A Reproduction of "Do Female Officers Police Differently? Evidence from Traffic Stops" (*American Journal of Political Science*, 2021)

Dianyi Yang [1] and Leike Huang [2]

[1]Department of Government, London School of Economics and Political Science
[2]Oxford School of Global and Area Studies, University of Oxford

December 11, 2023

Abstract

We reproduce Shoub, Kelsey, Katelyn E. Stauffer, and Miyeon Song (May 2021). "Do Female Officers Police Differently? Evidence from Traffic Stops," with alternative specifications and interpretation of the results. While our reproduction confirms that female police officers are less likely to search drivers than male officers and female officers are more likely to find contraband upon a search, we re-evaluate the authors' claims on the equality of effectiveness between male and female officers and find that female officers in the dataset confiscated less contraband than male officers.

Keywords— Police-Citizen Interaction, Feminism, Public Administration, Public Opinion, United States

1 Introduction

This paper re-examines the methodology and conclusion of Shoub, Stauffer, and Song's (2021) article: *Do Female Officers Police Differently? Evidence from Traffic stops.* As pointed out by the authors, police-citizen contact has long been the focus of political science and public administration for its political sensitivity and impacts on the legitimacy of the justice system and the broader political institutions, especially in the context of America. Traffic stops, being common in daily interaction and rich in data, have been studied extensively by the previous literature.

Whereas most previous research focused on the characteristics of the drivers, the impacts of *individual officer* characteristics on police-citizen contact have rarely been studied. The scope of those few articles that discuss this area solely focuses on officers' race and specific agency policies but ignores officer sex, another important factor that might have influenced how officers interact with the citizens. Shoub, Stauffer, and Song (2021) filled this gap in the literature by demonstrating how officers of different sexes behave differently in traffic stops.

To undertake this research, Shoub, Stauffer, and Song (2021) used data from traffic stop logs of the Florida State Highway Patrol (from 2010 to 2015) and the Charlotte (North Carolina) Police Department (from 2016 to 2017 and 2019 to 2020). The two offices are chosen due to the completeness of their data, which contains the necessary information to undertake this research. These include information such as whether a search is performed, the sex of the officer conducting the traffic stop, and other information that could be controlled for such as initial stop purposes, driver' s races, etc. These datasets, originating from two distinct law enforcement bodies —a municipal police force and a statewide institution —also enabled the researchers to discern if the institutional nature of a police agency might influence the study outcomes. Given that in both Florida and Charlotte, NC are marked by their diversity, and other studies indicate similar traffic stop trends nationwide, the choice of these locations arguably offers a robust representation of the broader U.S. landscape.

The authors have designated officers' behaviours when carrying out their duties as the primary dependent variable. More importantly, they investigate the occurrence of a search following a traffic stop and the frequency of finding contraband following a search. The scope of these searches is narrowed down to those based on probable cause, consent and those conducted within plain sight, deliberately sidelining those instances where officers weren't operating under discretionary powers. The foremost independent variable under consideration is the sex of the officer, identified as male or female.

The primary method used by Shoub, Stauffer, and Song (2021) is the Ordinary Least Squares (OLS) regressions with Two-Way Fixed Effects (TWFE). The TWFE OLS model estimates, on average, given the year and place, the linear change in the likelihood of a driver being searched given they are stopped, or contraband is found when a search is performed and if the officer is female as compared to if the officer were male. This is complemented by robustness checks using other statistical methods in the Supplementary Material, such as logistic regressions, matching and hierarchical linear models.

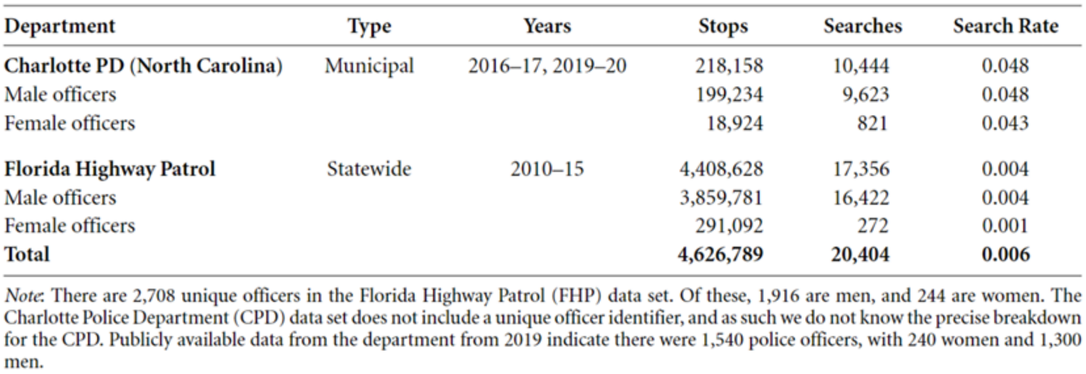
We therefore reassess the following main claims of the original paper:

1. Female police officers are less likely to search drivers than male officers given a traffic stop.
2. Female police officers are more likely to find contraband when they conduct a search.
3. Despite lower search rates, female officers confiscate the same net amount of contraband as male officers.
4. Female officers can minimise the number of negative interactions with citizens without compromising effectiveness.

Whereas our reproduction largely confirms the first two claims, it does not confirm the third claim of the original paper: our results show that female officers confiscate *less* contraband than male officers. Consequently, the fourth claim, which is based on the validity of the third claim, is also in question - rather than being a Pareto improvement, female officers could imply a trade-off between benign police-citizen contact and effectiveness.

2 Minor Discrepancy in Reproduction

We have successfully reproduced all figures and tables in Shoub, Stauffer, and Song (2021) except for Table 1. As shown in Tables 1a and 1b below, there are two major differences between the original paper and our reproduction. The total number of stops is 4,626,786 in our reproduction (Table 1b) but 4,626,789 in the original paper (Table 1a). Moreover, the total number of searches in our reproduction is 27,800, while the number in the paper is 20,404, which might be due to an error in calculation or typo. Furthermore, the search rate of 0.006 is incompatible with 20,404 searches divided by 4,626,789 stops as per the original table (which yields a false search rate of 0.004), while it matches our correction.



Department	Type	Years	Stops	Searches	Search Rate
Charlotte PD (North Carolina)	Municipal	2016–17, 2019–20	218,158	10,444	0.048
Male officers			199,234	9,623	0.048
Female officers			18,924	821	0.043
Florida Highway Patrol	Statewide	2010–15	4,408,628	17,356	0.004
Male officers			3,859,781	16,422	0.004
Female officers			291,092	272	0.001
Total			**4,626,789**	**20,404**	**0.006**

Note: There are 2,708 unique officers in the Florida Highway Patrol (FHP) data set. Of these, 1,916 are men, and 244 are women. The Charlotte Police Department (CPD) data set does not include a unique officer identifier, and as such we do not know the precise breakdown for the CPD. Publicly available data from the department from 2019 indicate there were 1,540 police officers, with 240 women and 1,300 men.

Table 1a: Summary of Stops and Searches by Agency (original)

Department	Type	Years	Stops	Searches	Search Rate
Charlotte PD (NC)	Municipal	2016-17, 2019-20	218,158	10,444	0.048
Male Officers			199,234	9,623	0.048
Female Officers			18,924	821	0.043
Florida Highway Patrol	Statewide	2010-15	4,408,628	17,356	0.004
Male Officers			3,859,791	16,422	0.004
Female Officers			291,092	272	0.001
Total			**4,626,786**	**27,800**	**0.006**

Table 1b: Summary of Stops and Searches by Agency (reproduction)

3 Corrections for Heteroskedasticity, Autocorrelation and Bias

A potential minor issue with the original analysis by Shoub, Stauffer, and Song (2021) is that it does not account for heteroskedasticity and autocorrelation for the linear probability model or correct for the incidental parameter bias for the fixed-effect logistic model. Although our refined analysis shows virtually the same results as the original paper, such corrections are necessary with regard to methodological validity.

The OLS model is known as a linear probability model when it is applied to a binary outcome (Y) rather than a continuous one (Stock and Watson 2014, p.434). The analysis of the sex difference in policing by Shoub, Stauffer, and Song (2021) belongs to this category as whether a search takes place or whether contraband is found is a binary outcome. However, unlike when the OLS model is applied to a continuous outcome, the errors of the linear probability model are *inherently* heteroskedastic[1] (Stock and Watson 2014, p.434). Empirically, this can be shown via a simple Non-Constant Variance Test, which is also known as the Breusch-Pagan test (Breusch and Pagan 1979).

Here, we use the regressions from Table 2 of the original paper (OLS Regressions Explaining Searches Following a Traffic Stop) as examples and test for their heteroskedasticity; results are summarised in Table 2a. The p-values of less than 0.001 suggest that we have sufficient evidence for the heteroskedasticity of errors, as opposed to the homoskedastic null.

Therefore, the original results which assume homoskedasticity of errors (the results from the original paper, presented in the "IID" columns of Table 2b) should be replaced by our results using heteroskedasticity-robust standard errors (the "Robust" columns of Table 2b). However, it can be seen that the difference between the IID and robust standard errors does not manifest in three decimal places (results in the table are identical).

Moreover, in the dataset used by Shoub, Stauffer, and Song (2021), the same units are observed across different time periods, which is known as a panel data structure. The authors have controlled for time and

[1] i.e. variance of errors depend on regressors.

Department	Chi-squared Statistic	Degrees of Freedom	p-value
Charlotte PD (NC)	82,355.1	1	<0.001
Florida Highway Patrol (FHP)	2,570,064.2	1	<0.001

Table 2a: Test for Non-Constant Variance for OLS Regressions Explaining Searches Following a Traffic Stop

	CPD			**FHP**		
	IID	Robust	Clustered	IID	Robust	Clustered
Female Officer	−0.026***	−0.026***	−0.026**	−0.004***	−0.004***	−0.004***
	(0.002)	(0.002)	(0.006)	(0.000)	(0.000)	(0.001)
(Intercept)	0.086***	0.086***	0.086***	0.026***	0.026***	0.026***
	(0.004)	(0.004)	(0.012)	(0.001)	(0.001)	(0.002)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Division fixed effects	Yes	Yes	Yes	No	No	No
County fixed effects	No	No	No	Yes	Yes	Yes
R^2	0.071	0.071	0.071	0.009	0.009	0.009
Adjusted R^2	0.071	0.071	0.071	0.009	0.009	0.009
N	150,547	150,547	150,547	2,712,478	2,712,478	2,712,478

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Table 2b: OLS Regressions Explaining Searches Following a Traffic Stop

	CPD - WCB
Female Officer	−0.026 (<0.001)***
	[−0.040, −0.013]
Control	Yes
Year fixed effects	Yes
Division fixed effects	Yes
County fixed effects	No
R^2	0.071
Adjusted R^2	0.071
N	150,547

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

P-values are in parentheses; 95% confidence intervals are in square brackets.

CIs are calculated through Wild Cluster Bootstrapping by Division with Webb weights.

Table 2c: OLS Regression with WCB inference Explaining Searches Following a Traffic Stop for CPD

geographical fixed effects to exploit the panel nature of the data. However, one potential problem with panel data is autocorrelation (or serial correlation), as later observations may be correlated with earlier ones. One solution to this problem is the *clustered standard errors*, which are robust to both heteroskedasticity and autocorrelation (Stock and Watson 2014, p.413). Therefore, we also provide results using clustered standard errors at the Division and County levels for the CPD and the FHP datasets, respectively. These results are shown in conjunction with the previous results (see the "Clustered" columns of Table 2b). Although the clustered standard errors are larger, the results are still very statistically significant, which implies the robustness of the original results.

One concern about using clustered standard errors with the CPD dataset is that the number of clusters (13) is small. Webb (2014) showed that clustered standard errors and their previous bootstrap variants perform poorly with few clusters and proposed a 6-point bootstrap weight distribution for inference. We use this alternative inference method for the CPD dataset in Table 2c. Note that a 95% confidence interval and p-value are provided in place of standard errors, as wild cluster bootstrapping methods do not assume a predetermined distribution.

Since the OLS model is not dedicated to binary outcomes, Shoub, Stauffer, and Song (2021) also provided results using the logistic regression in the Supplementary Material, which is more suited for such scenarios but the results of which are more difficult to interpret. However, the use of the non-linear models (including the logistic regression) *in tandem with* fixed effects leads to the **incidental parameter problem** (Neyman and Scott 1948). The intuition is that the fixed effects cannot be eliminated through first differencing in non-linear models, which is viable in linear models. Therefore, we correct for such bias using the method developed by Fernandez-Val and Weidner (2018). We display the original coefficient results using the logistic regression from the authors alongside the bias-corrected coefficients and the bias-corrected average partial effects in Table 2d. The results suggest that the difference between the original and the bias-corrected results is not discernible, and the average partial effects from the logistic regressions are comparable to those from the linear probability model.

	CPD			**FHP**		
	Logistic	Bias Corrected	APE	Logistic	Bias Corrected	APE
Female Officer	-0.470***	-0.470***	-0.018***	-1.425***	-1.425***	-0.004***
	(0.048)	(0.048)	(0.002)	(0.067)	(0.067)	(0.000)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Division fixed effects	Yes	Yes	Yes	No	No	No
County fixed effects	No	No	No	Yes	Yes	Yes
Deviance	49844.205	49844.205	49844.205	137716.157	137716.157	137716.157
N	150,547	150,547	150,547	2,712,478	2,709,925	2,709,925

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2d: Logistic Regressions Explaining Searches Following a Traffic Stop

	OLS		Logistic (Bias Corrected)		Logistic (Original)
	IID	Clustered	APE	Coefficient	Coefficient
Female Officer	0.103***	0.103**	0.094**	0.496**	0.499**
	(0.029)	(0.030)	(0.030)	(0.155)	(0.155)
(Intercept)	0.112**	0.112*			-17.981
	(0.042)	(0.048)			(148.705)
Controls	Yes	Yes	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes	Yes	Yes
County fixed effects	Yes	Yes	Yes	Yes	Yes
R^2	0.135	0.135			
Adjusted R^2	0.127	0.127			
Deviance			13267.483	13267.483	13267.483
N	12,782	12,782	12,777	12,777	12,782

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4a: Regressions Explaining Probability of Finding Contraband Given a Search

We also inspect the results in Table 4 of the original paper (OLS Regressions Explaining Probability of Finding Contraband Given a Search) in a similar fashion.

We first reproduce the results in Column 1 (Pr(Contraband Found|Search)) of the original Table 4 with clustered standard errors and the bias-corrected logistic regression. The results are summarised in Table 4a. Again, our additional robustness checks confirm the results produced by Shoub, Stauffer, and Song (2021). Not only do the results using clustered standard errors match the original results (using IID standard errors), but the estimated average partial effect from the bias-corrected logistic regression suggests a similar magnitude to the original OLS estimate.

We then move on to inspect the results on the hit rate per 10 searches. Here, each observation is an individual officer operating in a specified context. Since no county-fixed effects are added to the original OLS regression, we do not cluster by county and only provide robust standard errors for the OLS model. Shoub, Stauffer, and Song (2021) did not provide logistic regression result for this scenario (presumably because the outcome is a proportion), and we have thus tried to provide a comparable logistic regression[2]. Here, the coefficient still indicates the effect on the log odds of finding contraband (unaffected by the ”per 10 searches”), but the APE is multiplied by 10 to be comparable to the OLS results per 10 searches. The results are summarised in Table 4b, which suggests that the OLS and logistic regression results are similar.

We similarly reproduce the results on hit rate per 100 stops with robust standard errors and the logistic regression. The results (in Table 4c) show that the estimated average partial effect (APE) from the logistic regression is greater in magnitude than the OLS estimate. The interpretation of the results on hit rate per 100 stops is further discussed in Section 5.

4 Alternative Predictions

Although our reproduction largely confirms Shoub, Stauffer, and Song’s (2021) results on the effects of officer sex on traffic stop outcomes, our analysis provides different *predictions* using the same results.

[2] The logistic regression results are not corrected for bias as the R package for bias correction (alpaca) does not allow for proportions as outcomes for the logistic regression

	Hit Rate per 10 Searches			
	OLS		Logistic	
	IID	Robust	APE	Coefficient
Female Officer	1.122***	1.122***	1.089***	0.572***
	(0.276)	(0.314)	(0.290)	(0.147)
(Intercept)	0.301	0.301		-18.491
	(0.215)	(0.191)		(154.419)
Controls	Yes	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes	Yes
County fixed effects	No	No	No	No
R^2	0.131	0.131		
Adjusted R^2	0.128	0.128		
Deviance			10454.616	10454.616
N	9,677	9,677	9,677	9,677

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4b: Regressions Explaining Frequency of Finding Contraband per 10 Searches

	Hit Rate per 100 Stops			
	OLS		Logistic	
	IID	Robust	APE	Coefficient
Female Officer	-0.077***	-0.077***	-0.099***	-1.085***
	(0.012)	(0.009)	(0.006)	(0.101)
(Intercept)	0.138***	0.138***		-23.224
	(0.018)	(0.019)		(96.223)
R^2	0.004	0.004		
Adjusted R^2	0.004	0.004		
Deviance			30881.458	30881.458
N	747,784	747,784	747,784	747,784

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4c: Regressions Explaining Frequency of Finding Contraband per 100 Stops

Shoub, Stauffer, and Song (2021) claimed that, on average, male officers are 225% and 272%[3] more likely than female officers to conduct a search upon a traffic stop in the Charlotte Police Department and Florida Highway Patrol, respectively. However, these relative odds are predicted based on Table 2b regressions, with variables other than officer sex held to their means and *modes*.

Despite being common practice, the use of *modes* for categorical variables is problematic in this case as they do not represent a ”typical” case as intended. Taking the Charlotte Police Department as an example, the predicted scenario is when the officer and driver are both white, the driver is driving a local car on an investigatory search, and the stop took place in Jan 2019, in “South Division”. However, a closer examination of the dataset reveals that “South Divison”, as the modal value of the police division variable, appears barely more than other values (See Figure 1a).

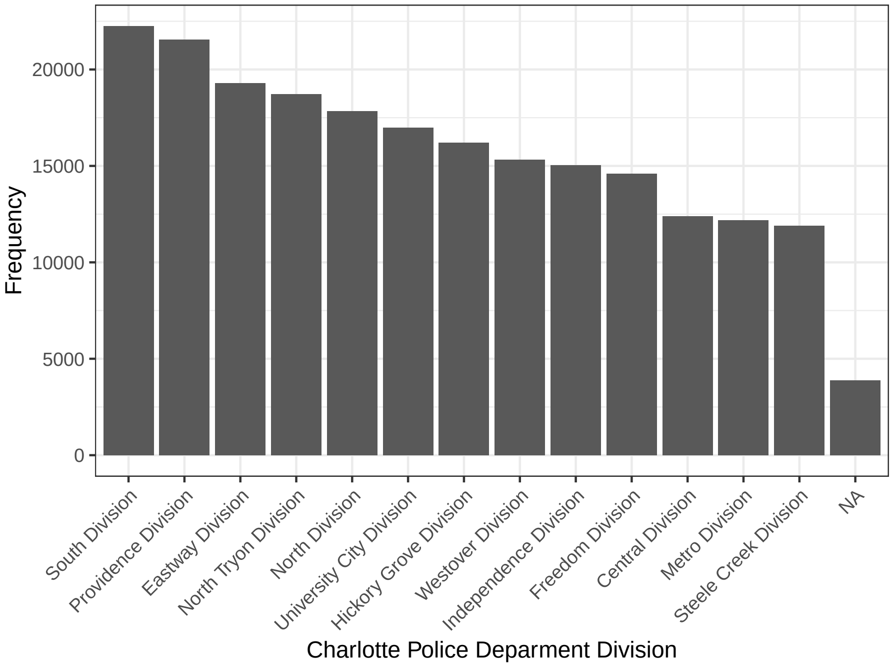


Figure 1a: Frequency by Police Division in the Charlotte Police Department Dataset

Moreover, as for the estimated fixed effects, “South Division” ranks the third lowest among all divisions, as shown in Figure 1b. This implies that Shoub, Stauffer, and Song’s (2021) predictions based on the South Division as the modal value would have most likely *underestimated* the probabilities of being searched upon a traffic stop by police for the Charlotte Police Department.

To remedy this bias, we draw 1000 random samples from each dataset with replacement (bootstrapping) and predict these sample values for male and female officers separately. Finally, we calculate the mean values for point estimates and prediction variances to construct more representative predictions and provide an alternative Figure 2 to the original paper. Here, we use the logistic regressions[4] alongside the OLS regression used by the original authors for comparison. In Figure 2, the original estimates from Shoub, Stauffer, and Song (2021) are shown in red dashed lines with triangles; predictions using the same mean and mode values are presented in blue dashed lines with triangles; Bootstrapped OLS and logit estimates are given in red and blue solid lines with round points, respectively. It can be seen that our bootstrapping method provides estimates that are more consistent between logistic and OLS models and has addressed the underestimation problem in the predictions for the Charlotte Police Department.

According to our bootstrapped predictions with logistic regressions, male officers are, on average, 51% and 308%[5] more likely than female officers to conduct a search upon a traffic stop in the Charlotte Police Department and Florida Highway Patrol, respectively.

5 Alternative Interpretation

Our main divergence from the original paper lies in the interpretation of the results on hit rate per 100 stops (see Table 4c). The original paper dismissed the difference in hit rate per 100 stops between officer sexes as trivial despite the coefficient from the OLS regression being significant at the 0.001 level. Shoub, Stauffer,

[3] In the original paper, the authors wrote “Male officers are over 272% *more* likely to conduct a search than female officers...” (p.762) Drawing from the results, we think the authors meant 272% *as* likely or 172% *more* likely.

[4] We confess that we did not use the bias-corrected versions of the logistic regressions as the R package does not allow predictions with new data. However, as we show in Table 2d, such bias is negligible in this case.

[5] or 84% and 335% using OLS regressions.

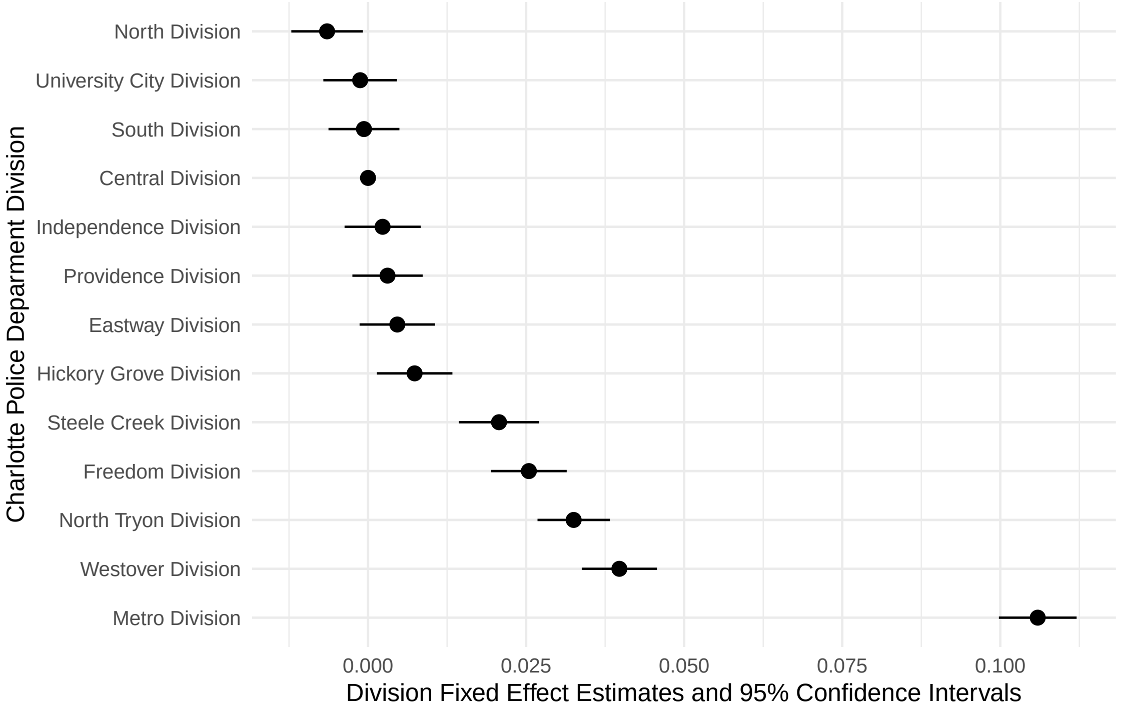


Figure 1b: Estimated Police Division Fixed Effects in the Charlotte Police Department OLS Regression Explaining Searches Following a Traffic Stop

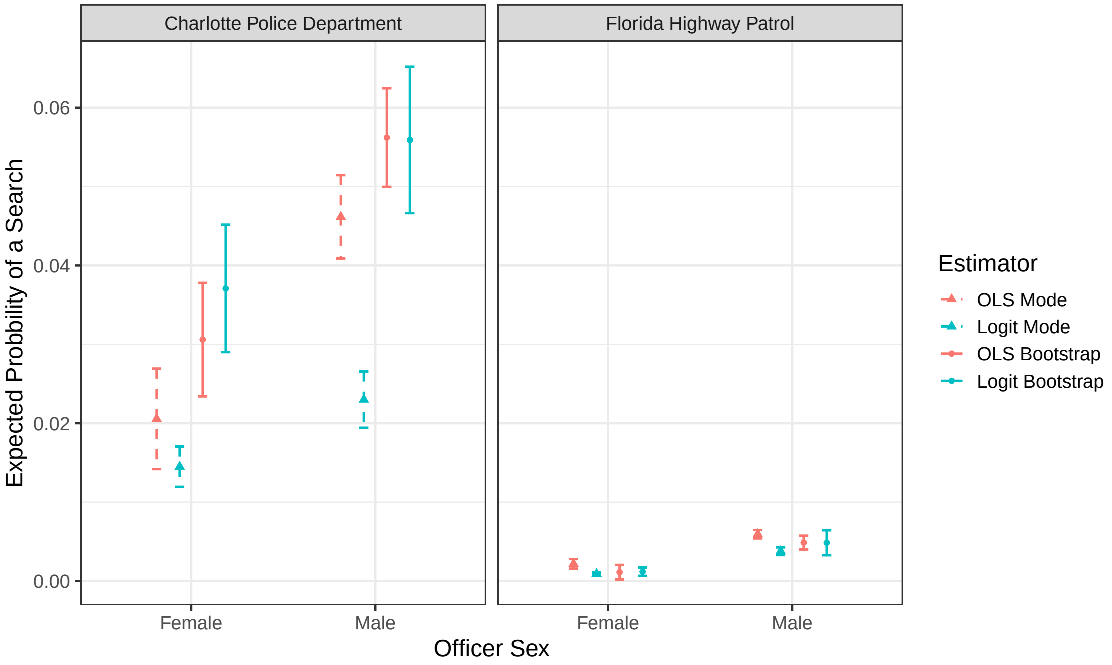


Figure 2: Expected Probability of Being Searched by a Male or Female Officer, by Agency and Estimator

and Song (2021) acknowledged that "[m]ale officers are expected to find contraband approximately 0.08 more times per 100 stops than female officers," and "we would still expect men to find contraband only 0.65 times per year on average given their average number of stops, whereas women would find contraband 0.18 times per year on average given their average number of stops made." (p.764) However, they concluded that all of the figures above can be rounded to zero so that the difference between officer sexes in finding contraband per 100 stops disappears and both sexes "are expected to find little or no contraband while conducting traffic stops.".

We instead insist that this negative, statistically significant relationship, albeit seemingly small, needs reinterpretation for three reasons. Firstly, Shoub, Stauffer, and Song (2021) did not round any coefficients to zero in their previous analyses, despite some of them are equally small in magnitude. For example, in Table 2 of the original paper (Table 2b in this paper), the coefficient for female officers in the FHP is -0.004. Even when similarly multiplied by 100, it would still be as small as -0.4. This means that, on average, the differences between officer sexes in conducting a search following a traffic stop is "merely" 0.4 times per 100 stops, which could as well be rounded to zero. Here, we are not calling for rounding these previous results to zero. On the contrary, we call for a *coherent* approach in respecting the statistical significance of *all* coefficients, regardless of their numerical magnitudes. The triviality of a result can only be, given sufficient statistical power (i.e. sample size), revealed by a large standard error relative to the coefficient estimate, which would render the result statistically insignificant.

Secondly, given the context of the result (hit rate per 100 stops), the difference of 0.08 is not small or trivial at all. This is because finding contraband upon a traffic stop is an event of extremely low probability. This can be demonstrated by calculating the expected hit rate per 100 stops similarly to Figure 2. Here we predict, in Figure 3, such hit rates for both officer sexes with OLS and logistic regressions (in Table 4c) and the bootstrapping technique. The plot indicates that the difference of 0.08 [6] is large relative to the predicted hit rates per 100 stops: The expected hit rate per 100 stops is 0.047 for female officers, and 0.139 for male officers, respectively, predicted using the logit bootstrapping method[7].

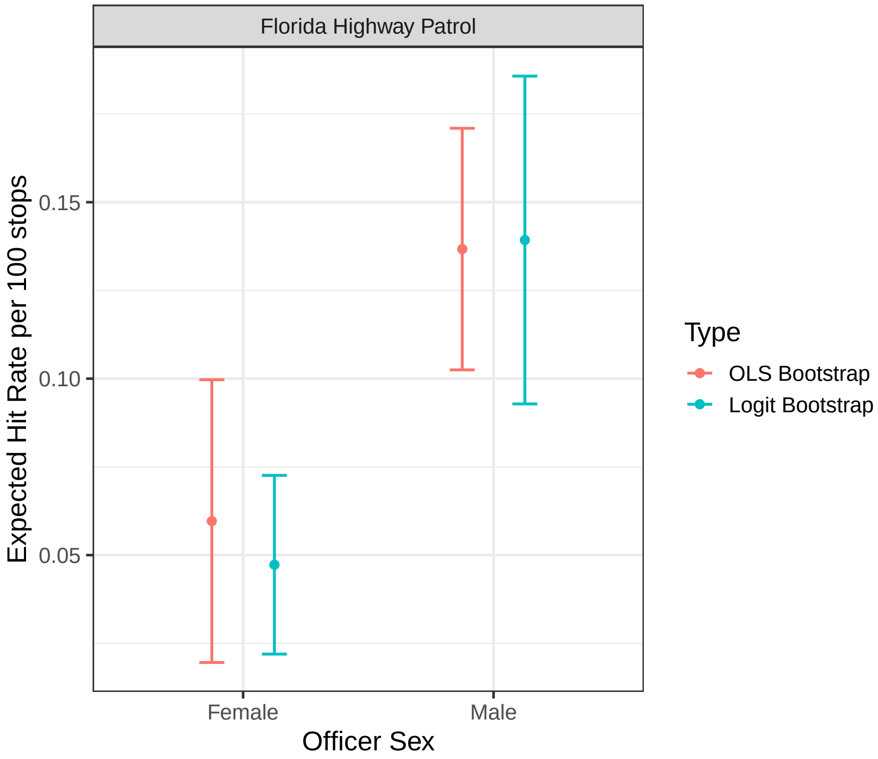


Figure 3: Expected Hit rate per 100 Stops by a Male or Female Officer, by Estimator

Thirdly, the relative odds of finding contraband based on the prediction results are comparable to those in Section 4: male officers are 295% as likely as female officers to find contraband upon a (or 100) traffic stop(s), shown by the logit bootstrapped results[8]. This magnitude is no less than the relative odds of a search being conducted reported by Shoub, Stauffer, and Song (2021) (225% and 272%).

Furthermore, readers of Shoub, Stauffer, and Song (2021) may pay attention to the following statement from the original paper:

> "Further, the statistically significant relationship disappears if officer effects are included in the model, which indicates that the statistical relationship is unstable." (p.764)

This presumably refers to the hierarchical linear model including officer random effects. The corresponding results were provided in Column 4 of Table C2 of the Supplementary Material. Here, we reproduce such results in the "Original" column of Table C2.

However, one caveat about using random effects models is the assumption that the random effects are *independent* of the independent variables (Vaisey and Miles 2017, p.47). The assumption is likely violated in this case because many independent variables are actually officer characteristics. Therefore, we follow the protocol proposed by Chatelain and Ralf (2021) to address this possible source of inconsistency

[6] or 0.10 from the logistic regression

[7] or 0.060 for female officers, and 0.137 for male officers, respectively, using the OLS bootstrapping method

[8] or 229% according to the OLS bootstrapped results.

	Original	Correlated	Hausman-Taylor
Female Officer	−0.053	−0.042	−1.039*
	(0.033)	(0.033)	(0.467)
Black Officer	−0.089**	−0.070*	−0.072
	(0.031)	(0.031)	(0.253)
Officer Age: 30-64	−0.023	−0.019	−0.028
	(0.016)	(0.020)	(0.019)
Officer Age: 65+	−0.124	−0.061	−0.084
	(0.214)	(0.230)	(0.228)
Experienced Officer	0.069***	0.076***	0.073***
	(0.015)	(0.019)	(0.018)
White Male	−0.055***	−0.055***	−0.055***
	(0.010)	(0.010)	(0.010)
Black Male	0.085***	0.084***	0.084***
	(0.010)	(0.010)	(0.010)
Black Female	−0.063***	−0.065***	−0.065***
	(0.012)	(0.012)	(0.012)
Latino Male	−0.009	−0.010	−0.010
	(0.011)	(0.011)	(0.011)
Latino Female	−0.068***	−0.068***	−0.068***
	(0.013)	(0.013)	(0.013)
Driver Age: 30-64	−0.119***	−0.118***	−0.118***
	(0.007)	(0.007)	(0.007)
Driver Age: 65+	−0.169***	−0.166***	−0.167***
	(0.012)	(0.012)	(0.012)
Investigatory Stop Purpose	0.243***	0.243***	0.243***
	(0.007)	(0.007)	(0.007)
Out of State	0.032***	0.031***	0.031***
	(0.008)	(0.008)	(0.008)
(Intercept)	0.071**	−0.332	0.195***
	(0.023)	(0.508)	(0.054)
Correction for Endogeneity of Time-Varying Variables	No	Yes	Yes
Correction for Endogeneity of Time-Invariant Variables	No	No	Yes
N	747.784	747.784	747.784

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table C2: Hierarchical Linear Models Including Officer Random Effects for Hit Rate per 100 Stops (FHP data)

in the model. We first add the cluster means of the time-varying variables (i.e. driver characteristics and stop information, in this case) to the mixed-effects model to correct for the correlation between the time-varying variables and the officer random effects. This is known as the Correlated Random Effects (CRE) technique or the Mundlak (1978) estimator. The results from the CRE are reported in the second column (Correlated) of Table C2. Note that at this stage, the correlation between *time-invariant* variables and officer random effects may still be present, and the results should be interpreted cautiously. However, time-varying variables for which the coefficients of the cluster means[9] are insignificant at the 0.05 level and can be considered exogenous to random effects.

We then perform the Hausman-Taylor (1981) estimator, which consistently estimates the effects of both time-varying and time-invariant variables that are endogenous to random effects. The Hausman-Taylor (1981) uses independent variables that are *exogenous* to random effects as internal instruments for their endogenous counterparts. In our case, those identified to be exogenous in the CRE results are used as internal instruments, as recommended by Chatelain and Ralf (2021).

The results for the Hausman-Taylor (HT) model are reported in Column 3 of Table C2. It can be seen that the HT model estimates a statistically significant impact of officer sex on the hit rate per 100 stops. However, the coefficient is much larger than our previous estimates. We cannot explain this discrepancy and welcome replications and discussions from future researchers.

6 Conclusion

Our reproduction of Shoub, Stauffer, and Song (2021) confirms the first and second claims of the paper:

1. Female police officers are less likely to search drivers than male officers.
2. Female police officers are more likely to find contraband when they do conduct a search.

Nevertheless, we predict the relative odds of being searched differently from the original paper. Whereas Shoub, Stauffer, and Song (2021) claimed that male officers are 225% and 272% as likely as female officers to conduct a search following a traffic stop in the two respective police departments, we estimate such odds to be 151% and 408%.

Furthermore, we think the third and fourth claims of Shoub, Stauffer, and Song (2021) need re-evaluation and further discussions:

3. Despite lower search rates, female officers confiscate the same net amount of contraband as male officers.
4. Female officers can minimise the number of negative interactions with citizens without compromising effectiveness.

For the FHP dataset, we argue that female officers specifically appear to have confiscated significantly *less* contraband than male officers between 2010 and 2015. This could imply a *trade-off* between negative police-citizens and effectiveness in traffic stops for contraband. While we are happy to see a growing literature on the impact of police officer sex on the use of coercive actions or force (Ba et al. 2021; Rabe-Hemp 2008; Schuck and Rabe-Hemp 2005 for positive results; Sun 2007 for a null result) and domestic violence (Poteyeva and Sun 2009 for a weak effect), we call for more scholarly attention on the difference in *effectiveness* as it is currently under-researched.

Data Availability Statement The replication files for this paper are available through this link.

Competing Interests The authors declare none.

[9] Means for all observations from the same officer

References

Ba, Bocar A. et al. (Feb. 2021). “The role of officer race and gender in police-civilian interactions in Chicago”. In: *Science* 371.6530, pp. 696–702. DOI: 10.1126/science.abd8694.

Breusch, T. S. and A. R. Pagan (1979). “A Simple Test for Heteroscedasticity and Random Coefficient Variation”. In: *Econometrica* 47.5, pp. 1287–1294. ISSN: 0012-9682. DOI: 10.2307/1911963. URL: https://www.jstor.org/stable/1911963 (visited on 09/15/2023).

Chatelain, Jean-Bernard and Kirsten Ralf (Apr. 2021). “Inference on time-invariant variables using panel data: A pretest estimator”. In: *Economic Modelling* 97, pp. 157–166. DOI: 10.1016/j.econmod.2021.01.014.

Fernandez-Val, Ivan and Martin Weidner (Dec. 2018). *Individual and Time Effects in Nonlinear Panel Models with Large N, T*. Tech. rep. arXiv:1311.7065 [econ, stat] type: article. arXiv. DOI: 10.48550/arXiv.1311.7065. URL: http://arxiv.org/abs/1311.7065 (visited on 09/15/2023).

Hausman, Jerry A. and William E. Taylor (May 1981). “Panel data and unobservable individual effects”. In: *Journal of Econometrics* 16.1, p. 155. ISSN: 0304-4076. DOI: 10.1016/0304-4076(81)90085-3. URL: https://www.sciencedirect.com/science/article/pii/0304407681900853 (visited on 09/18/2023).

Mundlak, Yair (1978). “On the Pooling of Time Series and Cross Section Data”. In: *Econometrica* 46.1, pp. 69–85. ISSN: 0012-9682. DOI: 10.2307/1913646. URL: https://www.jstor.org/stable/1913646 (visited on 09/18/2023).

Neyman, J. and Elizabeth L. Scott (1948). “Consistent Estimates Based on Partially Consistent Observations”. In: *Econometrica* 16.1, pp. 1–32. ISSN: 0012-9682. DOI: 10.2307/1914288. URL: https://www.jstor.org/stable/1914288 (visited on 09/15/2023).

Poteyeva, Margarita and Ivan Y. Sun (Sept. 2009). “Gender differences in police officers’ attitudes: Assessing current empirical evidence”. In: *Journal of Criminal Justice* 37.5, pp. 512–522. ISSN: 0047-2352. DOI: 10.1016/j.jcrimjus.2009.07.011. URL: https://www.sciencedirect.com/science/article/pii/S0047235209000956 (visited on 09/18/2023).

Rabe-Hemp, Cara (Sept. 2008). “Female officers and the ethic of care: Does officer gender impact police behaviors?” In: *Journal of Criminal Justice* 36.5, pp. 426–434. ISSN: 0047-2352. DOI: 10.1016/j.jcrimjus.2008.07.001. URL: https://www.sciencedirect.com/science/article/pii/S0047235208000895 (visited on 09/18/2023).

Schuck, Amie M. and Cara Rabe-Hemp (July 2005). “Women Police”. In: *Women & Criminal Justice* 16.4, pp. 91–117. DOI: 10.1300/j012v16n04_05.

Shoub, Kelsey, Katelyn E. Stauffer, and Miyeon Song (May 2021). “Do Female Officers Police Differently? Evidence from Traffic Stops”. In: *American Journal of Political Science* 65.3, pp. 755–769. DOI: 10.1111/ajps.12618.

Stock, James and Mark Watson (2014). *Econometrics, Update PDF Ebook, Global Edtion*. Harlow, United Kingdom, UNITED KINGDOM: Pearson Education, Limited. ISBN: 9781292071367. URL: http://ebookcentral.proquest.com/lib/londonschoolecons/detail.action?docID=5174962 (visited on 07/19/2023).

Sun, Ivan Y. (Dec. 2007). “Policing domestic violence: Does officer gender matter?” In: *Journal of Criminal Justice* 35.6, pp. 581–595. DOI: 10.1016/j.jcrimjus.2007.09.004.

Vaisey, Stephen and Andrew Miles (Jan. 2017). “What You Can—and Can’ t—Do With Three-Wave Panel Data”. en. In: *Sociological Methods & Research* 46.1, pp. 44–67. ISSN: 0049-1241. DOI: 10.1177/0049124114547769. URL: https://doi.org/10.1177/0049124114547769 (visited on 09/17/2023).

Webb, Matthew D. (Nov. 2014). *Reworking Wild Bootstrap Based Inference For Clustered Errors*. Working Paper 1315. Economics Department, Queen’s University. URL: https://ideas.repec.org/p/qed/wpaper/1315.html.