nscr

Citizen profiling during weapon controls in Amsterdam: An observational analysis of practices and potential biases

Marie Rosenkrantz Lindegaard
Peter Ejbye-Ernst
Mara van Dalen
Hans Myhre Sunde
Carlijn van Baak
Melissa Sexton
Virginia Pallante
Fabienne Thijs
Lea Echelmeyer
Steve van de Weijer
Laura Pighini
Gabriele Chlevickaite
Jo Thomas
Lasse Suonperä Liebst

Amsterdam, December 2022

Nederlands Studiecentrum
Criminaliteit en Rechtshandhaving

Citizen profiling during weapon controls in Amsterdam: An observational analysis of practices and potential biases

DECEMBER 2022

Marie Rosenkrantz Lindegaard [1, 2]
Peter Ejbye-Ernst [1]
Mara van Dalen [1]
Hans Myhre Sunde [1, 2]
Carlijn van Baak [1, 2]
Melissa Sexton [1]
Virginia Pallante [1]
Fabienne Thijs [1]
Lea Echelmeyer [1]
Steve van de Weijer [1]
Laura Pighini [1]
Gabriele Chlevickaite [1]
Jo Thomas [1]
Lasse Suonperä Liebst [1, 3]

[1] NSCR: Netherlands Institute for the Study of Crime and Law Enforcement
[2] Department of Sociology, University of Amsterdam
[3] Department of Sociology, University of Copenhagen

Contact
Prof. Dr. Marie Rosenkrantz Lindegaard, mrlindegaard@nscr.nl

Acknowledgements
We would also like to thank the coordinators of the project at the City of Amsterdam and at the police. We also thank Wim Bernasco for his contributions in launching the study.

Author contributions
Conceptualization: MRL, LSL, PEE. Data Curation: MD. Formal Analysis: LSL, PEE, VP, MD. Funding Acquisition: MRL. Measurement instrument: MRL, CB, MR, PEE, MS. Field observations: MRL, PEE, HMS, VP, MS, CB, FT, LE, SW, LP, GC, JT. Methodology: MRL, LSL, PEE. Project Administration: MRL. Resources: MRL. Supervision: MRL. Validation: LSL. Visualization: PEE. Writing – Original Draft Preparation: MRL, PEE, LSL, VP, MR. Writing – Review & Editing: MRL, PEE, HMS, VP, MS, CB, FT, LE, SW, LP, GC, JT, LSL.

Key findings

- We found that younger persons, especially those in their mid to late 20s, were disproportionally likely to be selected for weapon control compared to the persons in the same location that did not get selected.
- We did not find that non-white persons and men were disproportionally selected for weapon control compared to the persons present in the same location that did not get selected.

Summary

For the legitimacy of police work, unbiased search actions are essential. Existing self-reported evidence about search practices of the police in the Netherlands and beyond, indicates that citizens experience officers as biased in their target selection during search actions. In the current project, we evaluated the existence of possible gender, ethnicity and age biases during a pilot test of weapon controls in Amsterdam using an observational objective measure of citizen selections. This pilot was terminated after unauthorized practices by the police, and thus the current study was likewise put to an end during data collection. However, we did manage to collect a fair amount of observational data to draw some preliminary conclusions. Our statistical analysis indicated that younger persons at around 20 years were disproportionally likely to be selected for weapon control compared to the persons in the same location that did not get selected. No such biases were noticeable for Non-White people or men. As a caveat, it should be mentioned that it is unclear whether the lack of ethnic and gender profiling is due to the actual non-existence of this bias or rather is due to the current dataset being too small and noisy to detect this possible bias.

INTRODUCTION

For the legitimacy of police work, unbiased search actions are essential (Bradford & Loader, 2016). However, research from a variety of national contexts, including the Netherlands (Hesseling & Wilde, 2022), indicates that citizens experience officers as biased in their target selection during search actions, particularly due to a disproportional focus on young men of color (e.g., Dennison & Finkeldey, 2021). A problem with these studies—and with studies on discrimination in general—is that they often depend on the subjective experiences of the people involved, which are vulnerable to interpretation disagreement. This raises the need for more objective measures of offender profiling, but unfortunately, it is challenging to find an accurate method to evaluate it (Farrell & McDevitt, 2010). One promising but underutilized approach to do so is to observe the searches as they unfold on-site. A rare attempt to do so was conducted in Paris, and the behavioral evidence of that study confirmed the subjective experiences: During search actions, police officers disproportionately selected young men of color, despite instructions to use neutral profiling strategies (Jobard & Lévy, 2011). In the current report, we followed this promising behavioral approach by observing preventive weapon search actions in Amsterdam.

Due to an increase in weapon-related crimes in the city of Amsterdam the city mayor decided to run a pilot on weapon controls. Weapon controls had for years been abandoned by the City Council because the impact was interpreted as disproportional, including the potential problem of ethnic profiling biases in the selection procedures. As such, there is a tension between the potential negative impact controls have on people being searched and the potential benefits of finding weapons and preventing weapon-related crimes. As part of the pilot, the mayor addressed the negative impact by asking citizens to register as third-party 'observers' of the police during controls. In total, 37 citizens joined the police during controls and described their impression of the procedures in a report (Hesseling & Wilde, 2022). Only six percent of the observers experienced that the police used ethnic profiling in their selection procedures, and only one official complaint was submitted. On the other hand, a survey asking citizens about their general opinion about the controls indicated that about 25 percent of the respondents expected ethnic profiling to occur. These subjective evaluations of third-party observers and citizens were attention paid to the question of ethnic profiling in the report on the first pilot.

As the question of ethnic profiling during the first pilot was only based on self-reported measures, it thus remained unclear whether the police had selected disproportionally more people of color for control. Furthermore, the evaluation of the first pilot showed that more men than women were selected for control, which could indicate a bias toward selecting men. Additionally, it also showed that more people younger than 30 years old were selected than above 30 years old, which could indicate a bias toward selecting youth. However, rather than an indication of a male-biased or youth-biased selection practice, this result may simply be attributed to the presence of more men versus women, and more young than older people, in the controlled public places. From the evaluation of the first pilot study, it is also unclear what the ethnicity were of the people chosen for selection, as the police do not register this category.

After evaluating the first pilot, the City of Amsterdam concluded that the pilot could be continued, and the City of Amsterdam decided that instead of installing citizens as observers, independent researchers should take on the observer role during a second pilot—this task was carried out by the current project. Specifically, our role became to investigate whether the police selection procedures were taking place in unbiased ways—that is, everyone present in the location where the controls were taking place would have an equal chance of being selected for control.

The current study focuses on the second pilot of weapon controls in Amsterdam, in which 25 controls were planned between October 2022 and January 2023. However, after only five observed controls, the pilot was terminated due to unauthorized practices of the police, and, as such, the current project only includes data from these five conducted controls. Before carrying out the controls, the police had identified five urban areas that—according to their reported crime statistics—were particularly prone to weapon-related crimes. Those areas were: North, South, South-East, North New-West, and South New-West. Within each area, the police had identified a certain number of 'hotspots' in which they expected to find more weapons. It was communicated to the public that the police sometimes would work with hand scanners (i.e., to detect weapons), a detection port, and a randomization selection pole (i.e., a technical device where the citizens have to press a button to pass the pole and a proportion of persons are then automatically selected at the pure chance).

More specifically, the current study aims to statistically examine whether the police were biased towards selecting on ethnicity, gender, or age categories during weapon controls. To examine this question, we first had to understand how the selection procedure of the police worked in practice. We were informed beforehand that they used two ways of selecting people: Either single citizens would be selected by the police agents or by the randomization selection pole—we refer to these as "ordinary" and "pole" selection, respectively. When observing their working procedures, however, it turned out that a third selection method was used too. We refer to this method as "place selections" as it involved closing off single places, and not allowing people to leave the place until they had been searched. In our analysis of the potential biases in the selection procedures, we take these three methods into consideration.

METHOD

We decided to observe the police controls on-site because this allowed highly ecologically valid and qualitatively fine-grained data on how the controls actually took place. On the other hand, a potential drawback of on-site observation is the problem of 'reactivity,' meaning that police officers may alter their practice due to their awareness of being observed. In dialogue with the police and the City of Amsterdam and after having sought out research ethical advice, we thus decided to conduct the on-site observations covertly. Note that the code of ethics of the American Sociological Association (1999) and similar associations stress that scholars may conduct naturalistic observations in public places, as in the current case, without obtaining consent. Because of the covert observations, the police officers conducting the controls were not aware of whether

or not they were directly observed. That is, in our experience, we could typically blend in without any indications of being disclosed.

However, it should be mentioned that the police on the ground may have had some awareness because it had been announced—both internally in the police organization and in the public media—that a research group were to evaluate the policing practices. Adding to this, we were also spotted twice while observing, and we were further told that the officers had suspected being observed on a few other occasions (when, in fact, we had not been present). So, although our design did not exclude the risk of reactivity, we probably minimized this potential bias by creating uncertainty about whether the ongoing control was observed or not. We note that we would have preferred that the officers were fully blinded from the project and were observed with video security data without any possibility of reactivity. However, for practical reasons, this was not feasible (e.g., the police found that video observations were too intrusive compared to on-site observations).

In total, we carried out on-site observations during six conducted controls: three in South, one in North, one in South-East, and one in New-West. We worked in a team of 12 researchers who were instructed in using the measurement tool and took turns observing the controls. Typically four researchers were involved in each police shift and made observations for around four hours. On the day of each control, we were informed by the responsible police coordinator where it would take place. We did not disclose to this police coordinator whether we would actually be present, and we had instructed the police not to inform the officers on the shift about our potential presence. After arriving at the site of the control, the researchers observed the everyday behavior in the particular space in order to figure out how they could blend in during their observations. Since people rarely engage in prolonged activities in public space, they were instructed to shift their activities regularly not to get noticed. Typical blending-in activities were eating, drinking coffee, speaking on the phone, waiting on public transport, waiting with a suitcase, and chatting with a co-observer. The observers were instructed to terminate the observations the moment the police appeared to notice them.

Sampling procedure

During controls, the police should shift between checking pedestrians, bikes, and cars, but for feasibility reasons and to make our results comparable to prior profiling research (Jobard & Lévy, 2011), we only made records of pedestrians. Further, the observers took records on a smartphone using an online survey design tool. The selection of persons for coding depended on the control methods used by the police. First, in the situations where single citizens were selected by a police officer or by the randomization selection pole, we randomly selected approximately every third person crossing an imaginary line as they moved toward the control area. Alternatively, in settings with a low level of crowding, we sampled everyone present if practically feasible.

Second, this sampling procedure changed in the course of the study, as we realized that the police sometimes selected all persons present in a certain area for visitation—rather than

single persons, as in the case of the ordinary and randomization pole controls. For these place-based controls, we instead applied a case-control sampling approach, where we initially recorded as many persons as possible from within the area where the police searched everybody (i.e., cases). We then sampled a comparison category of persons from the non-searched areas immediately around the searched area (i.e., controls). This sampling procedure followed the recommendation that comparable control persons could be sampled from the same time and space as cases (Grimes & Schulz, 2005).

In practice, however, we could not sample a sufficiently large subset of controls—i.e., a ratio of 4 to 1 controls to cases is often recommended. So to approximate this control-to-case ratio, we decided to sample controls from the non-searched area one week after the police search had taken place. It must be admitted that these (i.e., the majority of) controls may be less comparable than if they had been sampled during the same time as the police searches (e.g., the weather was substantially worse and colder during the following week). However, under the assumption that these control persons are not different in terms of age, gender, and ethnicity than persons drawn one week earlier, we decided to use these control cases in the analysis. We also decided to do so because it allowed us to construct a more statistically well-powered dataset even after the city mayor had decided to terminate the police search pilot. As such, sensitivity analysis of the area-based dataset of, in total, 393 observations could identify an even small effect size ($f^2 = 0.02$) with a power of 80%.

By comparison, both datasets of the ordinary and especially the randomization observations had substantially lower power. Specifically, the dataset of 185 ordinary controls could identify a small-to-medium effect size ($f^2 = 0.04$), and the pole randomization dataset of only 30 cases could identify a medium-to-large effect ($f^2 = 0.28$). When analyzing the results, it should be kept in mind that the less well-powered datasets increase the risk of false negatives and noisy estimates in general.

Coding and measures

The systematic coding of, e.g., gender and ethnicity required the formulation of well-defined definitions of these social categories. If possible, we relied on prior research that had field-tested similar measures. This was the case with respect to the ethnicity definition, which was similar to the one used in the aforementioned Paris-based study (Jobard & Lévy, 2011). Further, prior studies have also shown that age, gender, and relationship ties may be accurately coded in real-life public settings (e.g., Liebst et al., 2022). However, to ensure the generalizability of these measures to the current study context, we conducted an intercoder reliability test of the included measures. This was done by comparing the agreement between two independent records of the same persons randomly selected in public places. The scores of each interrater reliability test are reported below. Note that we acknowledge that our interrater reliability testing does not guarantee that our measures are valid from the perspective of the rated persons. For example, we may be systematically wrong in our assessments (e.g., the raters may underestimate the actual

age of older persons). Further, our assessments may diverge from the self-perceived identity of the persons.

All measures were based on all visual cues available in the situation (e.g., clothing, hair or skin color): *Ethnicity* was measured with five categories: White, Arab, Black, Indo-Pakistani, and Asian. This measure had a 'fair' intercoder reliability (Krippendorff's alpha = 0.68) when assessed using standard thresholds for interrater reliability (Fleiss, 1981). However, when re-coded into a White versus Non-white dichotomy, the reliability score was 'good' (0.80). *Gender* was measured as Male versus Female and had a 'close to perfect' (1.0) interrater score. *Age* was measured on a continuous scale and reached a 'near perfect' (.94) interrater score. *Family affiliation* captured whether the persons were in the company of young children (0-12 years old) and hence were not supposed to get checked. This measure had a 'good' (0.80) interrater score.

RESULTS

This section presents our quantitative results. Table 1 shows how the data were collected during the three different selection procedures: pole, ordinary, and place selection. We sampled 30 individuals when the police used the randomization pole, and out of these, three persons were searched for weapons. We sampled 185 individuals in situations where the police used an ordinary selection procedure, of which 58 were searched. We sampled 393 individuals for analyzing place-based selection, including 81 who were searched[1].

During the use of the randomization pole, about one in three observed individuals was non-white, just under half were men, and the average age was 35 years. The individuals observed during ordinary selections were just above 40% non-white, about half were men, and had an average age of 30 years. Of the individuals observed during place selection, about one-fourth were non-white, just above half were men, and the estimated age was around 35 years. Therefore, the gender distribution and the average age varied only slightly across the three selection methods. The largest difference across the three samples was seen in the proportion of non-white people observed. This difference might be attributable to the variation in the ethnic composition of the neighborhoods where the different types of searches were observed. The different methods of selection were observed in different areas of the city: the selection pole was only observed in Amsterdam South. The ordinary selection was observed in Amsterdam North and South-East. The place-based selection procedure was observed in Amsterdam New-West and South. It follows from this that non-white persons could not be equally likely to be randomly selected into our samples across the methods if the areas differ in ethnic composition.

[1] For the analysis below we excluded families with children young than 12 and individuals who were estimated to be older than 65. Both of these groups were excluded because the police do not search either for weapons.

Table 1. Descriptive statistics of the people sampled during different selection procedures

	Pole selection	Ordinary selection	Place selection
The total number of individuals sampled	30	185	393
Number of individuals sampled who were searched for weapons	3	58	81
Non-white	30 %	42.7 %	26.7 %
Men	43.3 %	49.7 %	54.2 %
Average age	35.0	30.6	35.4
Part of Amsterdam	South	North and South-east	New West and South

Since we unfortunately only managed to observe a small number of individuals in weapons searches where the police used the selection pole (n = 30), we did not perform any further analysis on this particular selection method. For the remaining two selection methods, we investigated how the demographic characteristics of ethnicity, gender, and age were associated with the probability of being selected for a weapon search.

Ordinary selection

Figure 1 shows the selected age distribution of the individuals and those who were not selected when the police used the ordinary selection procedure. This figure shows that people in their late 20s were disproportionally more likely to be selected compared to other age groups. The figure also shows that individuals over 35 make up a larger proportion of non-searched individuals than those searched. A regression model indicated that this association between age and selection probability was statistically clear (OR = 0.96, p = 0.033), although only marginally so. With this caveat in mind, the odds ratio (OR) suggests that for each year younger, a person is 1/0.96 = 1.04 times more likely to be selected. The association is visualized in Figure 2, with lower age predicting a substantial increase in the likelihood of being selected—persons at around 20 years had around 40 probability of being selected while the probability drops to approximately 20% for persons around 50 years. It should be stressed that these regression results may be interpreted as statistically unclear if the p-value threshold was Bonferroni corrected to a 1% level (versus a traditional threshold of 5%) due to a multiple comparison problem (i.e., we test the same hypothesis in several tests). Note: If not otherwise mentioned, all regression analyses reported were specified as bivariate logit models with robust standard errors.

Figure 1. Age distribution of people who do and not get selection via ordinary selection

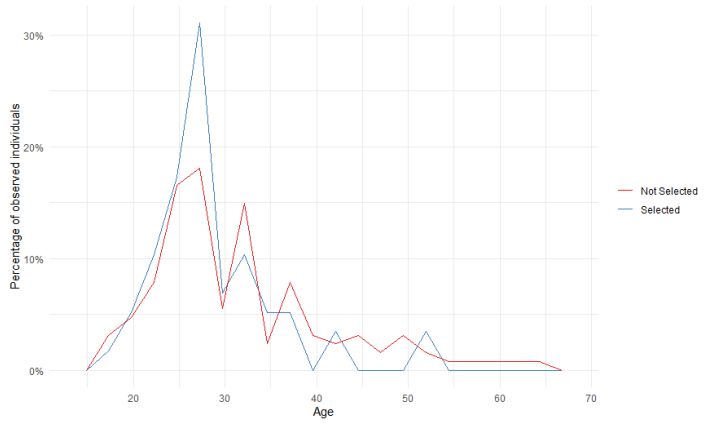
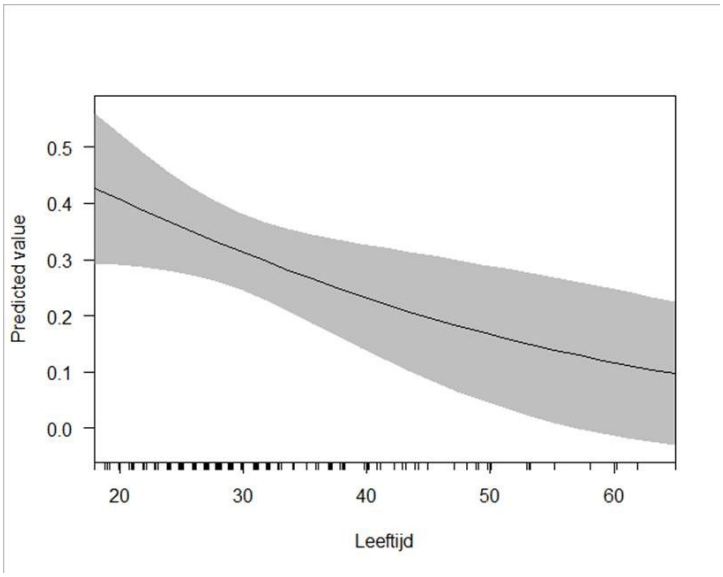


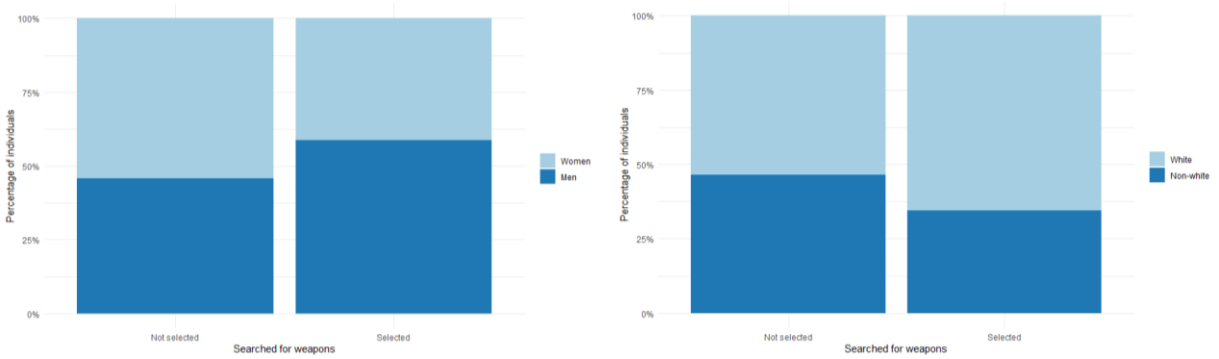
Figure 2. The estimated association between age and the predicted probability of being selected during ordinary searches, with the grey area depicting the 95% confidence



interval

Figure 3 shows the distribution of men and women who got selected and white and non-white individuals who got selected for weapon searches. The left panel of Figure 3 shows the percentage of people who got selected in situations during ordinary searches. This graph shows that men appeared to comprise a larger proportion of those selected compared to those who were not selected. However, this pattern was not statistically clear (OR = 0.59, p = 0.110). The right panel of Figure 3 shows that the proportion of white people selected during the ordinary selection procedure seemed disproportionally large. However, this pattern was also not statistically clear (OR = 0.61, p = 0.130).

Figure 3. The left and right panel depicts, respectively, the gender and ethnicity of people who do and do not get selected through ordinary selection



Next, we ran a multivariate analysis to see if the above findings remained the same when we included all three demographic predictors (i.e., age, gender, and ethnicity) in one model simultaneously. Here, only age showed some relationship with the likelihood of being selected for weapon visitation (OR = 0.96, p = 0.020), albeit the association was again only statistically clear at a 5% level, while not at a 1% level.

Place selection

Figure 4 shows the age distribution of the individuals searched for weapons and those not selected during the place-based selection procedure. This figure shows that a disproportionally large proportion of persons around 20 years was selected. Further, a higher proportion of the individuals who were not searched was over 35 years compared to those who were searched. This association between age and selection likelihood was very statistically clear (OR = 0.91, p < 0.001). This odds ratio suggests that for each year younger, a person is 1/0.91 = 1.10 times more likely to be selected. This pattern was similar, although more pronounced, when compared to the age pattern observed during the ordinary selection procedure.

Figure 4. Age distribution of people who do and do not get selected through place selection

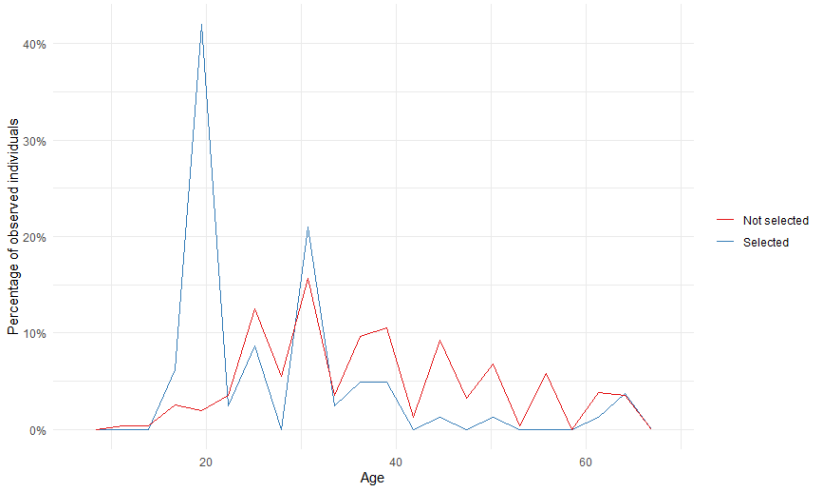
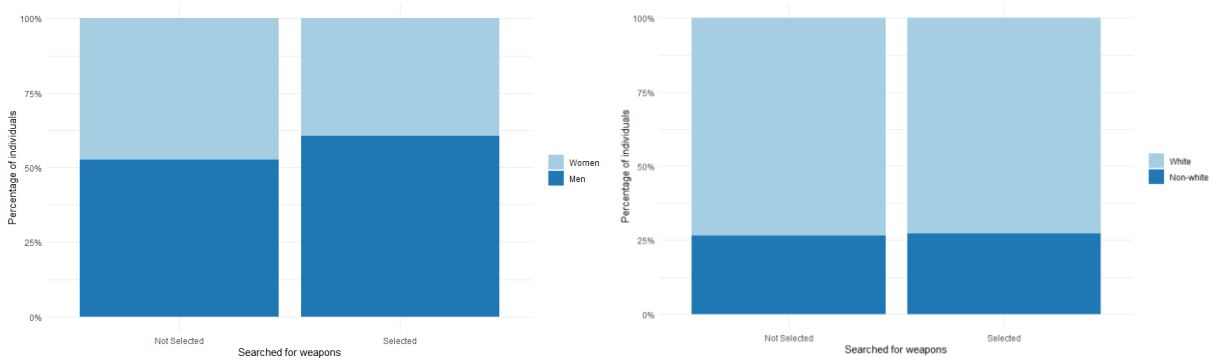


Figure 5 shows the distribution of gender and ethnicity of the individuals who were searched and those not searched during place selection. The left panel of Figure 5 shows the proportion of men and women among the individuals who were searched and those who were not. It appears that more men than women were selected for weapon searches, but this difference was not statistically clear (OR = 0.72, p = 0.204). The right panel of Figure 5 shows the proportion of white and non-white individuals searched and not searched. There was no visual or statistically clear (OR = 1.03, p = 0.920) difference between the selection proportion of these two groups.

Figure 5. The left and right panel depicts, respectively, the gender and ethnicity of people who do and do not get selected through place selection



To check the robustness of the above findings, we ran a multivariate logistic regression model containing age, gender, and ethnicity. Again, age was found to have a very statistically clear relationship with the selection likelihood (OR = 0.91, $p < 0.001$). Furthermore, this robustness analysis indicated that men had a higher probability of being selected compared to women (OR = 0.55, p = 0.036), although this estimate was only marginally statistically clear at 5% p-value

threshold or unclear if assessed against a Bonferroni corrected threshold. Note that this latter result was different from the bivariate analysis and what appeared from the above data visualization. Finally, we note that the above case-control results remained the same if the unauthorized controls were excluded from the sample.

CONCLUSION

In the current analysis, we have examined whether and why there is evidence suggesting a biased section based on ethnicity, age, or gender during weapon searches. Data did not provide robust evidence that the gender or ethnicity of citizens were associated with a disproportional likelihood of being selected for weapon searching. By comparison, data did offer more evidence that younger persons, especially those in their mid to late 20s, were disproportionally likely to be selected. This result is persuasive because it was replicated in two separate samples measuring ordinary and place-based selection procedures. As such, this replication mitigates the inconclusive result in the analysis of the ordinary selection data: A borderline statistically clear (i.e., ordinary selection data) and a highly statistically clear (i.e., place selection data) estimate pointing in the same direction should, in general, be interpreted as converging rather than conflicting evidence. The difference in statistical clarity between the two datasets might, to some extent, be due to the relatively smaller number of people observed in the situations with ordinary selection compared to place selection (i.e., fewer observations means lower statistical power to identify actual effects).

However, it should be mentioned that data does not exclude the risk that ethnic profiling could take place at the level of the neighborhoods when the police planned where to conduct the weapon searches. This could be the case if the police target neighborhoods with a high proportion of non-white people on the street. This would, in turn, inflate the count of non-white persons selected, even if the non-white and white people have the same probability of being selected. Also, it should be stressed that although the analysis revealed certain trends in the data, the generalizability of these patterns to other contexts is not certain. We only managed to observe five weapon controls in total before they were discontinued. Therefore, each of the three selection methods are based on only a few days of observation. Whether the findings reveal something about the specific days we observed or the general practices of the police is thus unclear. Finally, it should be stressed that the lack of evidence for ethnic and gender profiling could be due the current data being too small and noisy to detect such possible bias.

LITERATURE

American Sociological Association. (1999). *American Sociological Association code of ethics*. https://www.asanet.org/sites/default/files/savvy/images/asa/docs/pdf/CodeofEthics.pdf

Bradford, B., & Loader, I. (2016). Police, Crime and Order: The Case of Stop and Search. In B. Bradford, B. Jauregui, I. Loader, & J. Steinberg (Eds.), *The SAGE Handbook of Global Policing* (pp. 241–260).

Dennison, C. R., & Finkeldey, J. G. (2021). Self-reported experiences and consequences of unfair treatment by police. *Criminology*, *59*(2), 254–290. https://doi.org/10.1111/1745-9125.12269

Farrell, A., & McDevitt, J. (2010). Identifying and Measuring Racial Profiling by the Police. *Sociology Compass*, *4*(1), 77–88.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions*. Wiley.

Grimes, D. A., & Schulz, K. F. (2005). Compared to what? Finding controls for case-control studies. *The Lancet*, *365*(9468), 1429–1433.

Hesseling, N., & Wilde, S. de. (2022). Directie Openbare Orde en Veiligheid. https://openresearch.amsterdam/nl/page/79166/onderzoek-pilot-wapencontroles

Holmberg, L. (2003). *Policing stereotypes: A qualitative study of police work in Denmark*. Galda + Wilch.

Jobard, F., & Lévy, R. (2011). Racial profiling: The Parisian police experience. *Canadian Journal of Criminology and Criminal Justice*, *53*(1), 87–93.

Liebst, L. S., Baggesen, L., Dausel, K. L., & Lindegaard, M. R. (2022). *Human observers have a high accuracy in age estimating strangers in public settings*. PsyArXiv. https://doi.org/10.31234/osf.io/zbpkw