basis of intersectionality theory (Fader & Traylor, 2015) one could argue that Black women are particularly at risk. On the basis of driving while black literature one could argue that young black males are particularly at risk.

---

[6] LR $\chi 2$ (df=2) = 36.87; $p < .001$.
[7] LR $\chi 2$ (df=1) = 5.5; $p <.05$.

Following up on this suggestion requires examining models which add, after taking into account the main effects of age, gender, race, ethnicity, and the random effects of district context, additional interaction effects.

Therefore, for this outcome of stop sufficiency, the following series of models were run. [8]

A. A null or ANOVA model determines whether average scores on the outcome differ significantly across districts. Does district context matter?
B. Age and gender main effects are added jointly.
C. Race (black=1) and ethnicity (Hispanic=1) are added to learn whether race and ethnicity result in a markedly better fitting model will controlling for model complexity.
D. A two way interaction of female x Black indicates whether this race/gender combination links to the outcome.
E. To set up for testing the three way interaction (Black and young and male), a model with the constituent two way interactions is run.
F. The three way interaction is added to see if fit while controlling for complexity improves markedly.

When comparing models against one another, the Bayesian Information Criterion (BIC) is the preferred metric for choosing a "better" model (Raftery, 1995). A substantially lower BIC suggests that the model with the lower BIC provides a better fit to the data, while simultaneously controlling for model complexity. Drops of at least 2, 6, and 10 provided, respectively, positive, strong, and very strong evidence of a better model.

Results from the ANOVA model appear in Table 12. The significant likelihood ratio chi squared values confirm that district context should be taken into account.

Results comparing other models in the series are shown in Table 13. The following points emerge. The models adding a specific two way interaction of gender and race, either the male X Black interaction or the female X Black interaction, worsened fit while controlling for model complexity. BIC values went up substantially. Similarly, a model with all two way interactions relevant to the young x Black x male interaction also produced less fit while controlling for complexity, compared to the main effects model. BIC values went up quite substantially. Finally, the three way young x Black x male interaction does not improve model fit compared to model with the constituent main effects and two way interactions already included.

The upshot is simple. The model with only main effects for age, gender, race, and ethnicity, and controlling for district context will be used. [9] This is true for all three treatments of bicycle sidewalk violations

Each model in the series was run with weighted data. In each case models were run three times.

- Once with bicycle sidewalk violations included but considered properly premised.
- Once with bicycle sidewalk violations included but considered improperly premised.
- Once with bicycle sidewalk violations excluded.

[8] Time did not permit examining the interaction question with other outcomes.
[9] Although these same tests were not conducted for the other outcomes, main effects models are used there as well for consistency in modeling across different outcomes.

Table 12. Null mixed effects logit model: Stop sufficiency

Weighted data, bicycle sidewalk violations included and treated as IMproperly premised

	B	SE	OR	95% CI of OR LCL	UCL
				11.226	19.584
Fixed effects					
Constant	2.696	0.142	14.827	11.226	19.584
Proportions			0.937	0.918	0.951
Random effects					
	Variance	SE of variance			
District	0.258	0.138			

LR χ2(df=1) vs. logistic model: = 19.65; p < .001

n = 3,376

Weighted data, bicycle sidewalk violations included and treated as properly premised

	B	SE	OR	95% CI of OR LCL	UCL
Fixed effects					
Constant	2.766	0.140	15.892	12.088	20.892
Proportions			0.941	0.924	0.954
Random effects					
	Variance	SE of variance			
District	0.212	0.123			

LR χ2(df=1) vs. logistic model: = 16.05; p < .001

n = 3,376

Table 12, continued

Weighted data, bicycle sidewalk probable cause violations included

	B	SE	OR	95% CI of OR LCL	95% CI of OR UCL
Fixed effects	2.751	0.141	15.665	11.890	20.637
Constant			0.940	0.922	0.954
Proportions					
Random effects					
	Variance	SE of variance			
District	0.218	0.125			

LR χ2(df=1) vs. logistic model: = 16.42; $p < .001$

n = 3,310

Note. Outcome: 0 = stop improperly premised; 1 = stop properly premised on reasonable articulable suspicion. Source: Jan.-Jun. 2016 legal narratives equal race sample.

Table 13 BIC values different models

	Bicycle sidewalk violations					
	Included – NOT ok		Included-ok		Excluded	
Model	BIC	BIC Δ	BIC	BIC Δ	BIC	BIC Δ
Null (random effects for districts only)	1,964.4		1,696.53		1,688.05	
+ race and ethnicity	1,974.0		1,699.40		1,691.31	
+ age and gender (Full main effects model)	1,972.9		1,710.48		1,701.48	
+ interaction (male x Black)	1,980.9	8.0	1,718.49	8.01	1,709.47	7.99
Full main effects model + 2 way interactions (young, male, Black)	1,993.6	20.7	1733.125	22.65	1723.87	22.39
Above + 3 way interaction	1,996.0	2.4	1,735.47	2.35	1,726.27	2.40

Note run=117. Weighted data. Source: Jan.-Jun. 2016 legal narratives equal race sample.

8.5 PREDICTED PROBABILITIES BASED ON MODEL RESULTS

Once the statistical model is run, each case in the sample has a predicted probability that *that* stop, based on the factors used in the model, is properly premised. This indicates the *predicted* likelihood, between 0 and 1, that the stop in question was properly premised. Each case's score on the predictors in the model, combined with the parameters from the model, generate these predicted probabilities. A *higher* predicted probability means, according to model results, a *greater* likelihood that the investigatory stop was properly premised.

To repeat, these differences in predicted probabilities by race, ethnic groups and gender inform us of overall or ***gross*** *race, gender, and ethnicity effects based on the contributions of all the factors considered by the model. A later investigation of marginal probabilities illuminates* ***net*** *racial, gender and ethnic effects, controlling for other factors.*

Since there are only two outcomes, one minus these predicted probabilities reflects the predicted chances that stops were *improperly premised*. Of interest will be the differences, between gender-and-race/ethnicity-based groups, in these predicted probabilities of an improperly premised investigatory stop.

The predicted probabilities based on model results will be presented only under the bicycle scenarios that result in a significant net impact of race. If the results show no net significant impact of race, predicted probabilities are not pursued.

8.5.1 Net race impacts

Net race impacts get presented under the three different bicycle sidewalk violation scenarios: included and treated as improperly premised investigatory stops; included and treated as properly premised stops; and excluded.

Whether a significant net race impact shows depends on which of the bicycle scenarios are being examined.

8.5.1.1 Bicycle sidewalk violations included, treated as improperly premised

Table 14 shows the results of a model with only main effects. Bicycle sidewalk violations are included as improperly premised. The odds ratio for race suggests that controlling for other factors, Black non-Hispanic civilians' expected odds of having a [*properly* premised stop vs. an improperly premised stop] are about (1-.639=) .361 lower, or 36.1 percent *lower*.

This impact of race, however, is not statistically significant either with a two tailed or even a more generous one tailed test; in both cases $p > .05$. Nor is it significant in the model with only race and ethnicity entered as predictors (results not shown). With this set of investigatory stops, there is no suggestion of a net race impact on stop premise sufficiency after taking district context into account.

Table 14. Main effects model predicting sufficient stop premise: Bicycles on sidewalk included and treated as IMproperly premised

Variable	Variable name	B	SE	OR	z	p <	OR: 95 % confidence interval LCL	UCL
Fixed effects								
Black non-Hispanic (= 1; white non-Hispanic = 0)	dblack	-0.448	0.292	0.639	-1.53	ns	0.361	1.133
Hispanic (= 1; white non-Hispanic = 0)	dhisp	-0.099	0.318	0.906	-0.31	ns	0.486	1.688
Female (=1; 0 = male)	dfemale	0.560	0.231	1.751	2.43	.05	1.115	2.752
Age (centered by sample mean)	c_age2	-0.014	0.004	0.986	-3.2	.01	0.977	0.994
Constant		2.825	0.296	16.859			9.444	30.097
Random effects		Variance	SE of variance					
District		0.204	0.115				0.067	0.616

LR $\chi 2(df=1)$ vs. logistic model: = 18.40; $p < .001$
Note. Outcome = stop sufficiently premised (=1) or not (=0). Weighted data. Bicycles on sidewalk included. n = 3,376. Source: Jan.-Jun. 2016 legal narratives equal race sample. Probabilities are two tailed. Results from mixed effects logit model, investigatory stops grouped by police districts.

Age and gender each significantly influence stop premise sufficiency. Women are *more* likely to be in a *sufficiently* premised stop ($p < .05$), but *older* stopped civilians are *less* likely ($p < .01$). [10] The suggestion is that gender and age each influence stop premise sufficiency in the full sample when bicycle sidewalk violations are included.

Table 15 shows the predicted probabilities based on the factors shown in the above table, and including district context. Figure 2 shows the differences graphically.

[10] Since age goes up to 100, this model was repeated with only 79 or younger, and again with only those 69 or younger. The significance pattern for age did not change, and the OR for age was unchanged for the first two decimal places.

Table 15 Predicted probabilities for stop premises by gender and race/ethnicity: Bicycle sidewalk violations included and treated as IMPROPERLY premised

Predicted probability stop properly premised

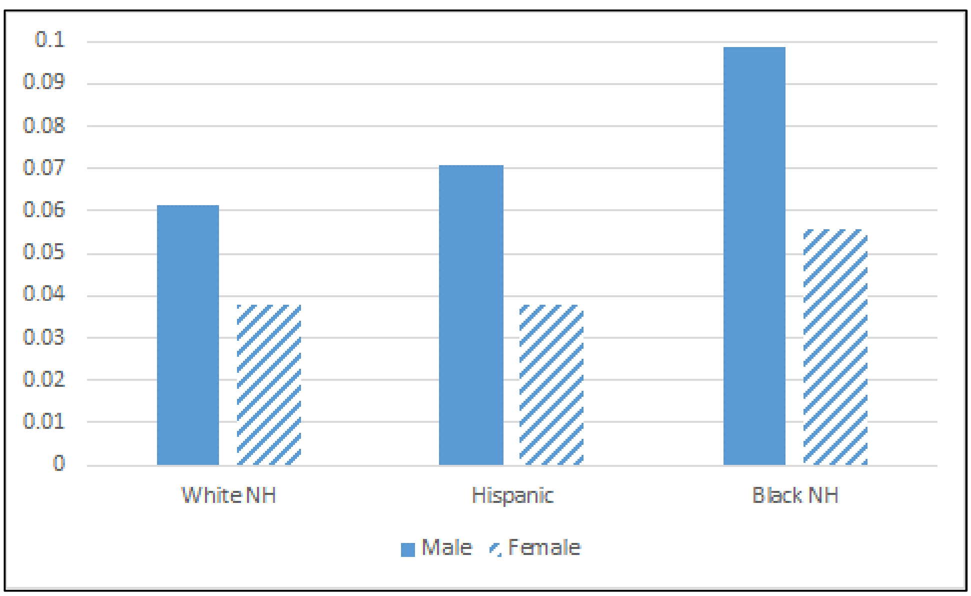
Gender	White NH	Hispanic	Black NH	Total
Male	0.9386	0.9292	0.9012	0.9226
Female	0.9624	0.9623	0.9443	0.958
Total	0.9436	0.9341	0.9062	0.9282

Predicted probability stop improperly premised

	White NH	Hispanic	Black NH	Total
Male	0.0614	0.0708	0.0988	0.0774
Female	0.0376	0.0377	0.0557	0.042
Total	0.0564	0.0659	0.0938	0.0718

Note. N = 3,376. Predicted probabilities from model with main effects with bicycle sidewalk violations INCLUDED and treated as IMPROPERLY premised. Higher probability means greater likelihood that stop was properly premised on reasonable articulable suspicion factors. Source: Jan.-Jun. 2016 legal narratives equal race sample.

Figure 2 Graphical depiction of predicted probabilities for stop premises by gender and race/ethnicity: Bicycle sidewalk violations included and treated as IMPROPERLY premised



Note. N = 3,376. Predicted probabilities from model with main effects. Higher probability means greater likelihood that stop lacked reasonable articulable suspicion factors. Source: Jan.-Jun. 2016 legal narratives equal race sample.

This picture shifts, however, if bicycle sidewalk violations are excluded as investigatory stops. The picture also shifts if bicycle sidewalk violations are treated as proper stops.

8.5.1.2 Bicycle sidewalk violations included, treated as properly premised

Table 16 shows the results with bicycle sidewalk probable cause violations included and treated as properly premised. Race significantly affects the chances of being in a properly premised stop. Non-Hispanic Black stopped civilians' predicted chances were significantly lower ($p < .05$) than the chances of Non-Hispanic Whites. **This is a significant net impact of race.** Stopped Black civilians' odds of being in a [sufficiently vs. improperly premised] stop were predicted to be (1-.448=) 55 percent lower than the corresponding chances for stopped White Non-Hispanic civilians.

Table 16 Main effects model predicting sufficient stop premise: Bicycles on sidewalk Included and treated as properly premised

							OR: 95 % confidence interval	
		B	SE	z	OR	p <	LCL	UCL
Fixed effects								
Black non-Hispanic (= 1; White Non-Hispanic = 0)	dblack	-0.802	0.363	-2.208	0.448	.05	0.220	0.914
Hispanic (= 1; White Non-Hispanic = 0)	dhisp	-0.207	0.395	-0.525	0.813	ns	0.375	1.763
Female (=1; 0 = male)	dfemale	0.388	0.240	1.615	1.474	ns	0.920	2.358
Age (centered by sample mean)	c_age2	-0.007	0.005	-1.460	0.993	ns	0.983	1.003
	Constant	3.333	0.362					
Random effects								
District		Variance	SE of variance					
		0.180	0.108					

LR $\chi 2(df=1)$ vs. logistic model: = 13.35 (df=1); $p < .001$
Note. Outcome = stop properly premised (=1) or not (=0).Weighted data. Bicycles on sidewalk excluded. n = 3,376. Source: Jan.-Jun. 2016 legal narratives equal race sample. Probabilities are two tailed. Results from mixed effects logit model, stops grouped by police districts (run=117)

8.5.1.3 Bicycle sidewalk violations excluded

Weighted results for the main effects model when treating stopped bicyclists on public sidewalks as probable cause rather than investigatory stops, and thus removing them, also yielded a statistically significant impact of race. Results of this model appear in Table 17.

Table 17 Main effects model predicting sufficient stop premise: Bicycles on sidewalk EXcluded.

							OR: 95 % confidence interval	
		B	SE	OR	z	p <	LCL	UCL
Fixed effects								
Black non-Hispanic (= 1; white non-Hispanic = 0)	dblack	-0.791	0.363	0.454	-2.18	.05	0.223	0.924
Hispanic (= 1; white non-Hispanic = 0)	dhisp	-0.209	0.395	.0.811	-0.53	ns	0.374	1.760
Female (=1; 0 = male)	dfemale	0.408	0.240	1.503	1.7	ns	0.939	2.406
Age (centered by sample mean)	c_age2	-0.009	0.005	0.991	-1.69	ns	0.982	1.001
	Constant	3.306	.363	27.270			13.390	55.538
Random effects								
District		Variance	SE of variance					
		0.186	0.111				0.057	0.599

LR $\chi 2$(df=1) vs. logistic model: = 13.71; p < .001
Note. Outcome = stop sufficiently premised (=1) or not (=0).Weighted data. Bicycles on sidewalk excluded. n = 3,300. Source: Jan.-Jun. 2016 legal narratives equal race sample. Probabilities are two tailed. Results from mixed effects logit model, stops grouped by police districts

8.5.1.4 Short aside on bicycle sidewalk violations

To learn a bit more about the impact of how bicycle violations on between-group disparities on observed probabilities that a stop was properly premised or not, the observed proportion of properly vs. improperly premised stops was gauged under two scenarios: with the 66 bicycle sidewalk violations excluded, and with them included but coded as **improperly** premised. This descriptive exploration helps us understand why the net race impact is significant under one option and not under the other.

If the 66 bicycle PC stops are excluded, the percentage of White non-Hispanic stops that were bad was 3.5 (39 out of 1,107 White non-Hispanic stops; 3,310 total). Adding in the 66 PC bicycle stops and coding them as bad jumped the percentage of White non-Hispanic stops that were bad sizably, up to 5.8 percent (66 out of 1,134 white stops; 3,376 total). In contrast the percent of bad black stops was unaffected in these two situations. If the 66 stops are excluded, the percentage of Black non-Hispanic stops that were bad was 8.2 percent (89 out of 1,084 Black non-Hispanic stops; 3,310 total). If the 66 stops are included and coded as bad, the percent of Black non-Hispanic bad stops remains virtually the same at 8.1 percent (89 out of 1,100 Black non-Hispanic stops; 3,376 total). So the difference between the two groups in their respective percentages of bad stops has diminished markedly (8.2-3.5=4.7 percent difference; down to 8.1-5.8=2.3 percent difference). Indeed, the difference in percent bad stops between these two race groups, White vs. Black non-Hispanic individuals, has been cut in half.

8.5.2 Understanding the net statistical impacts of race

To better understand the significant race effect the pattern of marginal race effects over age and gender are examined.

These indicate the "partial change in the probability" of the outcome when race shifts from one group to another and other factors are held constant (Long, 1997: 71). [11] Stated more simply, these are about **just** the **net** impacts of race. The race impact is shown for stopped civilians of different age and gender combinations. Only non-Hispanic civilians are considered. Figure 3 shows the marginal probabilities for non-Hispanic stopped civilians when bicycle probable cause violations are included. Figure 4 shows the same effects from the same main effects model when bicycle sidewalk violations that were probable cause were excluded.

Males appear on the left in each figure, and females on the right. The line shows the estimated marginal net impact of race for persons of different ages.

For example, looking at Figure 3, and considering 15 year old males, the model says the following. The probability that a 15 year old male would be involved in a properly premised stop goes down about five percent if that individual is Black and Non-Hispanic instead of White and non-Hispanic. This predicted impact is due **just** to race **after controlling for** other factors in the model.

The figure also shows that the net race effect becomes somewhat larger as stopped civilian age increases. For example if a stopped male non-Hispanic civilian aged 45 rather than 15, and is Black rather than white, his probability of being in a properly premised stop goes down about six percent rather than five percent.

One more point about the left hand panel in the figure. The lines extending up and down from the net race impact line represent the upper and lower bounds of the 95 percent confidence interval. In the case of males, these intervals do not cross the zero value. This means that in the **full population** of stop records from which this sample was drawn, there is likely to be a **significant net race effect** for **males** regardless of age.

That is not true for females. There, the confidence intervals touch or barely cross zero. So there appears to be **no** significant **net** race effect in the **full population** for non-Hispanic **females.**

In short,

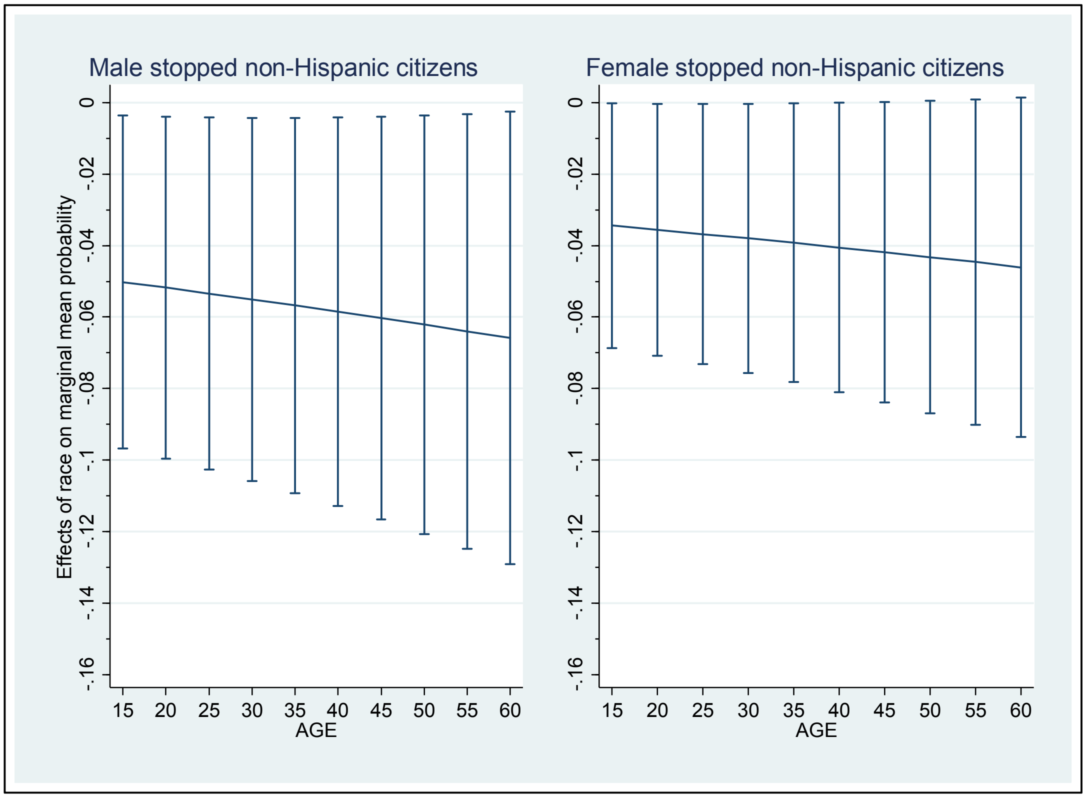
- The results show a five to six percent probability penalty for males who are Black rather than white. Their chances of being involved in a properly premised stop go down by that amount according to the model. This **net** race impact probably applies to the entire population of stops of non-Hispanic civilians.
- The results show a smaller probability penalty for females who are Black rather than White, as compared to males who are Black rather than White. Females' chances of being involved in a properly premised stop go down by about four percent if they are Black rather than White. The net race impact for females may **not** apply to the population of

[11] In Stata, these are generated using the dydx option in marginsplot.

female non-Hispanic records from which the sample was drawn because the confidence intervals for women cross zero. So in the population the estimated “true” net race impact might be zero for non-Hispanic women.

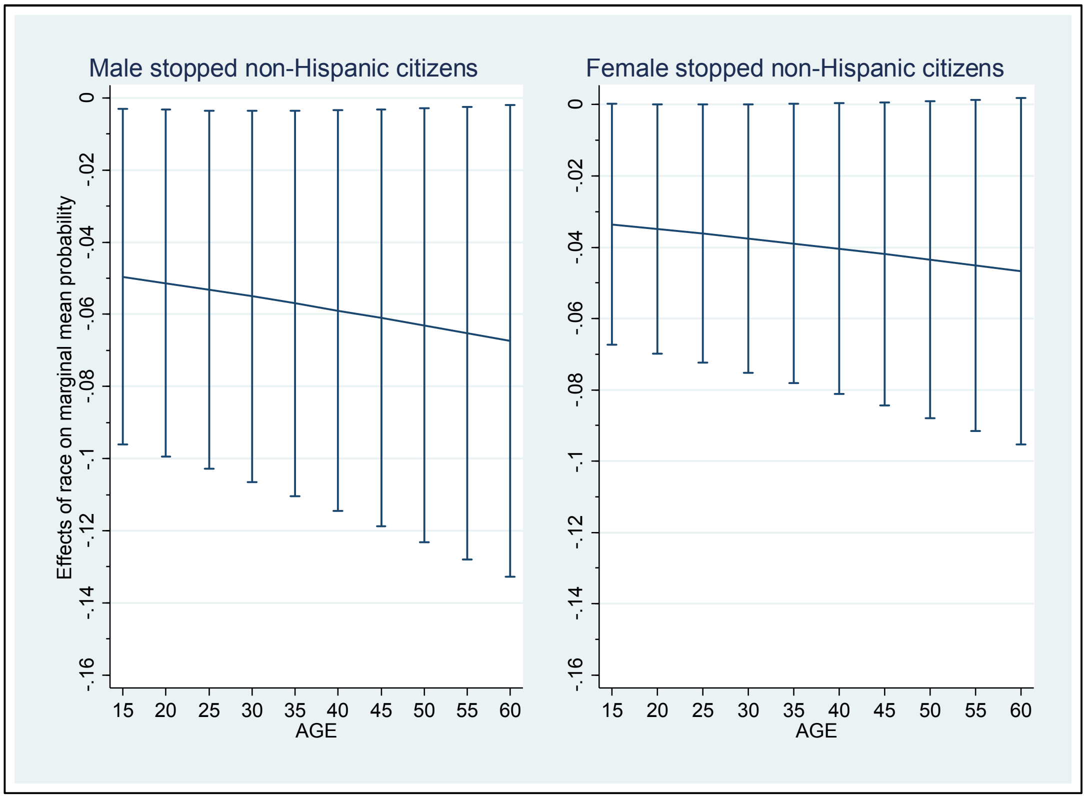
Figure 4 shows the same information when probable cause bicycle sidewalk violations are excluded. The pattern is identical to that already described.

Figure 3. Partial change in probability of a properly premised stop due to race variable: Net impact, bicycles included



Note. Sidewalk bicycle violations included and treated as properly premised (n=3,376). Margins and margin plot generated from full model of main effects, weighted data. 95% upper and lower confidence limits shown. Hispanic stopped civilians excluded. Each data point reflects a predicted impact of switching from a White stopped civilian to a Black stopped civilian on the probability that the stop is properly premised. For males, none of the upper confidence limits cross zero, this is a significant race impact for all the ages shown, for males. Some of the 95 % confidence interval limits appear to cross zero, suggesting the predicted race effect may not be significant for females of all ages.

Figure 4. Partial change in probability of a properly premised stop due to race variable: Net impact, bicycles excluded

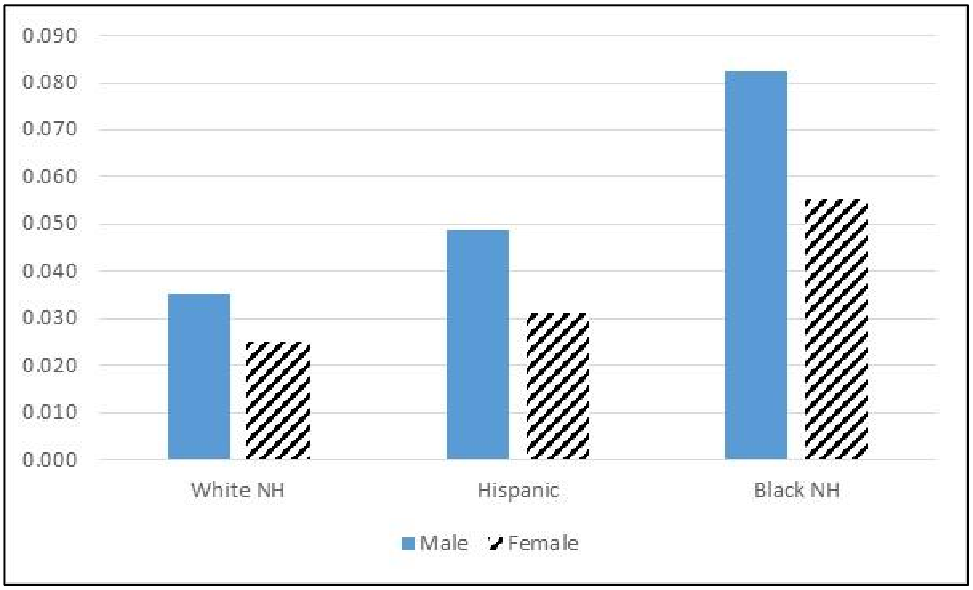


Note. Sidewalk bicycle violations excluded (n=3,310). Margins and margin plot generated from full model of main effects, weighted data. 95% upper and lower confidence limits shown. Hispanic stopped civilians excluded. Each data point reflects a predicted impact of switching from a White stopped civilian to a Black stopped civilian on the probability that the stop is properly premised. Because none of the upper confidence limits cross zero, this is a significant race impact for all the ages shown, for both males and females.

8.5.3 Modeled gross race/ethnicity and gender impacts: Description using predicted probabilities

One can gain a closer appreciation of these patterns of **modeled gross** impacts by examining predicted probabilities from the full model separately for different race, ethnicity and gender combinations. These predicted probabilities, with bicycle sidewalk violations included and treated as properly premised, expressed as the chances that the stop **lacked** reasonable articulable suspicion factors are displayed graphically in Figure 5 and numerically in Table 18. With bicycle sidewalk violations excluded, those patterns are displayed graphically in Figure 6 and numerically in Table 19.

Figure 5 Predicted probabilities stop improperly premised by gender and race/ethnicity: Bicycle sidewalk violations included and treated as properly premised



Note. N = 3,376. Predicted probabilities from model with main effects. Higher probability means greater likelihood that stop lacked reasonable articulable suspicion factors. Source: Jan.-Jun. 2016 legal narratives equal race sample.

Table 18 Predicted probabilities for stop premises by gender and race/ethnicity: Bicycle sidewalk violations included and treated as properly premised

Bicycle sidewalk violations included and treated as properly premised

Predicted probability stop Properly premised

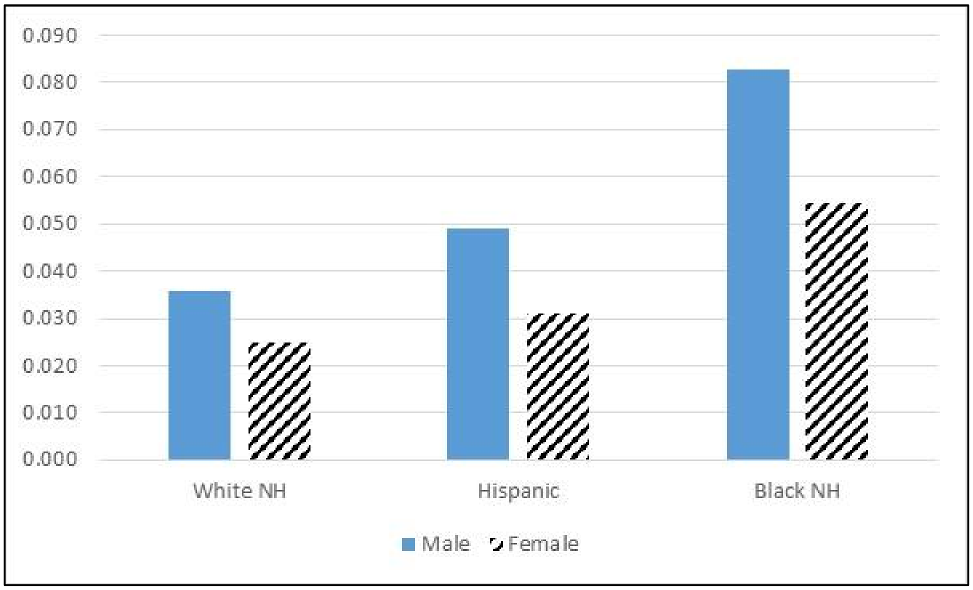
Gender	White NH	Hispanic	Black NH	Total
Male	0.965	0.951	0.918	0.944
Female	0.975	0.969	0.945	0.966
Total	0.967	0.954	0.921	0.948

Predicted probability stop Improperly premised

	White NH	Hispanic	Black NH	Total
Male	0.035	0.049	0.083	0.056
Female	0.025	0.031	0.055	0.034
Total	0.033	0.046	0.079	0.053

Note. N = 3,376. Predicted probabilities from model with main effects. Higher probability means greater likelihood that stop lacked reasonable articulable suspicion factors. Source: Jan.-Jun. 2016 legal narratives equal race sample.

Figure 6 Predicted probabilities stop improperly premised by gender and race/ethnicity: Bicycle sidewalk violations EXcluded



Note. N = 3,310. Predicted probabilities from model with main effects. Higher probability means greater likelihood that stop lacked reasonable articulable suspicion factors. Source: Jan.-Jun. 2016 legal narratives equal race sample.

Table 19 Predicted probabilities for stop premises by gender and race/ethnicity: Bicycle sidewalk violations EXcluded

Bicycle sidewalk violations Excluded

Predicted probability stop Properly premised

Gender	White NH	Hispanic	Black NH	Total
Male	0.964	0.951	0.917	0.944
Female	0.975	0.969	0.946	0.966
Total	0.966	0.954	0.921	0.947

Predicted probability stop Improperly premised

	White NH	Hispanic	Black NH	Total
Male	0.036	0.049	0.083	0.057
Female	0.025	0.031	0.054	0.034
Total	0.034	0.046	0.079	0.053

Note. N = 3,310. Predicted probabilities from model with main effects. Higher probability means greater likelihood that stop lacked reasonable articulable suspicion factors. Source: Jan.-Jun. 2016 legal narratives equal race sample.

These displays show the following.

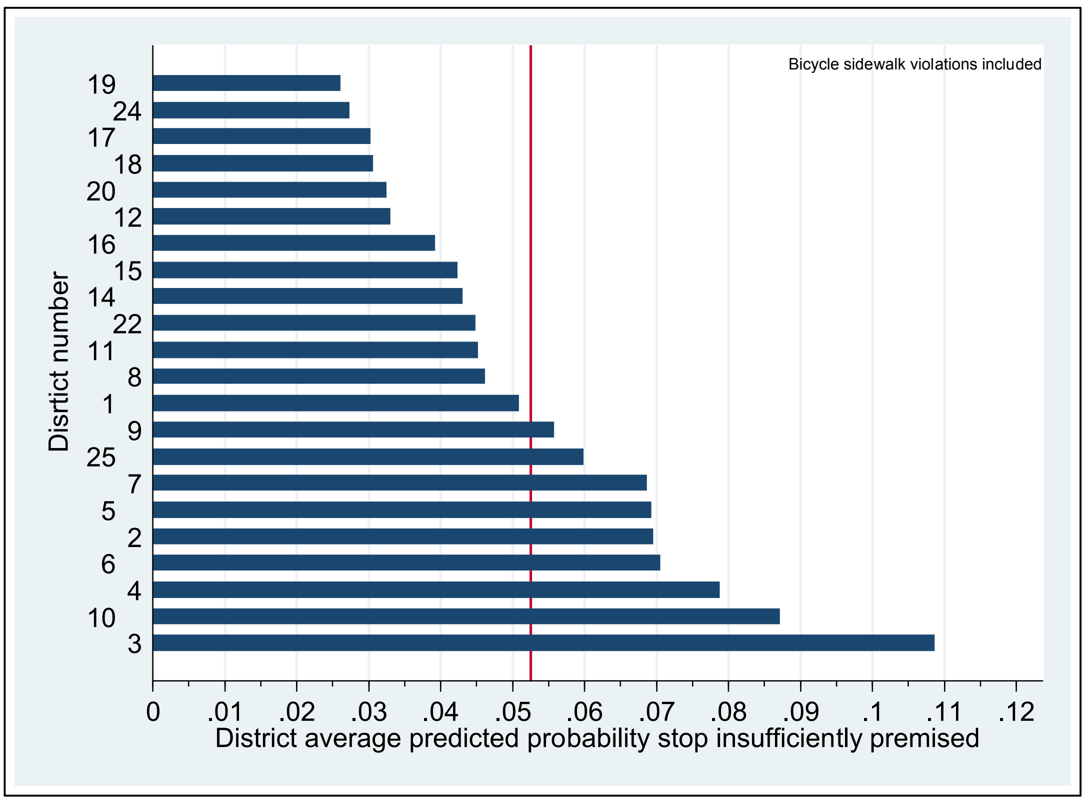
- The pattern proves consistent regardless of whether bicycle sidewalk violations presenting probable cause (n=66) are included or excluded.
- As a group, stopped Black Non-Hispanic civilians' predicted chances of being in an improperly premised stop on average were more than twice the average predicted chances of stopped White Non-Hispanic civilians. This held for both males and females.
- Further, especially for Non-Hispanic Black stopped civilians, average chances of being in an improperly premised stop appeared higher for males than females.

These are gross race and ethnicity impacts which means factors associated with race have not been controlled, nor has district context.

8.5.4 Describing overall geographic patterns in predicted probabilities

District variation in predicted chances that a stop would lack reasonable articulable suspicion factors appears in Figure 7 with bicycle violations included, and in Figure 8 with those records removed. Districts ranged from predicted insufficiency rates that were about half the average predicted insufficiency rate (Districts 19, 24), to those that were about twice the average predicted insufficiency rate (District 3). The pattern was essentially equivalent regardless of how bicycle sidewalk violations were treated.

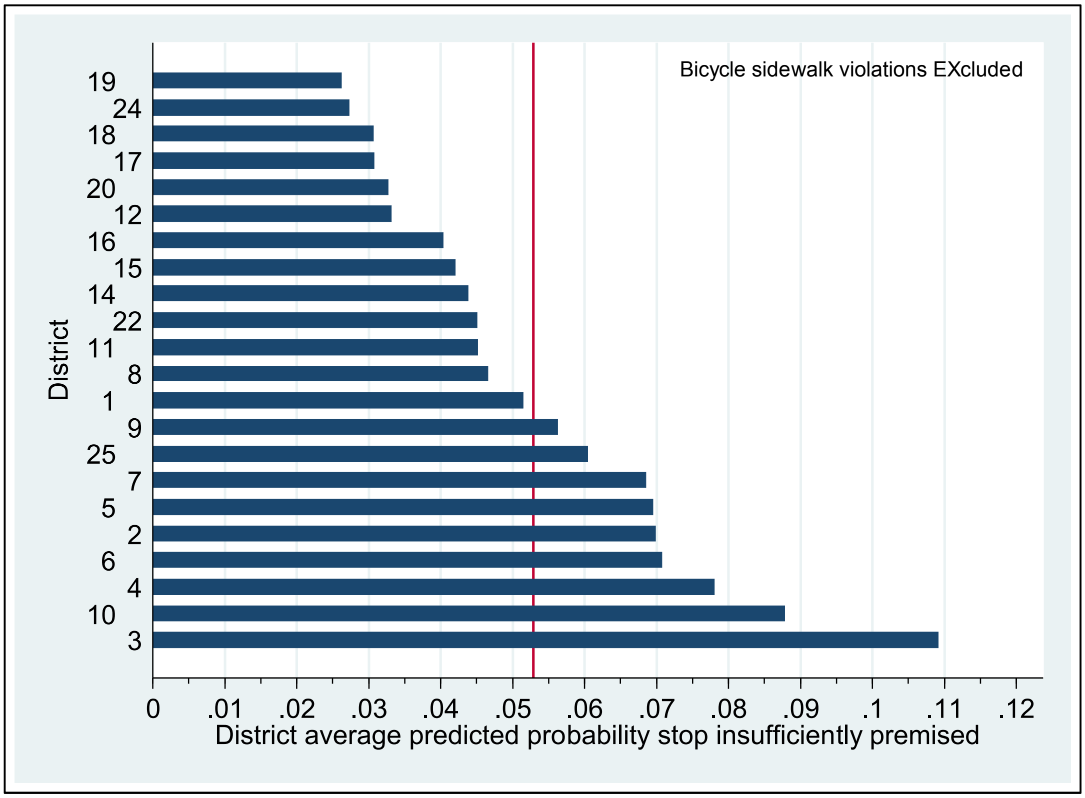
Figure 7 District level average predicted probabilities stops improperly premised: Bicycle sidewalk violations included and treated as properly premised



Note. N = 3,376. Predicted probabilities from model with main effects and weighted data. Higher probability means greater likelihood that stop lacked reasonable articulable suspicion factors.

Source: Jan.-Jun. 2016 legal narratives equal race sample. Vertical reference line represents overall average predicted probability.

Figure 8 District level average predicted probabilities stops improperly premised: Bicycle sidewalk violations EXcluded



Note. N = 3,310. Predicted probabilities from model with main effects and weighted data. Higher probability means greater likelihood that stop lacked reasonable articulable suspicion factors. Source: Jan.-Jun. 2016 legal narratives equal race sample. Vertical reference line represents overall average predicted probability.

The above figures, in essence, display how model predictions play out across different districts, given the factors taken into account by the model: race, ethnicity, age, gender, and district context. They do not indicate what is responsible for these variations. These figures merely describe the variations.

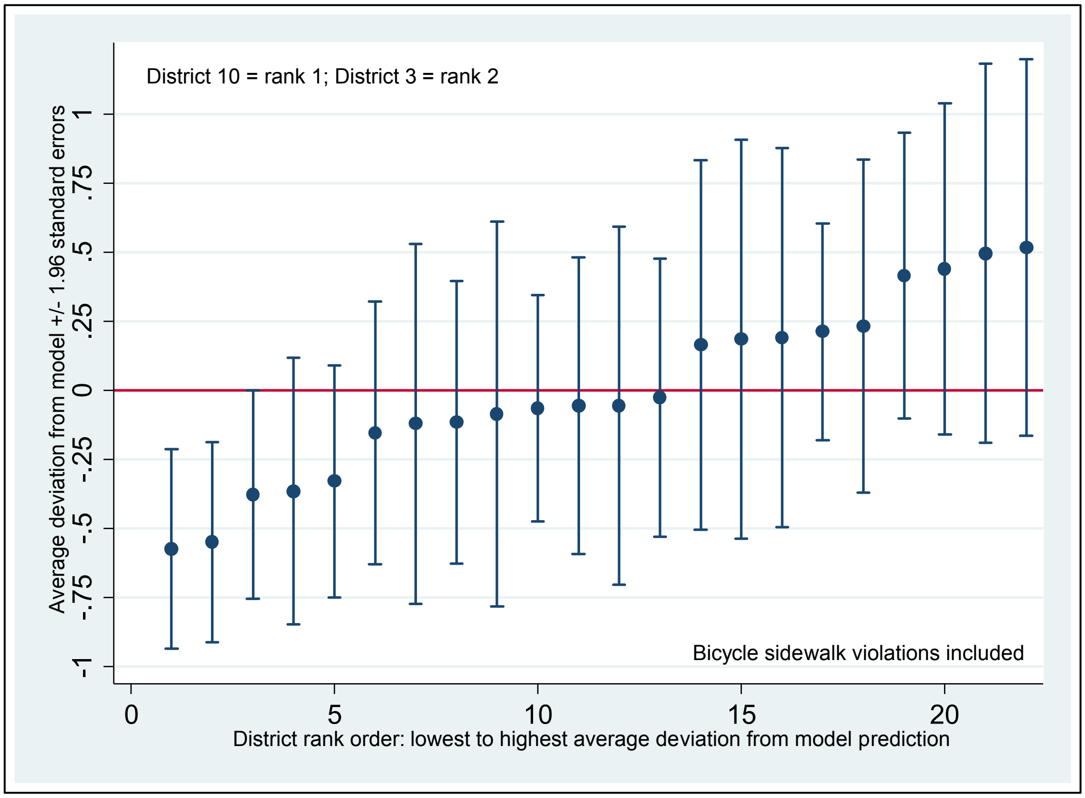
8.5.5 Geographic unexplained variation

The discrepancies between what the model **predicted** should happen with stop basis, and what **actually** happened, are called residuals. These are generated on a case by case basis. These residuals represent deviations from the model prediction. They can be averaged at the district level to capture the district-level average discrepancy from model predictions.

Residuals = [observed score (0 or 1)] – [predicted score (predicted probability)]

Since the outcome was scored zero if a stop was **improperly premised,** a **negative** average residual at the district level suggests that in that district there was a **higher** proportion of stops **lacking** RAS factors.

Figure 9 District-level discrepancies from model predictions of stop sufficiency: Bicycle sidewalk violations included and treated as properly premised



Note. N = 3,376. Residuals capture deviations from predicted probabilities from model with main effects. *Lower* average residual represents *higher* fraction of stops *improperly premised*. Bars capture 95 percent confidence interval around each district's average residual. Horizontal line at zero represents average residual. Confidence intervals not crossing the reference line are significantly different from the average. Source: Jan.-Jun. 2016 legal narratives equal race sample.

District average residuals with bicycle sidewalk violations included appear in Figure 9. The average discrepancy for both District 10 and District 3 was significantly below average. **This means that improperly premised stops in these districts occurred more frequently than expected by the model.**

Potential reasons for the discrepancies are numerous. Only one was examined here: the proportion of stopped civilians who were non-Hispanic Black. That factor at the district level **did not** correlate with these discrepancies.

With bicycle sidewalk violations excluded, the caterpillar plot showed the exact same pattern of significance; Districts 10 and 3 each had an average deviation from the model that fell significantly below zero (results not shown).

8.5.5.1 Robustness tests

The mixed effects models presented here are potentially problematic given the low number of higher level units; there are only 22 districts (Bryan & Jenkins, 2016; Schmidt-Catran & Fairbrother, 2016). Therefore, as a robustness test of the main race effects, single level logit models were run with dummy variables entered for all districts save District 1. These models were problematic in that stops from two districts (18, 19) were dropped because there was no variation on the outcome there. [12]

With the single level models the previously significant net impacts of race (p < .05) became only marginally significant (p < .10) (detailed results not shown). The **size** of the net race effect was closely comparable to what was seen earlier:

- Black OR = .501; z = -1.83; p = .067 (two tailed) with bicycles included and treated as properly premised
- Black OR = .507; z = -1.80; p = .072 (two tailed) with bicycles Excluded

But the impact was no longer significant using a conventional two tailed hypothesis test.

8.6 SUMMARY AND LIMITATIONS: STOP PROPERLY OR IMPROPERLY PREMISED

These analyses of stop premises suggest the following:

1. The likelihood that a stop is properly premised on RAS factors varies significantly across districts (Table 12).
2. A statistically significant (p < .05) net impact of race on stop premise sufficiency emerges. BUT:
3. This significant net race impact on stop premise sufficiency **depends on how probable cause bicycle violation stops are treated**. If they are seen as properly premised, then the net race impact is observed. If they are seen as improperly premised **investigatory** stops, then there is **no significant net race impact.**
4. The significant net effect of race appears, based on the marginal plots, to depend somewhat on gender. Statistically significant net race impacts on the outcome surface regardless of age for males, but not for females (Figure 3, Figure 4).
5. The marginal plots display the **size** of the marginal net race impact, holding other factors constant. **It is a difference of about 4 or 5 percent in the predicted probability that the stop is sufficiently premised. The practical implications of this sized net impact deserve careful discussion. That discussion should take into account the overall rates of properly and improperly premised stops.**
6. Switching from net to gross impacts, average predicted probabilities that a stop lacked sufficient grounds depend on both gender and ethnoracial combinations (Figure 5, Figure 6).

[12] Alternate modeling leaving district 18 as the reference string, and thereby losing only 64 observations had no effect on the net race impact.

7. The group with the highest average model prediction that their stops would be improperly premised are Black Non-Hispanic males (Figure 5, Figure 6).
8. The average model prediction that stops would be improperly premised varies markedly across districts (Figure 7, Figure 8)
9. Two districts (10 and 3) have a higher than expected fraction of improperly premised stops, even after taking into account the factors used by the model (Figure 9).

These analyses have limitations, so results should be interpreted with caution.

10. Most importantly, analyses rely mainly on predicted probabilities and those predicted probabilities rely on a specific model with a specific set of predictors and random effects for districts. Different results could appear with different predictors.
11. **The significant net race impact failed to replicate if we controlled for district context a different way.** Instead of treating districts as random effects, an alternate model entered dummies for each district. This resulted in losing observations from at least one district. It is not clear if the alternative analytics, or the lost observations were the cause of the different result pattern.
12. Mixed effects models presume that random effects of the higher level units represent a normal distribution of effects. This assumption may not be warranted with a relatively small number of higher level units such as we have here. Future analyses probably should be conducted at the beat-within-district level as the geographic unit of analysis.
13. Some might object that by controlling for geography, gender and age we committed the partialling fallacy (Gordon, 1968). The factors controlled for, some might argue, especially geography, were standing in as proxies for race. We don't think this applies for the stop premise outcome because the gross impact of race is about comparable in size to the net impact of race.

9 RESULTS: REASONABLE ARTICULABLE SUSPICION FOR A PAT DOWN

This section examines the relationship between pat down basis and race/ethnicity, before and after controlling for civilian age and gender, and district context. Because some stopped civilians were selected to receive a pat down, and others were not, analyses of the pat down basis need to take that into account. Whether a pat down occurred depended on CPD officers' checking the appropriate box.

There are three possible outcomes:

A. Stopped civilian receives a properly premised pat down.
B. Stopped civilian receives an improperly premised pat down.
C. Stopped civilian does not receive a pat down.

Each model run will generate, for each stopped civilian, a predicted probability, based on the factors in the model, for each of these three outcomes. For each stopped civilian, the three probabilities necessarily sum to 1 or 100 in percent terms. Of greatest interest here are effects of race/ethnicity, controlling for age and gender, on the predicted probability the stopped civilian was subjected to an improperly premised search.

For each specific model, these three outcomes lead to two predictions:

- The chances of B vs. A: The relative risk of receiving an [improperly premised vs. properly premised pat down]. This is called Contrast 1.
- The changes of C vs. A: The relative risk of receiving [no pat down vs. a properly premised pat down]. This is called Contrast 2.

Because only a sample of stops were coded, and because the full report on post stop outcomes analyzes in detail whether a pat down occurs or not, discussion here centers on Contrast 1. [13]

Further, district context also must be taken into account if the proportion of properly premised pat downs varies across districts. As will be seen, it does. The appropriate type of model, therefore, is a multilevel multinomial model with the data weighted so that results reflect the overall population of stops (Rabe-Hesketh & Skrondal, 2012). This is carried out using generalized structural equation models.

9.1 Descriptive pattern

Using weighted data, but excluding probable cause stops including bicycle sidewalk "on view" violations, Table 20 shows differences across the three ethnoracial groups on this outcome.

The group with the highest percentage of records involved an improperly premised pat down was Black non-Hispanic civilians. In the weighted data, 75 out of 2,327 stops in this group or 3.2 percent involved an improperly premised pat down lacking reasonable articulable suspicion. This contrasts with 2.4 percent of the stops of White non-Hispanics that involved a pat down lacking RAS. Least likely to be involved in an improper pat down were Hispanic civilians where only 1.5 percent of their stops involved an improper pat down.

Overall, the weighted data suggest that about 2.8 percent of all investigatory stops of members of these three groups involved an improper pat down. (Of course, there is sampling error around this overall percentage, but it is not shown here.)

[13] Nevertheless, all three outcomes need to be considered simultaneously in one model rather than two models of pairwise comparisons. Otherwise different civilians are in different analyses, and predicted probabilities across the three outcomes for a civilian may not total to 100 percent. (Long, 1997: 151).

Table 20 Descriptive differences, pat down premise

Pat down and basis		Ethnoracial category: White NH	Hispanic	Black NH	Total
Pat down, RAS	Weighted N	51	210	794	1,055
	Percent	20.23	30.05	34.14	32.19
Pat down, no RAS	Weighted N	6	11	75	92
	Percent	2.35	1.52	3.24	2.8
No pat down	Weighted N	196	478	1,457	2,131
	Percent	77.42	68.43	62.63	65
Total	Weighted N	253	698	2,327	3,278
	Percent	100	100	100	100

Note. Equal race sample, Jan-June 2016, weighted data. Investigatory stops only; probable cause stops excluded. NH = non-Hispanic. Percentages shown are column percentages for each group.

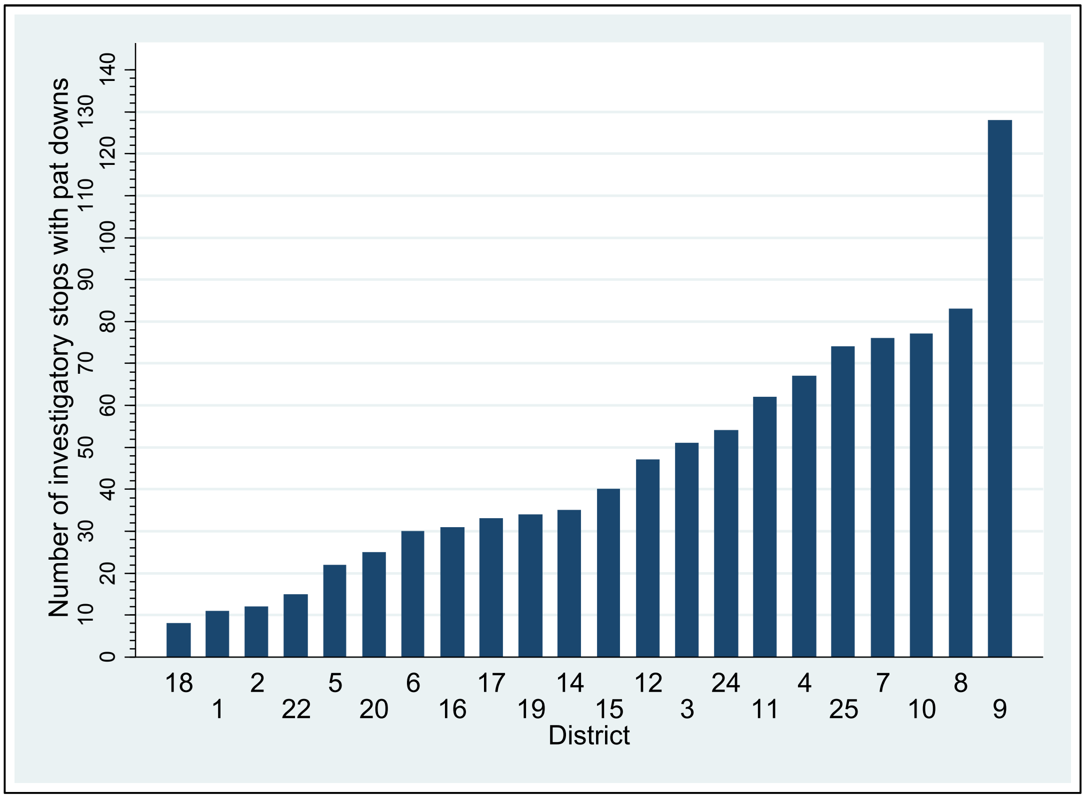
9.2 Patterns of Pat Downs and Pat Down Basis Across Districts

The number of ISRs varies across districts from 44 to 383 with bicycle sidewalk violations included and 44 to 375 with bicycle sidewalk violations excluded. Counts of pat downs across districts appear in Figure 10 with bicycle sidewalk violations included. In the sample, the number of pat downs per district, like the number of stops per district, varies widely across the city.[14] The numbers range from around 10 (Districts 1, 2, 18), to over 100 (District 9).

The number of stops that included an improperly premised pat down in each district appears in Figure 11. The numbers range from zero (Districts 5, 22) to eight (Districts 9, 11, 19).

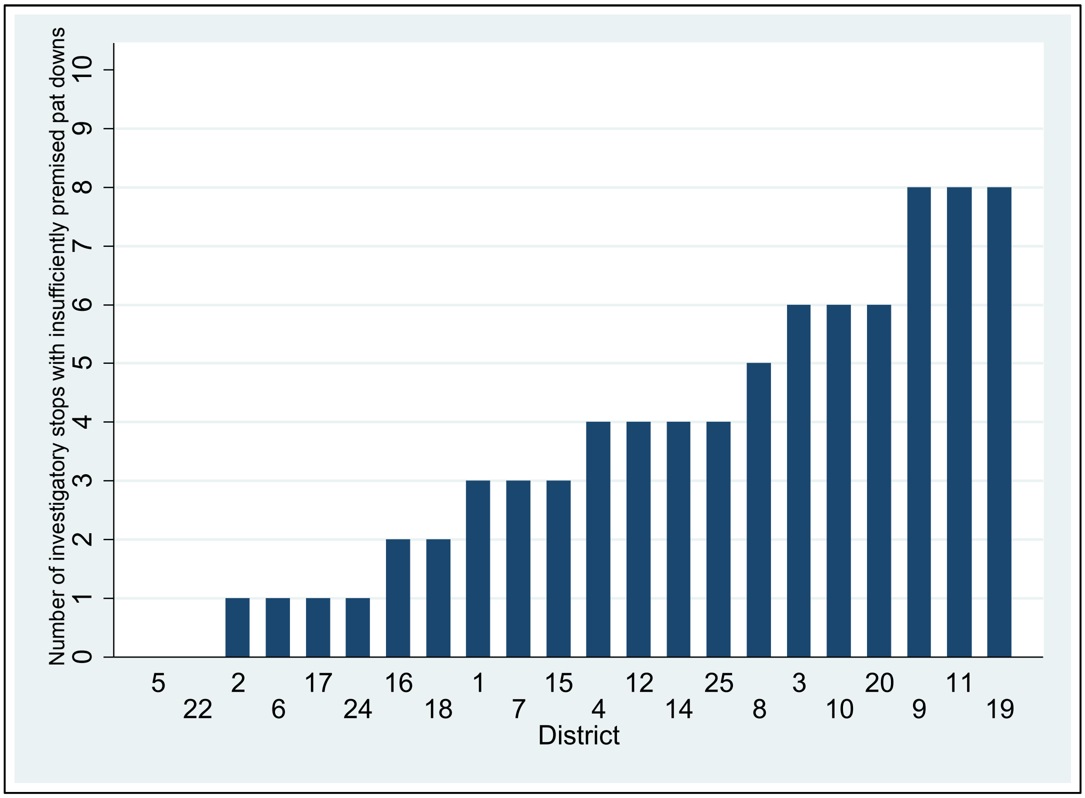
[14] Whether a pat down occurs at all is predicted in the post stop outcomes report.

Figure 10 Number of pat downs per district in sample



Note. Bicycle sidewalk violations included. Source: Jan.-Jun. 2016 legal narratives equal race sample.

Figure 11 Number of improperly premised pat downs per district in sample



Note. Bicycle sidewalk violations included. Source: Jan.-Jun. 2016 legal narratives equal race sample.

9.3 Stop-Level predictors of pat down basis

As mentioned earlier, with the multinomial model and three groups there are two contrasts.

Contrast 1:	Those receiving an improperly premised pat down vs. a properly premised one
Contrast 2:	Those receiving no pat down vs. a properly premised one

For both contrasts, receiving a properly premised pat down is the base category.

Results for Contrast 1 appear in Table 21. There is no significant impact for ethnicity or age. There is, however, a significant impact for gender.

Table 21 Net impacts of race, ethnicity and gender of likelihood of receiving an improperly vs. properly premised pat down, while controlling for district context.

Predictor	B	SE	Z	p <	LCL (95%)	UCL (95%)
Bicycle sidewalk violations included						
Black	0.0308	0.4572	0.07	ns	-0.8653	0.9270
Hispanic	-0.7727	0.5422	-1.43	ns	-1.8354	0.2900
Female	0.8075	0.3679	2.2	.05	0.0865	1.5285
c_age2	-0.0061	0.0099	-0.62	ns	-0.0255	0.0132
_cons	-2.3369	0.4489				
-2 x log likelihood	-2318.67					
Bicycle sidewalk violations excluded						
Black	0.0334	0.4575	0.07	ns	-0.8632	0.9300
Hispanic	-0.7710	0.5425	-1.42	ns	-1.8344	0.2923
Female	0.8071	0.3679	2.19	.05	0.0861	1.5281
c_age2	-0.0057	0.0099	-0.57	ns	-0.0251	0.0137
-2 x log likelihood	-2285.88					

Note. Results from stop-level multilevel multinomial model with stops nested within districts. Results shown only for Contrast 1. Results for contrast 2 not shown. N = 3,372 with bicycle sidewalk violations included, and 3,296 with bicycle sidewalk violations excluded. These numbers of cases differ from other tables because of missing values on this outcome. C_age2 = age centered on sample average (29.55 years). Source: Jan.-Jun. 2016 legal narratives equal race sample

9.3.1 Gender

9.3.1.1 Net impact

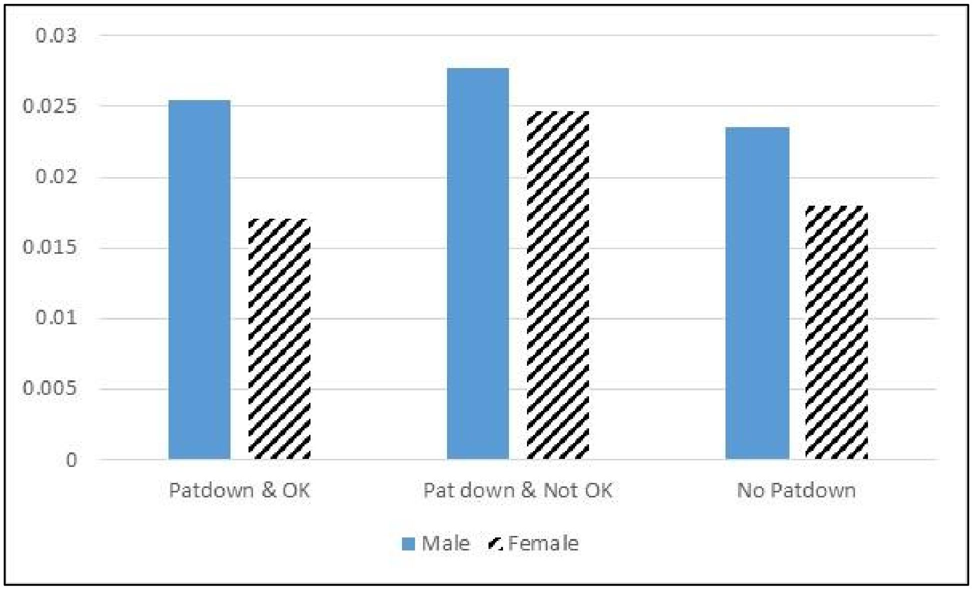
Stopped women are less likely to be patted down than men. But if a stopped civilian of either gender is patted down, the pat down is significantly ($p < .05$) more likely to be improperly premised if the stopped civilian is female. **The chances are better than 95 out of a 100 that in the full set of investigatory stops there is a net gender impact on whether a pat down is properly premised. Pat downs of women are significantly more likely to be improperly premised compared to pat downs of men, controlling for other factors.**

Figure 12 helps clarify. The chart shows the average predicted probability for six groups. Gender is crossed with what actually happened: no pat down, a good pat down, or a bad pat down. The chart shows how gross gender impacts play out in predicted probabilities.

Note the disparity between the two left most bars for females. When *actually* in a *good* pat down, women's predicted probabilities of being in a bad pat down were quite low, about 1.7 percent. But when *actually* in a *bad* pat down, their predicted chances of being in a bad pat down were markedly higher, averaging almost 2.5 percent. By contrast, males' predicted chances of being in a bad pat down were quite similar, regardless of whether they actually were or were not in a good or a bad pat down.

This net gender impact, controlling for civilian race, ethnicity, age, and district context, appears regardless of whether bicycle sidewalk violations are included as investigatory stops (results not shown).

Figure 12 Full model predicted probabilities of a problematic pat down



Note. Actual outcome shown on horizontal axis. Bicycle sidewalk violations included (n=3,376). Predicted probabilities from multilevel multinomial model with stops nested within districts and main effects for race, ethnicity, age, and gender. Source: Jan.-Jun. 2016 legal narratives equal race sample.

9.3.1.2 CAUTION: Descriptive background

More descriptive background may help better contextualize this impact.

Counts by gender for each outcome category appear in Table 22. There are seven improperly premised pat downs of women and 71 of men.

Expressing those numbers as percentages: 1.3 percent of investigatory stops involving women include an improperly premised pat down compared to 2.5 percent of the investigatory stops of men.

The important point here is that t**he significant net gender impact noted in the above table is based only on seven properly premised pat downs of women.**

Table 22 pat down occurrence and premises by gender: Counts, percentages and relevant risks

		Male	Female	
Counts				
	Outcome			
	pat down: Properly premised (G)	865	68	
	pat down: Improperly premised (B)	71	7	
	No pat down (N)	1,900	461	
	Total	2,836	536	3,372
Percentages of total				
	Outcome			
	pat down: Properly premised (G)	0.305	0.127	
	pat down: Improperly premised (B)	0.025	0.013	
	No pat down (N)	0.670	0.860	
		1.000	1.000	

Note. Bicycle sidewalk violations included. Source: Jan.-Jun. 2016 legal narratives equal race sample. Unweighted data. G = "good"; B = "bad"; N = no pat down.

9.3.2 Predicted probability of an improperly premised pat down, geographic context, and race

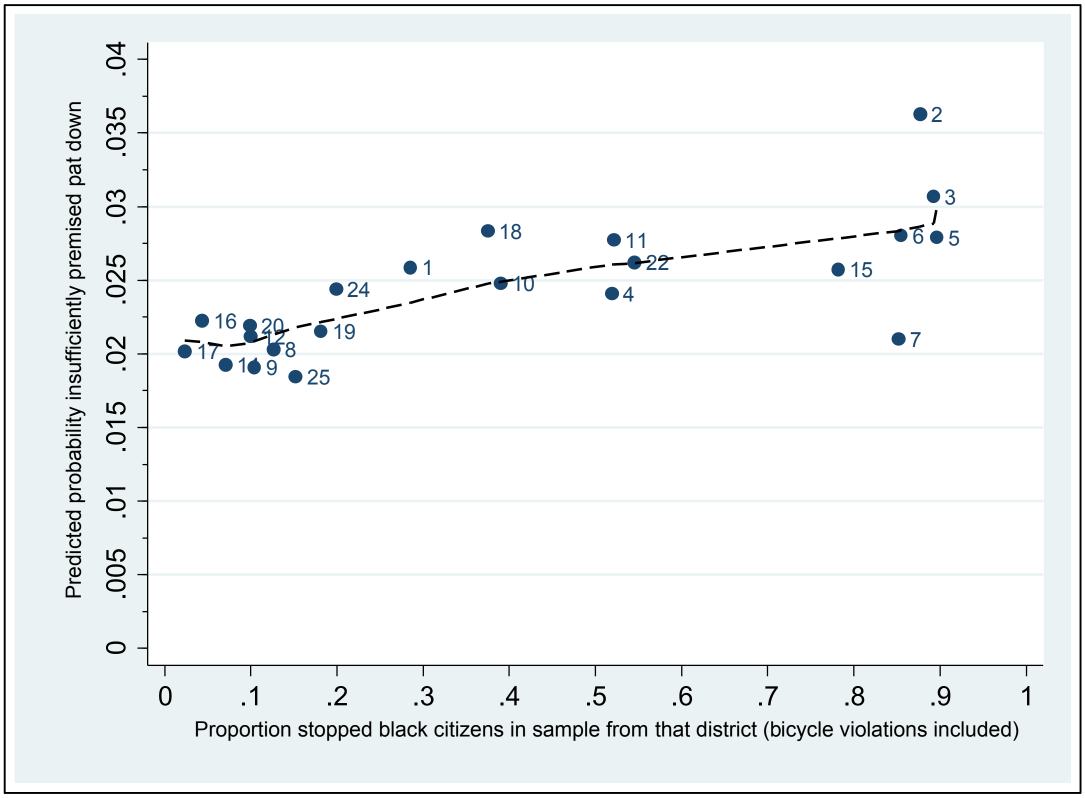
Although the individual level race variable does not link to the odds that a pat down experienced was [improperly premised vs. properly premised], **descriptively** at the **district level** there does appear to be a relationship with race. See Figure 13.

This figure shows the district average predicted probabilities that the pat downs occurring in the district are improperly premised. These range from a little less than two percent to more than four percent.

These district average predicted probabilities are organized by the percent of stopped civilians in the district who were non-Hispanic Black.

The pattern **descriptively suggests** a **district level gross rather than net connection** between the chances of a bad pat down and the racial composition of stopped civilians. The predicted probabilities that pat downs occurring would be poorly premised were higher in districts with a higher proportion of stopped Black non-Hispanic civilians.

Figure 13. District average predicted probabilities pat down improperly premised, and district percent stopped civilians who are black



Note. District average predicted probabilities based on multilevel full model with main effects for race, gender, age and ethnicity, random effect for district (weighted data). Bicycle sidewalk violations included. Line shown is a locally weighted smoothed regression line (Cleveland, 1979). Source: Jan.-Jun. 2016 legal narratives equal race sample.

9.4 Summary and limitations: Pat down basis

These analyses of pat down premise suggest the following points.

1. Regardless of whether bicycle sidewalk violations are included or excluded, women, although they have a lower chance of being patted down at all, if they are patted down, are significantly more likely than men to receive an improperly premised pat down. This is a net gender effect, controlling for district, race, ethnicity, and age. Although it is statistically significant, practically speaking it relies heavily on just seven unwarranted pat downs of women.
2. Further, although there is no stop-level relationship between civilian race and pat down basis.
3. There may, however, be something going on at the district level between pat down premising and race of stopped civilians. At the district level the average predicted probability of an improperly premised pat down does link positively with the fraction of stopped civilians in the district who are Non-Hispanic Black. This is a descriptive gross relationship at an ecological level.

The pat down basis analysis have limitations as well.

4. Most importantly, analyses rely mainly on predicted probabilities and those predicted probabilities rely on a specific model with a specific set of predictors and random effects for districts. Different results could appear with different predictors.
5. Mixed effects models presume that random effects of the higher level units represent a normal distribution of effects. This assumption may not be warranted with a relatively small number of higher level units such as we have here. Future analyses probably should be conducted at the beat-within-district level as the geographic unit of analysis.
6. Some might object that by controlling for geography, gender and age we committed the partialling fallacy (Gordon, 1968). The factors controlled for, some might argue, especially geography, might be standing in as proxies for race. It is just because of this concern that we examine predicted probabilities of a bad pat down by the racial composition of those stopped in the district. That examination suggests pat down basis is more likely to be inadequate in districts with higher proportions of stopped civilians who are Black and non-Hispanic. Because there are so few districts it is not possible to do a meaningful statistical test of this link. In order to better address this limitation for future periods analyses are planned at the beat-within-district level.

10 RESULTS: PROBABLE CAUSE FOR A SEARCH

Attention turns to whether a conducted search during an investigatory stop was properly premised on probable cause, given the descriptions provided in the narratives.

Many investigatory stops resulted in searches that were incident to arrest or transport. In these instances a search was mandated. Therefore, whether the search was premised on probable cause or not was irrelevant in these instances.

10.1 SEARCH FREQUENCY AND BASIS

Table 23 provides information about searches. In the unweighted sample (n=3,310) excluding probable cause bicycle sidewalk violations, searches were conducted for 15.5 percent of the stops (n=512).

Of the 512, focusing just on investigatory stops, and only on records where CPD officers checked the search box, 343 searches (67 percent of the 512) were incident to arrest (n=316) or transport (n=27). In these instances the question of the search being properly premised was irrelevant.

In an additional 44 (8.6 percent) investigatory stops where CPD officers checked the search box, the narratives on the ISR forms provided insufficient information to gauge whether the search was premised on probable cause.

That left 125 (24.4 percent of the 512) searches during investigatory stops, where CPD officers checked the search box, where search premise could be gauged.

In 120 of these 125 (96 percent), the narratives indicated the searches were properly premised on probable cause.

In only 5 instances (4 percent of 125) were the searches deemed improperly premised on probable cause. (With weighted data the number of improperly premised searches was 3.)

10.2 Search Basis and Race/ethnicity

Given the extremely low number of searches improperly premised on probable cause in the sample (n=5), it is not possible to conduct a meaningful analysis examining the relationship between race/ethnicity and this search premise variable.

10.3 Searches and Pat Downs

In the sample using unweighted data, the search outcome links to the previously analyzed pat down outcome. Of the 1,011 investigatory stops resulting in a pat down, in 18.8 percent of them a search also took place (n=190). This contrasts, in the stops without a pat down (n=2,299), where only 14 percent of those stops also involved a search (n=322).

Table 23

Search probable cause basis	Did a search take place? (police check box)		
	No	Yes	Total
0. Sufficient probable cause articulated	0	120	120
1. Sufficient probable cause NOT articulated	0	5	5
Custodial search	0	343	343
INAP (no search) / Insufficient information (search)	2,798	44	2,842
Total	2,798	512	3,310

Note. Unweighted data. Bicycle sidewalk violations excluded.

11 APPENDIX A: CODES FOR STOP RAS, PAT DOWN RAS, AND SEARCH PC

NARRATIVE CODED RESULT stop RAS	Freq.	Percent	Cum.
0. RAS sufficient	3,128	73.90	73.90
1. PC stop no RAS needed	923	21.80	95.70
2. time/distance too attenuated	2	0.05	95.75
4. hunch not personal observation	5	0.12	95.87
7. not enough facts	99	2.34	98.20
8. fleeing or avoidant subject only	2	0.05	98.25
9. no crim activity afoot	57	1.35	99.60
11. no basis for terry or PC stop	13	0.31	99.91
.i	4	0.09	100.00
Total	4,233	100.00	

12 REFERENCES

Bryan, M. L., & Jenkins, S. P. (2016). Multilevel Modelling of Country Effects: A Cautionary Tale. *European Sociological Review, 32*(1), 3-22. doi:10.1093/esr/jcv059

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association, 74*, 829-836.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159.

Fader, J. J., & Traylor, L., L. (2015). Dealing with differences in desistance theory: The Promise of intersectionality for new avenues of inquiry. *Sociological Compass, 9*(4), 247-260.

Gordon, R. A. (1968). Issues in multiple regression. *American Journal of Sociology, 73*, 592-616.

Klinger, D. A. (1997). Negotiating order in patrol work: An Ecological theory of police response to deviance. *Criminology, 35*(2), 277-306.

Laub, J. H., & Sampson, R. J. (2003). *Shared Beginnings, Divergent Lives : Delinquent Boys to Age 70*. Cambridge: Harvard University Press.

Long, J. S. (1997). *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks: Sage.

Rabe-Hesketh, S., & Skrondal, A. (2012). *Multilevel and longitudinal modeling using Stata Volume II: Categorical responses, counts and survival. Third edition (ISBN 9781597181044)*. College Station, TX.: Stata Press.

Raftery, A. E. (1995). Bayesian Model Selection in Social Research. *Sociological Methodology, 25*, 111-163.

Rosenfeld, R., Rojek, J., & Decker, S. (2012). Age Matters: Race Differences in Police Searches of Young and Older Male Drivers. *Journal of Research in Crime and Delinquency, 49*(1), 31-55. doi:10.1177/0022427810397951

Schmidt-Catran, A. W., & Fairbrother, M. (2016). The Random Effects in Multilevel Models: Getting Them Wrong and Getting Them Right. *European Sociological Review, 32*(1), 23-38. doi:10.1093/esr/jcv090