Diversity training and employee behavior: Evidence from the police[*]

Steven Mello Matthew Ross Stephen Ross Hunter Johnson

November 16, 2023

Abstract

We study the effects of cultural diversity training on employee behavior in the context of policing. Relying on administrative data covering all traffic stops made by early-career highway patrol officers in Texas and leveraging variation in the timing of a mandated diversity training, we find that troopers respond to training by adjusting their racial composition of stops. The probability that a stopped motorist is white increases by about 1.5 percentage points in the six months after training, and troopers achieve this by stopping additional white motorists. Search and arrest rates of stopped white motorists fall after training, suggesting that officers are stopping less "guilty" whites. Importantly, behavioral changes after training only persist for about one year.

JEL Codes: M53, J15, K42

[*] Mello: Dartmouth College and NBER. Ross, M: Northeastern University. Ross, S: University of Connecticut and NBER. Johnson: Claremont Graduate University. We are grateful to Bocar Ba, Leah Boustan, Jennifer Doleac, Felipe Goncalves, Erzo Luttmer, Emily Owens, Emma Rackstraw, Roman Rivera, Kevin Schnepel, CarlyWill Sloan, and Cody Tuttle, as well as seminar participants at the Online Economics of Crime Seminar, the Western Economics Association Annual Meeting, and Montana State University for helpful comments. Jackyln Pi and Zeyu Wang provided excellent research assistance. M. Ross thanks the Russell Sage Foundation (G-2011-30016) for financial support. Any errors are our own.

1 Introduction

Diversity training, or training aimed at improving employee understanding of the diverse backgrounds of individuals, has become a pervasive feature of employment in the United States. Over two thirds of organizations in the U.S. offer some form of diversity training[1], and diversity training has been the focus of significant public attention during several high-profile incidents over the past decade. In 2018, Starbucks announced the closure of all stores in the United States to allow employees to complete training after two Black men were arrested at a Philadelphia Starbucks for using the restroom but not placing an order (Calfas, 2018). Delta Air Lines provided diversity training for all 23,000 flight attendants following a 2016 incident in which a Black physician's credentials were questioned while providing medical attention to another passenger during a flight (Shen, 2017).

Despite the prevalence of diversity training, there is no consensus on its effectiveness. While some studies have shown positive effects on employee attitudes towards other groups (e.g., Chang et al. 2019; Bezrukova et al. 2012), others have documented a "backlash" effect, where attitudes worsen following training (Dobbin and Kalev, 2016). Moreover, evidence of effects on behavior, rather than self-reported attitudes, is scarce (Chang et al., 2019). Important obstacles in this literature include a lack of credible variation in exposure to training and an inability to reliably measure behavioral responses.

In this paper, we study the effects of a mandated cultural diversity training program on the behavior of highway patrol officers in Texas. This setting has several important advantages. First, we can directly measure on-the-job behavior using administrative records on troopers' enforcement activities. Second, policing is a particularly interesting setting given widespread concerns about racial discrimination by officers (e.g., Newport 2016; Pierson et al. 2020) and the need for policing reform more broadly (Crabtree, 2020). A potential downside of our setting is that diversity training for police officers focuses on behavior towards the public, whereas diversity training in much of corporate America targets employee behavior towards coworkers. However, we view the training we study as similar to that provided to client-facing employees or service providers in the public sector, such as healthcare workers or educators (e.g., Tumen et al. 2022).

After completing the police academy, Texas Highway Patrol officers are required to participate in a variety of in-service trainings in order to achieve proficiency levels mandated by the state's occupational licensing agency for law enforcement officers. One such training is in cultural diversity, which is an eight-hour training taken in two, four-hour blocks and emphasizing role-playing of interactions with individuals of diverse backgrounds. Training is delivered by a senior police officer, rather than an academic or social-service provider.

We study the impacts of this cultural diversity training on the behavior of early-career troopers using an event study approach, leveraging the staggered timing of training across

[1] See https://www.soocial.com/diversity-training-statistics/.

officers. To address the various identification concerns associated with two-way fixed effects approaches raised in the recent econometrics literature (e.g., Roth et al. 2022), we rely on the two-step imputation estimator proposed by Borusyak et al. (2022) and Gardner (2021). This approach accommodates both our data structure (our dataset is at the traffic stop level, rather than a standard unit × time panel) and our need for a more complex fixed effects structure than TWFE. Throughout our analysis, we condition on detailed location and shift fixed effects, in addition to officer and time effects. Our approach relies on the assumption that, following diversity training, the enforcement behavior of officers would have trended similarly to those who patrol the same areas and who take training in the future, had training not occurred at that date.

Importantly, we highlight that while all troopers are eventually required to receive diversity training, the variation we exploit in practice is "natural variation" driven by officer choices of precisely when to take up training. The timing and location of course offerings are at the discretion of a decentralized system of service providers and we show that that trooper characteristics do not systematically differ by treatment timing, that trooper patrol assignments do not change systematically around the timing of training, that the timing of diversity training rarely coincides with the timing of other mandated trainings, and that officer behavior does not change systematically in the weeks leading up to diversity training, all of which increase confidence in the validity of our empirical approach.

We find that troopers respond to diversity training by adjusting the racial and ethnic composition of their traffic stops. In the six months immediately following diversity training, the probability that a cited motorist is white increases by 1.3 percentage points ($se = 0.005$), or about four percent relative to the pre-treatment mean. This effect is concentrated entirely among white officers; there is no change in the racial composition of the stops of Black or Hispanic officers following training. Event study estimates revert to zero after ten months and remain statistically insignificant thereafter, suggesting that troopers revert to their pre-treatment behavior within a year of training.

To benchmark the magnitude of the short-run effect, we compare our estimate to a commonly-used test for police discrimination in the literature, the so-called "veil of darkness" (VOD) test, which compares stop composition during darkness and daylight. Our estimate of the effect of training is about sixty percent as large as a benchmark VOD estimate, suggesting that cultural diversity training erodes over half the discrimination against Black and Hispanic motorists among trained officers in the short run.

Collapsing the traffic stop data into a standard panel at the trooper × week level, we next show that troopers achieve the change in the racial makeup of their stops by stopping about one additional white motorist per week ($se = 0.26$), with minimal, concordant changes in the number of stops of Black or Hispanic motorists. To accompany this finding, we explore changes in stop outcomes after diversity training, separately for white and minority motorists. Following training, the likelihood that a stopped white motorist is searched or arrested falls, while the same probabilities remain constant for stopped minority motorists. The dynamics of these effects mirror those for racial stop composition, reverting to zero within a year.

Taken together, we interpret these results as informative about changes in the composition of a trooper's stops. Specifically, our findings suggest that troopers stop additional, less "guilty" whites after training. These marginal motorists are less likely to be searched or arrested, reducing troopers' average search and arrest rates of stopped white motorists. However, we cannot rule out the alternative hypothesis that diversity training changes a trooper's propensity to search or arrest motorists of different backgrounds.

Overall, our findings clearly suggest the potential for cultural diversity training to change employee behavior on-the-job but also come with several caveats. First, the evidence we present suggests that troopers respond to training by reducing their lenience towards white motorists, rather than changing their behavior towards minorities. In many settings, policing included, this outcome may not be the desired result of a diversity training initiative. Indeed, police reform proposals often advocate for reducing interactions between minorities and patrol officers given the potential risks of escalation (Woods, 2021). Second, we find that effects on behavior do not persist in the long-run. While this downside may be addressed by delivering training regularly, our empirical framework and institutional setting do not allow us to speak to the effectiveness of repeated trainings.

In addition to providing novel evidence on the effects of a cultural diversity training, our paper contributes to a broad literature in economics on worker training programs (e.g., Becker 1964). Given our focus on the effects of training on the job performance of public-facing, public sector workers, our paper is most related to the literatures on teacher and police training. Evidence on the effectiveness of a variety of teacher training programs on student outcomes is mixed (e.g., Angrist and Lavy 2001; Bressoux et al. 2009; Jacon and Lefgren 2004; Harris and Sass 2011). Most related to our analysis is Tumen et al. (2022), who study the effects of diversity training for teachers in Turkey and find that training reduces absenteeism among refugee students.

Police training is a topic that has received significant public attention in recent years as calls for police reform have increased in the wake of several high-profile, police-involved killings (Crabtree, 2020). While proposals such as defunding the police and eliminating police enforcement of nonviolent crimes are supported by less than half of Americans, 85 percent favor expanded training (Ipsos, 2021). Nonetheless, evidence on the effects of police training on enforcement behavior is both limited and mixed. Randomized control trials have yielded some evidence that procedural justice training and cognitive behavioral therapy can reduce officer use of force and low-level arrests (e.g., Wheller et al. 2013; McLean et al. 2020; Owens et al. 2018; Dube et al. 2023), while Adger et al. (2023) document the importance of field-training on officer outcomes. Our study is relatively unique in documenting the effects of a training aimed specifically at officer behavior towards minority groups.

In that vein, our paper also adds to a broad literature on racial disparities in the criminal justice system. While many studies have documented racial disparities in police behavior and tested for discrimination by the police (e.g., Doleac 2022; Knowles et al. 2001; Anwar and Fang 2006; West 2021; Goncalves and Mello 2021), little evidence on the potential for policy interventions to mitigate racial discrimination has emerged. One exception is a strand a literature showing the importance of the racial makeup of the police force in explaining racial disparities in outcomes (McCrary 2007; Ba et al. 2021; Rivera 2022). Our paper suggests a potential role for cultural diversity training programs in changing the racial attitudes of police officers, at least in the short run.

The remainder of our paper is organized as follows. In section 2, we describe the relevant institutional details and data. Section 3 lays out our empirical strategy and section 4 presents the results. We conclude in section 5.

2 Setting and Data

2.1 Institutional details

Training for highway patrol officers in Texas is divided into three distinct phases: basic academy training, field training, and in-service training. In the first and second phases of training, new recruits to the Texas Highway Patrol (THP) complete approximately 1,050 hours of basic academy training and 350 hours of field training. The third phase of training consists of legislatively mandated and unit-specific in-service training courses which are taken continuously throughout an officer's career. Our focus is on cultural diversity training taken during the third phase.

In-service training requirements depend on a trooper's proficiency status. Officers are granted a basic proficiency certificate after completing basic academy training and are then required to work towards achieving intermediate, advanced, and masters proficiency certificates. While the specific incentives vary across agencies, peace officers throughout Texas generally receive pay increases when reaching higher proficiency levels. To advance from basic to intermediate proficiency, THP troopers must reach two to four years of service (depending on prior education and military service) and complete a set of 17 courses.[2]

There are four courses (cultural diversity, crisis intervention, deescalation, and special investigation topics) which are required to be taken once in each four-year training cycle until intermediate proficiency is achieved.[3] The focus of our analysis is an officer's first

[2] See appendix B-3 additional institutional details, including an explanation of the intermediate proficiency requirements. Note that one of the required trainings is a course on racial profiling which is conceptually distinct from cultural diversity training. THP's profiling training is only four hours long and focuses on the legal environment in Texas, emphasizing the distinction between a "racially motivated" stop and reasonable suspicion under state and federal law.

[3] The relevant four-year training cycles during our study period are 09/01/2005–08/31/2009; 09/01/2009–08/31/2013; 09/01/2013–08/31/2017; and 09/01/2017–08/31/2021. Troopers begin-

cultural diversity training, taken prior to reaching intermediate proficiency. We discuss the potential complications associated with identifying effects of diversity training at a point in a trooper's career when other trainings are also occurring in more detail in section 3.

Officers schedule training, which is offered at a state-level academy in Austin and 50 other locations throughout the state, at their own discretion.[4] Course offerings with capped enrollments are scheduled in advance by each individual training academy, meaning that troopers have imperfect control over the precise date and location of their trainings. Given the large geographic dispersion of THP's jusrisdiction and officer residences, constraints on course availability, and the need to schedule training in advance, we argue that it is largely idiosyncratic where and when troopers complete in-service training courses and provide support for this view below in sections 3 and 4.

Note that the training received by THP troopers is broadly representative of police training throughout the country. Relative to the average police agency (Bureau of Justice Statistics 2018; Bureau of Justice Statistics 2020), THP troopers receive fewer hours of field training but more hours of classroom or simulation-based training, including about 200 percent more in-service training. Over the past several decades, the International Association of Directors of Law Enforcement Standards and Training (IADLEST) has issued a core set of recommendations for in-service training that have been broadly adopted across the country. In-service training requirements for the THP align closely with IADLEST's recommendations both generally and with specific regard to cultural diversity.

Cultural diversity training for law enforcement officers is largely geared towards making officers more effective at their job (which may differ from the goals of similarly-named public or private sector trainings aimed primarily at changing workplace culture). While cultural diversity training in law enforcement has been around since the 1960's, its modern incarnation came into existence during the 1990's (Hennessy, 2001). The stated goal of the modern form of cultural diversity training is to improve officers' understanding of their changing communities and to enhance their ability to effectively communicate with populations of varied backgrounds.

Rather than emphasizing racial or ethnic sensitivities or mitigating implicit bias, the training focuses pragmatically on teaching officers to understand cultural differences amongst the populations that they serve. The course is framed to officers as a key set of skills that can help them elicit community trust and better de-escalate volatile situations. In Texas, cultural diversity training is divided into two four-hour blocks which are typically taken sequentially. The first block consists of two modules that all officers are required to take: *introduction to diversity* and *cultural diversity*. These blocks include a lesson on the changing demographics

ning their careers at least two years into the current cycle are given until the end of the following cycle to complete required trainings.

[4] See https://www.tcole.texas.gov/law-enforcement-academies for a map of training locations. In our sample, about 60 percent take diversity training in Austin.

Texas and critical discussions of issues such as fairness versus equal treatment and the role of police as members of their communities. In the second block, officers learn about diversity from a variety of modules that include *generational diversity*, *workplace diversity*, *gender diversity*, and *law enforcement as a diverse culture*. These trainings are designed to be "hands on, interactive, and scenario based," with roleplaying that is oriented to day-to-day experiences on the job. The course is typically taught by a senior police officer with substantial real-world experience, a potentially important departure from other race-focused trainings which are often taught by civilians such as academics or social workers.[5]

While cultural diversity training is nominally focused on developing officers' ability to communicate with diverse populations, we note that such a training could have broad effects on officers' racial attitudes more generally. For example, prompting officers to consider their communication tactics when engaging with citizens of different backgrounds could make the racial and ethnic composition of their interactions more salient, could induce officers to more seriously consider citizen backgrounds before initiating an enforcement activity, or could affect latent racial biases. Given the lack of evidence on the effects of diversity training on behavior, we focus on documenting the responses to training in terms of enforcement behavior without a strong prior on exactly how training should affect enforcement decisions.

2.2 Data

Our analysis relies on administrative records of all traffic stops made by the Texas Highway Patrol over the period 2006–2019, provided by the Texas Department of Public Safety (TDPS). These records include officer badge numbers and detailed information about the stop, including date, time, and GPS coordinates, which we map to counties and census tracts. The data also include information about motorists including race, age, and gender, and information about the outcome of the stop: whether a citation was issued and, if so, for what violation, whether a search was conducted, whether contraband was found, and whether an arrest was made.[6]

[5] One of the coauthors of this study completed an online version of the training offered through a TCOLE-approved provider. In our opinion, the training was similar to diversity trainings we have encountered at academic institutions with a few key exceptions above and beyond the specific focus on law enforcement scenarios. First, at eight hours, the training was significantly longer. Second, each module was followed by examinations which were surprisingly challenging. The participant was only allowed two attempts and was required to complete each module-specific exam with a score of at least 70 percent. The training explicitly stated that failure would be reported to TCOLE and would require re-taking the entire course. In our opinion, the median trooper would need to take careful notes in order to pass the course on the first try.

[6] Luh (2022) presents evidence that THP troopers systematically misreport Hispanic motorists as white in order to mask their racial profiling behavior. In figure A-10, we repeat our main analysis using an alternative definition of Hispanic status based on surnames (e.g., Goncalves and Mello 2021) and find similar results.

We match officers in the traffic stops data to historical, course-level rosters for all trainings administered to THP troopers from the Texas Commission on Law Enforcement (TCOLE), the state-level agency that administers civil service requirements and certifications for law enforcement officers. We use these records to identify the date at which each trooper completes various trainings, including both academy and in-service trainings. Supplementary information on trooper demographics was provided by the Texas state comptroller's office. These data include race/ethnicity, gender, age, and hire date corresponding to each law enforcement employment spell for the troopers in our sample. For additional details on the data sources, see appendix B.

Our analysis sample is comprised of 1,723 troopers who we observe starting their careers in 2009 or after and who can be matched to both the demographics and TCOLE datasets. Table A-1 reports summary statistics for this analysis sample of troopers. Troopers are around 27 years of age at career start. The vast majority of troopers are male (92 percent) and either white (54 percent) or Hispanic (36 percent). During a typical week on the job, troopers in our sample make about 25 traffic stops and write 9 citations.

Note that our traffic stops data include both stops which result in a citation and stops which result in a formal (written) warning. Because officers can also issue informal (verbal) warnings which will not appear in our data, and because only citations result in actual sanctions for motorists, we focus primarily on the subset of our data where a citation is issued. However, we also examine the effects of training on the total number of stops (citations and written warnings) and on the share of these stops resulting in a citation. Whenever relevant, we repeat our core analyses in the appendix using both citations and formal warnings.[7]

3 Empirical approach

Our empirical approach examines changes in officer behavior around the precise timing of cultural diversity training using an event study design. Specifically, our goal is to estimate regressions of the form:

$$Y_{ijst} = \sum_{\tau} \theta_{\tau} + \alpha_j + \delta_t + \psi_s + u_{ijst} \tag{1}$$

where i indexes traffic stops, j indexes troopers, t indexes time, s indexes patrol assignments (e.g., location, day of week, hour), and $\tau = t - \tilde{t}_j$ indexes event time (where $\tilde{t}_j$ denotes the period in which officer j receives training). Our primary outcome of interest is an indicator for whether a stopped motorist is white.

Recent advances in the econometrics literature have documented the various empirical issues associated with estimating event studies with two-way fixed effects (TWFE) via OLS

[7] Senate Bill 1849 prohibited the use of informal (i.e., undocumented) warnings beginning in 2018, which would be relevant for the final two years of our sample, at least in theory. However, we are skeptical that such a law is enforced in practice. Consistent with our skepticism, figure A-13 documents no substantive increase in warnings appearing in our data beginning in 2018.

(e.g., Chaisemartin and D'Haultfoeuille 2020; Goodman-Bacon 2021; Sun and Abraham 2021; Callaway and Sant'Anna 2021; Borusyak et al. 2022; Roth et al. 2022). Important concerns raised in this literature include the contamination of treatment effect estimates created by comparisons between currently treated and previously treated units and underidentification issues in fully dynamic specifications.

To address these issues, we estimate event studies using the two-step imputation estimator proposed by Borusyak et al. (2022) and Gardner (2021). In the first step, the fixed effects are estimated by regressing the outcome on fixed effects using only the not-yet-treated observations. In the second step, differences between observed outcomes and predicted outcomes, based on the estimated fixed effects in the first step, are averaged to construct the event study estimates. For additional details, see appendix C.

This approach is particularly well-suited to our setting for three important reasons. First, the imputation estimator can be applied to data structures other than standard panel data; our dataset is at the level of the traffic stop, rather than at the level of the officer × time (although we also construct a panel dataset, described in more detail below). Second, the imputation estimator is more computationally manageable than traditional "stacking" approaches, which typically become very computationally burdensome with many different treatment timings. Third, this approach easily accommodates a more complex fixed effects structure than TWFE, which is important in our setting because we want to account for the fact that officers patrol very heterogeneous geographic areas, for example.

To that end, our baseline regressions include trooper fixed effects, exact date fixed effects, fixed effects at the level of the census tract $\times$ $\mathbf{1}$[weekend] $\times$ shift, where shifts are three eight-hour partitions of the day, and county × year × month fixed effects. Our census tract $\times$ $\mathbf{1}$[weekend] $\times$ shift effects ensure that comparisons underlying our estimates are only between treated and not-yet-treated officers patrolling the same beat and shift and our county × year × month fixed effects adjust for secular changes over time in the racial makeup of neighborhoods. Note that, while we condition on these detailed fixed effects in our analyses, figure A-3 illustrates that trooper assignments do not change systematically around the timing of training, which is reassuring for the validity of our research design.

We define event time using eight-week bins around the week of training and focus on the six periods ($\approx$ one year) before and after treatment.[8] To summarize the event study estimates, we report the average of the period-specific estimates for $\tau = 0$ through $\tau = 2$, or the first 24 weeks after training. For inference, we compute standard errors and confidence bands using a Bayesian bootstrap (Rubin, 1981), clustering at the trooper-level.

While analyzing the data at the traffic-stop level is ideal for examining the effect of training on the racial composition of stops, an important drawback of this approach is that

[8] As shown in figure C-2, the effective number of observations informing the computation of the event study parameters drops significantly beyond the six period window. We use eight-week bins, rather than a coarser definition of event time, to improve statistical precision.

we cannot examine effects of training on stop volume. To study the effects of training on troopers' number of traffic stops, we aggregate the traffic stops data into a panel dataset at the level of officer × week. We can then use the same imputation-based approach to assess how an officers number of weekly stops evolves after diversity training.

Importantly, once the data have been aggregated up to the officer × week level, adjusting for patrol assignments is less straightforward; while each traffic stop can be easily assigned to a location and time, each officer-week cannot. For our panel data regressions, we assign each officer × week to a county and shift using the empirical distribution of stops, as described in appendix B-2. We then condition on assigned county × assigned days of the week × assigned shift fixed effects and assigned county × year × month fixed effects, in addition to trooper and week fixed effects, in our event study estimates.

A relevant consideration for our analysis is the fact that troopers also receive other trainings in traffic stop and arrest protocol, as well as racial profiling and deescalation training, during this early-career period. Our estimates of the impact of cultural diversity training on trooper behavior will be biased if the timing of diversity training is correlated with the timing of other trainings *and* those other trainings impact behavior.

Panel (a) of figure A-4 verifies that this is a salient concern by documenting a statistically significant relationship between the timing of cultural diversity training and the timing of other trainings. However, this correlation is quantitatively small: about 2.5 percent of troopers receive a different training in the same eight week period as their diversity training. Moreover, figure A-5 illustrates that, if anything, these other trainings have the opposite effect as that of diversity training, meaning that any bias in our estimates attributable to other trainings should be towards zero.

4 Results

4.1 Main results

Figure 1 examines the racial composition of trooper citations around the timing of cultural diversity training. Specifically, this figure reports event study estimates where the outcome is an indicator for whether a cited motorist is white. As described above, we measure event time using eight-week bins, dropping the exact week of training. Hence, $\tau = 0$ includes weeks one through eight after training, $\tau = 1$ includes weeks 9 through 16 after training, and so on. First, we note that there is no evidence of differential changes in trooper behavior during the period leading up to training. Using the test prescribed by Borusyak et al. (2022), we cannot reject the null hypothesis of parallel trends in stop composition prior to training ($p = 0.27$).

However, immediately following training, the probability that a cited motorist is white increases and remains elevated for about 32 weeks. Our short-run point estimate, obtained by averaging the estimates over the first 24 weeks following training, is 0.013 ($se = 0.005$).

This represents about a four percent increase in likelihood that a stopped motorist is white, relative to a counterfactual mean of 34 percent.[9] In terms of dynamics, the estimated impact of training peaks at around two percentage points during the 16-24 weeks following training, then recedes back towards zero. Estimates for periods greater than 32 weeks after training are statistically indistinguishable from zero, suggesting that training has fairly short-lived effects on trooper behavior.[10]

Figure 2 repeats the analysis from figure 1, splitting the sample by trooper race, and reveals that nearly the entire effect of diversity training can be attributed to changes in the behavior of white troopers. While short-run effects for nonwhite troopers are statistically indistinguishable from zero, white troopers increase their white share of citations by about 1.8 percentage points ($se = 0.008$) in the 6 months following training.

A natural question is whether the effects we document are large or small. On one hand, a 1.5 percentage point (four percent) increase in the share of citations which are of white motorists may seem like a small effect in absolute terms, *prima facie*. On the other hand, if the effect of cultural diversity training is to mitigate discriminatory behavior by officers, then the magnitude or our treatment effect estimates should be bounded above by the extent of discrimination among troopers.

To benchmark our estimated magnitude, we compare our findings on the racial composition of stops and citations to results from a common test for police discrimination from the literature, the so-called *veil of darkness* (VOD) test (Grogger and Ridgeway 2006; Horrace and Rohlin 2016; Ross et al. 2023). Described further in appendix D, this test compares the racial composition of stops made during daylight and darkness but at the same time of day, with the idea being that an overrepresentation of minorities during daylight (e.g. during the spring and summer) suggests racial profiling when race is observable to troopers *prior* to making a stop.

Our benchmark VOD estimate implies that a stopped motorist is about two percentage points ($se = 0.0015$) less likely to be white during daylight. In other words, our estimate of the short-run causal effect of cultural diversity training on the fraction of white stops is about sixty percent as large as an estimate of share of white stops attributable to discrimination or racial profiling behavior by troopers.

[9] Post-treatment counterfactual means are estimated by regressing the outcome on the fixed effects using using only the sample of not-yet-treated observations and averaging predictions from this regression over event time (e.g., Kleven et al. 2020, Borusyak et al. 2022).

[10] In the appendix, we verify the robustness of this central finding. Specifically, estimates are similar when examining both citations and formal warnings as shown in figure A-6, when controlling for officer experience as shown in figure A-8, and when varying the set of fixed effects as shown in figure A-9.

4.2 Mechanisms

Exactly what changes in officer behavior underlie the observed changes in the composition of cited drivers that we document? To shed light on this question, we first explore the effects of training on a trooper's stop volume, using a panel dataset constructed by aggregating the stop data up to the trooper × week level. Note that here, we first focus on the total number of stops, which includes both citations and formal warnings. As described in appendix B-2, in these regressions we condition on a trooper's assigned county and shift which we infer from the data. Panel (a) of figure 3 documents a dramatic increase in a trooper's weekly number of stops of white motorists following diversity training. Over the six months after training, troopers make an average of 0.9 ($se = 0.26$) additional stops of white motorists per week, about a ten percent increase relative to the counterfactual mean.

On the other hand, changes in the number of stops of Black and Hispanic motorists are significantly attenuated, with magnitudes less than half as large, both in absolute terms and as a proportion of the relevant means. Moreover, the short-run estimates for Black and Hispanic stops are not statistically distinguishable from zero.

In panel (b) of figure 3, we return to our traffic stop-level specification and present event study estimates where the outcome of interest is whether a traffic stop results in a citation, as opposed to a formal warning, estimated separately by motorist race. The figure illustrates a clear pattern of officer behavior following training: the fraction of stopped white motorists receiving a citation declines slightly, while the fraction of stopped nonwhite motorists receiving a citation increases slightly. Nonetheless, as evidenced by figure 1, these changes in citation (versus warning) rates are not sufficient to "undo" the relative increase in the number of stops of white motorists shown in panel (a) of figure 3.

The patterns presented in figures 1 and 3 suggest that following training, troopers are more likely to stop "marginal" white motorists that they were letting pass prior to training. If so, we might also expect that these newly stopped white motorists are less "guilty." To further probe this hypothesis, we next estimate event studies focusing on additional stop outcomes such as whether a search or arrest is conducted, separately by motorist race.

As shown in panels (a) through (c) of figure 4, citations of white motorists made after training are less likely to result in searches, hits (defined as whether contraband is found in a search), and arrests. Combined short-run estimates over the 24 weeks after training are marginally statistically significant for searches and arrests, while the period-specific estimates for the second period after training are significant in all cases. On the other hand, there is no evidence of changes in the rate at which stops of nonwhite motorists result in searches, hits, or arrests.[11] Note that the dynamic patterns in the estimates for white motorists mirror those from the stop composition estimates in figure 1 and the stop volume estimates in figure 3: the rate at which citations of white motorists results in searches, hits, and arrests increases

[11] Figure A-11 documents very similar patterns when examining all stop, including warnings.

as the white share of citations increases, then trends back to pre-training rates as the white share estimates trend back towards zero.

We interpret the findings from the first three panels of figure 4 as signaling a change in the composition of cited white motorists after diversity training. Specifically, the evidence suggests that troopers' stop less "guilty" white motorists after training, which is consistent with the evidence on the racial composition of stops and stop volumes discussed above.

Alternatively, it may be that diversity training also changes a trooper's propensity to search or arrest differentially by motorist race. In this case, one cannot necessarily interpret the patterns in panels (a)-(c) of figure 4 as informative about changes in sample composition. We present a simple test of this alternative hypothesis in panel (d) of figure 4, which reports event study estimates for the *conditional* hit rate: the probability that a search turns up contraband. If training induces troopers to search white motorists less often, we should expect to see the conditional hit rate increase for white motorists relative to nonwhite motorists following training. Unfortunately, because this test relies only on the small fraction of citations resulting in a search, large standard errors render this test inconclusive. However, we acknowledge that the figure is at least suggestive of a relative increase in conditional hit rates for white motorists. Overall, we interpret the evidence in figure 4 as suggestive of a change in the composition of stopped white motorists after training, while noting that an effect of training on officer post-stop decisions may also play a role.

Note that the patterns in figures 3 and 4 for nonwhite motorists tell a less clear story. In particular, increases in citation rates for nonwhite motorist after training (which persist for a full year) are somewhat surprising in light of both the substance of the training and the other results we present. Potential explanations for the heightened citation rates of nonwhite motorists include the hypothesis that officers become more selective in terms of *which* nonwhite motorists to stop or that officers reduce their number of pre-textual stops, rates of which are typically higher among minority motorists. While both these hypotheses would also predict changes in the post-stop outcomes shown in figure 4 for nonwhites, whereas we find null effects for nonwhite motorists, West (2019) shows that inexperienced officers are much worse at predicting the outcome of minority vehicular searches. Hence, we may be unable to detect changes in post-stop outcomes for minorities even if officers were becoming more selective about which nonwhite motorists to stop.

5 Conclusion

In this paper, we study the effects of cultural diversity training on the enforcement behavior of early-career Texas Highway Patrol officers. Leveraging variation across officers in the precise timing of training using an event study approach, we find that the racial makeup of a trooper's set of stopped and cited motorists changes systematically in the six months following training. The probability that a cited motorist is white (as opposed to nonwhite) increases by 1.3 percentage points, or four percent relative to the control mean. Benchmarked against a standard estimate of racial bias in the literature (applied to our specific setting), our short-run estimates suggest that diversity training erodes over half of the discrimination practiced by the average officer.

Troopers achieve the change in stop composition that we document by stopping and citing additional white motorists, rather than stopping fewer minority motorists. We present suggestive evidence that these additional white motorists are less “guilty” by showing that the likelihood that a white motorist is searched or arrested falls after training. For nonwhite motorists, these post-stop outcome measures are unaffected by training. This aligns well with the theory that diversity training prompts troopers to stop marginal white motorists whom they were letting pass prior to training, potentially eroding an important margin of discrimination by officers: lenience towards whites (e.g., Goncalves and Mello 2021).

However, more stringent attitudes towards white motorists, coupled with no change in the enforcement of nonwhite motorists, may not reflect the desired outcome of cultural diversity training. Many modern proposals in police reform, for example, explicitly aim to reduce the number of police-civilian interactions (e.g., Woods 2021), and the implied goal of diversity training in most instances appears to be increasing sensitivity to and understanding of diverse groups. Moreover, we find that behavioral changes induced by diversity training only persist in the short-run, with officers reverting to their pre-training enforcement behavior within one year, suggesting a limited ability for one-off trainings to create long-run changes.

References

Adger, C., M. Ross, and C. Sloan (2023). The effect of field training officers on police use of force. *Unpublished manuscript*.

Angrist, J. and V. Lavy (2001). Does teacher training affect pupil learning? Evidence from matched comparisons in Jerusalem public schools. *Journal of Labor Economics 19*(2), 343–369.

Anwar, S. and H. Fang (2006). An alternative test of racial prejudice in motor mehicle searches: Theory and evidence. *American Economic Review 96*(1), 127–151.

Ba, B., D. Knox, J. Mummolo, and R. Rivera (2021). Diversity in policing: The role of officer race and gender in police-civilian interactions in Chicago. *Science 371*(6530), 696–702.

Becker, G. (1964). *Human Capital*. University of Chicago Press.

Bezrukova, K., K. Jehn, and C. Spell (2012). Reviewing diversity training: where we have been and where we should go. *Academy of Management Learning and Education 11*(2), 207–227.

Borusyak, K., X. Jaravel, and J. Spiess (2022). Revisiting event study designs: Robust and efficient estimation. *Review of Economic Studies*.

Bressoux, P., F. Kramarz, and C. Prost (2009). Teacher training, class size, and student outcomes: Learning from administrative forecast mistakes. *The Economic Journal 119*(536), 540–561.

Bureau of Justice Statistics (2018). State and Local Law Enforcement Training Academies, 2018 – Statistical Tables. https://bjs.ojp.gov/sites/g/files/xyckuh236/files/media/document/slleta18st.pdf.

Bureau of Justice Statistics (2020). Local Police Departments: Policies and Procedures, 2016. https://bjs.ojp.gov/content/pub/pdf/lpdpp16.pdf.

Calfas, J. (2018). Was Starbucks' racial bias training effective? Here's what these employees thought. *Time Magazine*.

Callaway, B. and P. Sant'Anna (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics 225*(2), 200–230.

Chaisemartin, C. and X. D'Haultfoeuille (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review 110*(9), 2964–96.

Chang, E., K. Milkman, and A. Duckworth (2019). The mixed effects of online diversity training. *Proceedings of the National Academy of Sciences 116*(16), 7778–7783.

Crabtree, S. (2020). Most Americans say policing needs major changes. *Gallup*.

Dobbin, F. and A. Kalev (2016). Why diversity programs fail. *Harvard Business Review*.

Doleac, J. (2022). Racial bias in the criminal justice system. *A Modern Guide to the Economics of Crime*.

Dube, O., S. MacArthur, and A. Shah (2023). A cognitive view of policing. *Unpublished manuscript*.

Gardner, J. (2021). Two-stage difference-in-differences. *Unpublished Manuscript*.

Goncalves, F. and S. Mello (2021). A few bad apples? racial bias in policing. *American Economic Review 111*(5), 1406–1441.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics 225*(2), 254–277.

Grogger, J. and G. Ridgeway (2006). Testing for racial profiling in traffic stops from behind a veil of darkness. *Journal of the American Statistical Association 101*(475), 878–887.

Harris, D. and T. Sass (2011). Teacher training, teaching quality, and student achievement. *Journal of Public Economics 95*(7), 798–812.

Hennessy, S. (2001). Cultural awareness and communication training: What works and what doesn't. *Journal of Police Chiefs 68*(11), 15–19.

Horrace, W. and S. Rohlin (2016). How dark is dark? bright lights, big city, racial profiling. *Journal of the American Statistical Association 98*(2), 226–232.

Ipsos (2021). USA Today/Ipsos Crime and Safety Poll. https://www.ipsos.com/sites/default/files/ct/news/documents/2021-07/Topline-USAT-Crime-and-Safety-070821.pdf.

Jacon, B. and L. Lefgren (2004). The impact of teacher training on student achievement: Quasi-experimental evidence from school reform efforts in Chicago. *Journal of Human Resources 39*(1), 50–79.

Kleven, H., C. Landais, and J. Sogaard (2020). Children and gender inequality: Evidence from Denmark. *American Economic Journal: Applied Economics 11*(1), 181–209.

Knowles, J., N. Persico, and P. Todd (2001). Racial bias in motor vehicle searches: Theory and evidence. *Journal of Political Economy 109*(1), 203–229.

Luh, E. (2022). Not so black and white: Uncovering racial bias from systematically misreported trooper reports. *Unpublished manuscript*.

McCrary, J. (2007). The effect of court-ordered hiring quotas on the composition and quality of the police. *American Economic Review 97*(1), 318–353.

McLean, K., S. Wolfe, and J. Rojek (2020). Randomized controlled trial of social interaction police training. *Criminology and public policy 19*(3), 805–832.

Newport, F. (2016). Public opinion contest: Americans, race, and police. *Gallup*.

Owens, E., D. Weisburd, K. Amendola, and G. Alpert (2018). Can you build a better cop? Experiental evidence on supervision, training, and policing in the community. *Criminology and public policy 17*(1), 41–87.

Pierson, E., C. Simoiu, and J. Overgoor (2020). A large-scale analysis of racial disparities in police stops across the United States. *Nature Human Behavior 4*, 736–745.

Rivera, R. (2022). The effect of minority peers on future arrest quantity and quality. *Unpublished manuscript*.

Ross, M., S. Ross, and J. Kalinowski (2023). Endogeneous driving behavior in tests of racial profiling in traffic stops. *Journal of Human Resources*.

Roth, J., P. Sant'Anna, A. Bilinski, and J. Poe (2022). What's trending in difference-in-differences? A synthesis of the recent econometrics literature. *Unpublished Manuscript*.

Rubin, D. (1981). The Bayesian bootstrap. *The Annals of Statistics 9*(1), 130–134.

Shen, L. (2017). Delta adds diversity training for 23,000 crew members. *Fortune Magazine*.

Sun, L. and S. Abraham (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics 225*(2), 175–199.

Tumen, S., M. Vlassopoulos, and J. Wahba (2022). Training teachers for diversity awareness: Impacts on school attendance of refugee children. *IZA Disucssion Paper 14557*.

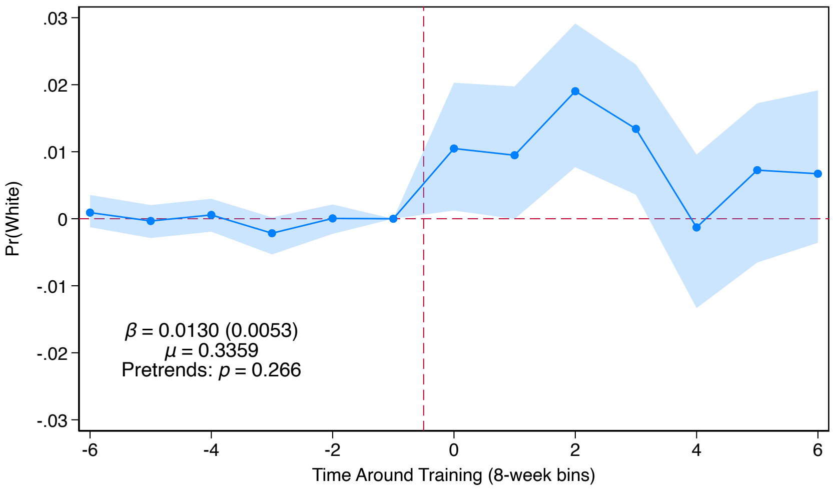
West, J. (2019). Learning by Doing in Law Enforcement . *Unpublished manuscript*.

West, J. (2021). Racial bias in police investigations. *Unpublished manuscript*.

Wheller, L., P. Quinton, A. Fildes, and A. Mills (2013). The greater Manchester police procedural justice experiment. *Coventry, UK: College of Policing*.

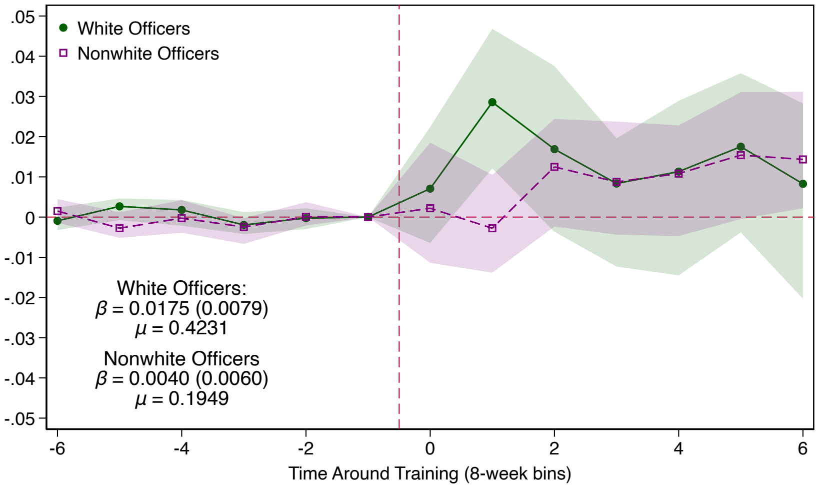
Woods, J. (2021). Traffic enforcement would be safer without police. Here's how it could work. *Washington Post*.

Figure 1: Racial composition of citations around training



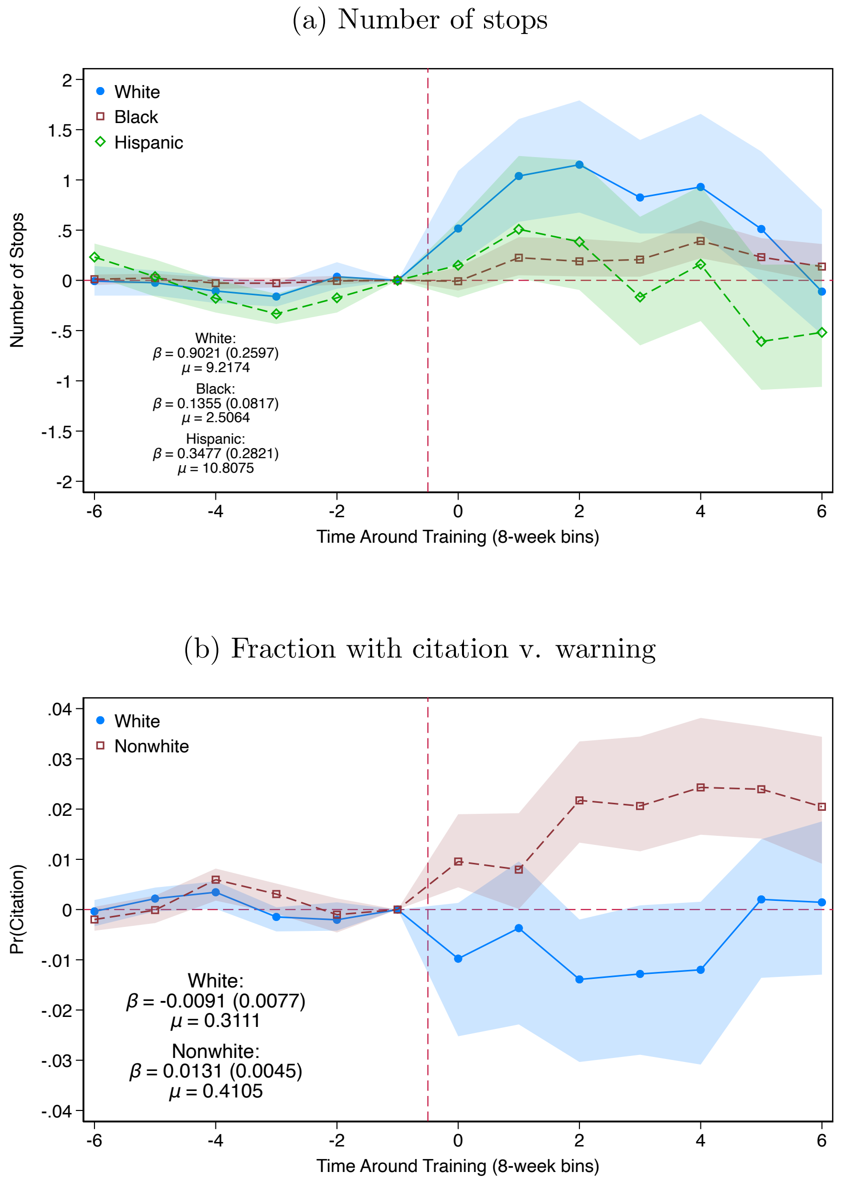
*N*otes: This figure plots imputation-based event study estimates and 95 percent confidence bands using the sample of citations where the outcome is an indicator for whether a cited motorist is white. Regression controls for trooper, date, census tract × weekend × shift, and county × year × month fixed effects. The figure reports the average of the event study coefficients for $\tau \in [0, 2]$ and associated standard error, as well as the estimated counterfactual mean over the same period and the p-value from a test of parallel pre-treatment trends.

Figure 2: Effects on racial composition by officer race



*N*otes: Same as figure 1 except that event studies are estimated separately by officer race.

Figure 3: Effects on stop volume and citation rates

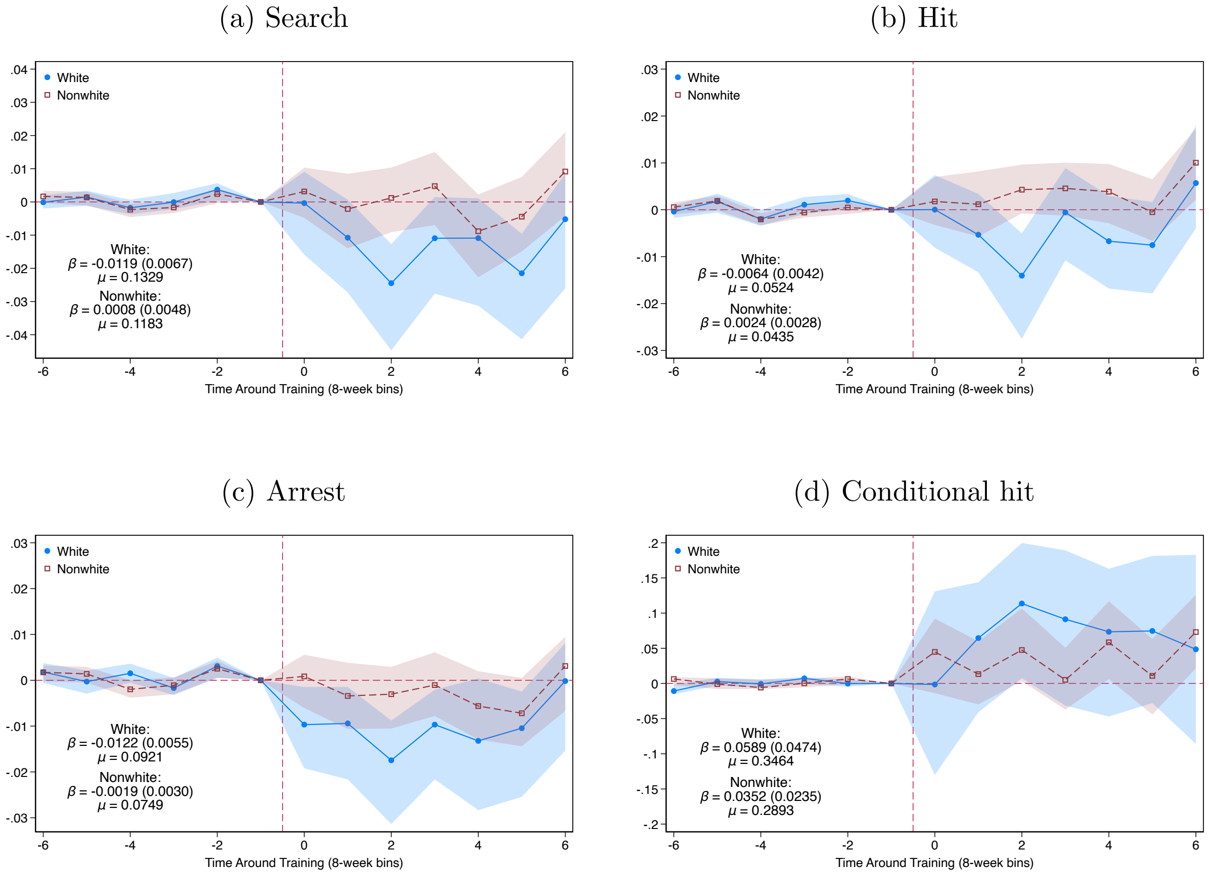


(a) Number of stops

(b) Fraction with citation v. warning

*N*otes: Panel (a) plots imputation-based event study estimates using an officer × week panel where the outcomes are the number of stops by motorist race. Panel (b) plots estimates identical to those in figure 1 except that we use all traffic stops and the outcome of interest is whether a traffic stop results in a citation (as opposed to a formal warning). In panel (b), regressions are estimated separately by whether a stopped motorist is white or any other race/ethnicity.

Figure 4: Post-citation outcomes by motorist race


(a) Search

(b) Hit

(c) Arrest

(d) Conditional hit

*N*otes: Each panel plots imputation-based event study estimates and 95 percent confidence bands using all traffic stops with a citation. Regressions are the same as those in figure 1 except that they are estimated separately by motorist race. In panel (a), the outcome is whether the motorist is searched. In panel (b), the outcome is whether contraband is found. In panel (c), the outcome is whether the stop ends in an arrest. In panel (d), the outcome is whether contraband is found using only stops where a search is made.

FOR ONLINE PUBLICATION: APPENDICES

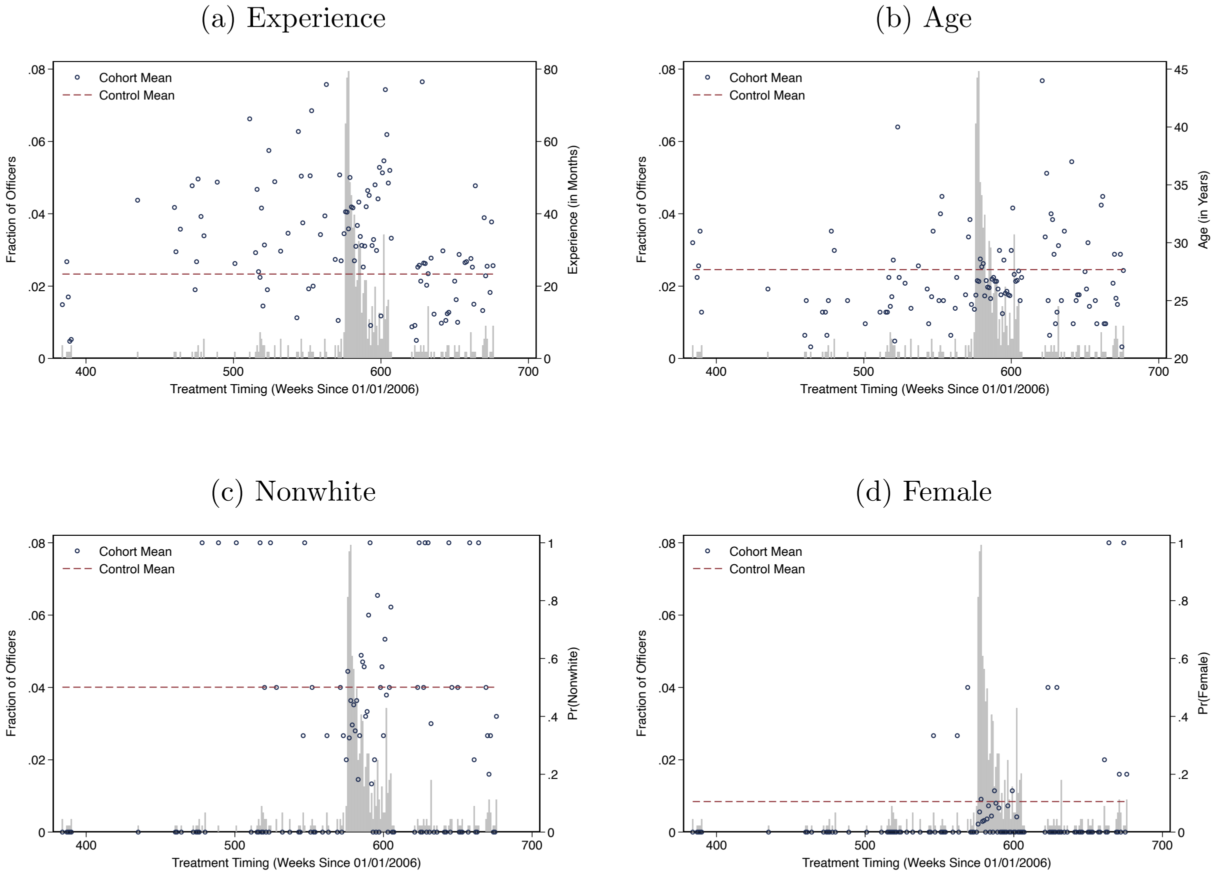
A Supplementary results

Table A-1: Summary statistics for troopers in analysis sample

				By Treatment Timing		
	All	Treated	Untreated	Cycle 1/2	Cycle 3	Cycle 4
	(1)	(2)	(3)	(4)	(5)	(6)
Age	27.61	27.45	27.83	29.06	27.01	27.80
Male	0.917	0.934	0.894	0.961	0.933	0.888
Race = White	0.543	0.579	0.494	0.650	0.552	0.663
Race = Black	0.071	0.059	0.087	0.028	0.065	0.079
Race = Hispanic	0.363	0.339	0.395	0.294	0.362	0.247
Race = Other	0.023	0.022	0.025	0.028	0.022	0.011
Experience at Training	28.43	28.43	–	0.661	35.96	23.12
Weekly Stops	25.03	24.93	25.18	25.34	24.71	25.85
Share White	0.451	0.495	0.391	0.562	0.478	0.501
Weekly Citations	8.91	9.00	8.80	9.47	8.89	8.96
Share White	0.384	0.430	0.319	0.506	0.413	0.417
Troopers	1,723	996	727	180	727	89

*N*otes: This table reports summary statistics for troopers in the analysis sample. Age is measured as of the officer's hire date. Experience at training is an officer's months of writing citations prior to cultural diversity training. Column (4) reports means for troopers who take training before September 2013, column (5) reports means for troopers who take training between September 2013 and August 2017, column (6) reports means for troopers who take training after Septermber 2017.

Figure A-1: Event study cohorts



(a) Experience

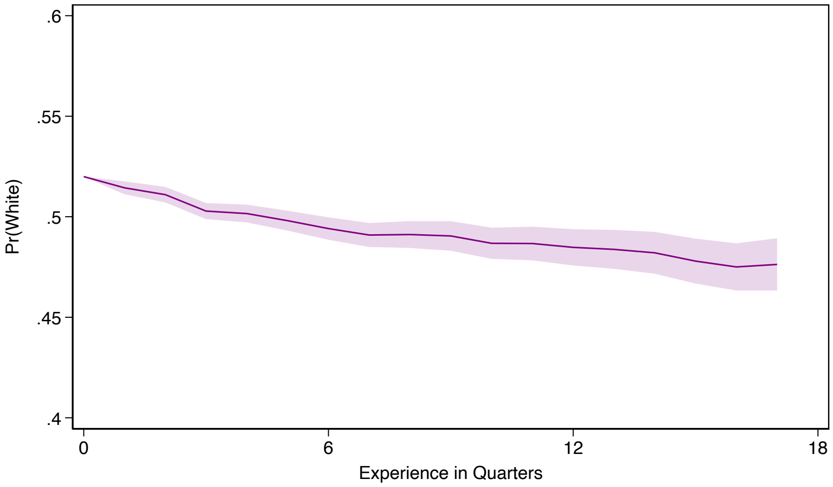
(b) Age

(c) Nonwhite

(d) Female

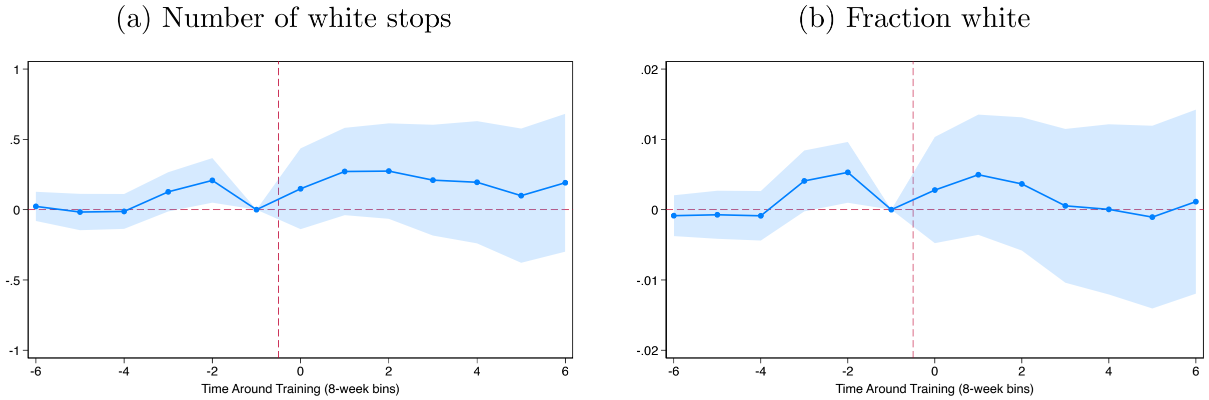
*N*otes: Each figure plots a histogram of treatment timing (gray bars; left axis) as well as the average characteristics of troopers in each treatment cohort (blue circles; right axis). Dashed line indicates the average outcome for the set of never-treated troopers. Experience is defined as months since an officer's first citation, computed at the time of training for treated troopers and computed as of the final cohort for untreated cohorts. Age is computed as of career start.

Figure A-2: Experience profile of racial stop composition



*N*otes: This figure illustrates the estimated experience profile in racial stop composition. Using our analysis sample of traffic stops, we regress **1**[white driver] on the same fixed effects as in our main analyses, as well as controls for whether an officer has received each of the five training types, and then indicators for experience in quarters, topcoded at 17. Standard errors are clustered at the trooper-level. Figure plots the estimated coefficients and 95 percent confidence intervals from estimated coefficients on the experience indicators.

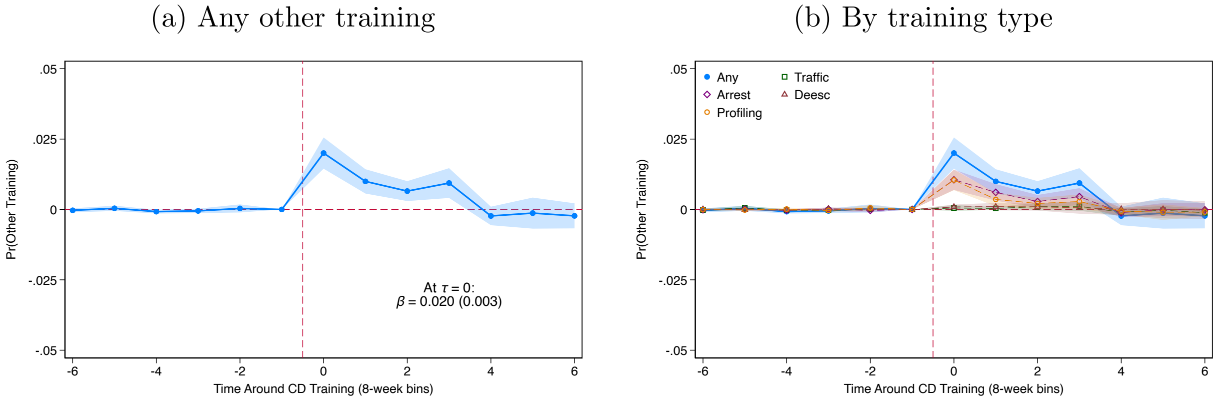
Figure A-3: Predicted outcomes based on patrol assignments


(a) Number of white stops

(b) Fraction white

*N*otes: Each panel plots event study estimates for predicted outcomes based on patrol assignments. For outcome Y (number of stops of white motorists or white share of stops), we first regress the outcome on trooper, week, and assignment fixed effects using only untreated observations, with the assignment defined as the county × weekend × shift each trooper is assigned to in a given week, computed using the approach described in appendix B-2. We then estimate event studies (using the Borusyak et al. 2022 imputation approach) where the estimated fixed effect for that officer's assignment in a given week is the outcome variable, conditioning only on trooper and week fixed effects. To account for estimation error in the initial fixed effects estimates, we bootstrap the entire procedure using a Bayesian bootstrap clustered at the trooper-level.

Figure A-4: Relationship between timing of cultural diversity training and other trainings

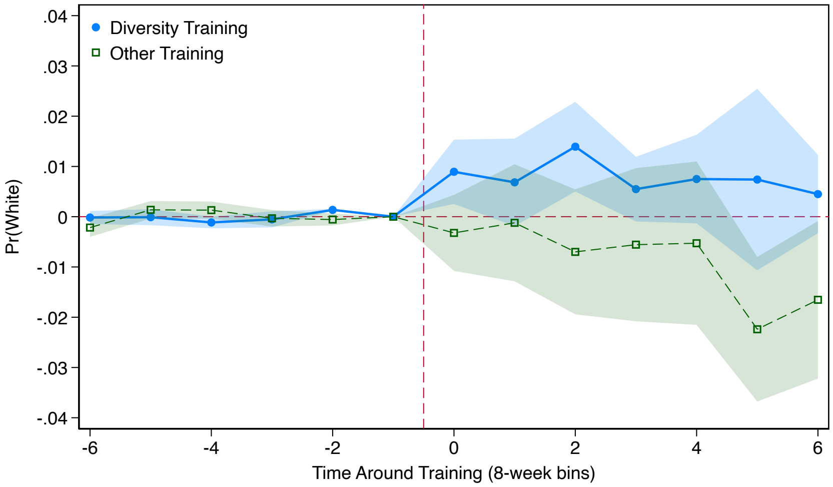


(a) Any other training

(b) By training type

Notes: Each panel plots event study estimates for predicted outcomes based on patrol assignments. For outcome Y (number of stops of white motorists or white share of stops), we first regress the outcome on trooper, week, and assignment fixed effects using only untreated observations, with the assignment defined as the county × weekend × shift each trooper is assigned to in a given week, computed using the approach described in appendix B-2. We then estimate event studies (using the Borusyak et al. 2022 imputation approach) where the estimated fixed effect for that officer's assignment in a given week is the outcome variable, conditioning only on trooper and week fixed effects. To account for estimation error in the initial fixed effects estimates, we bootstrap the entire procedure using a Bayesian bootstrap clustered at the trooper-level.

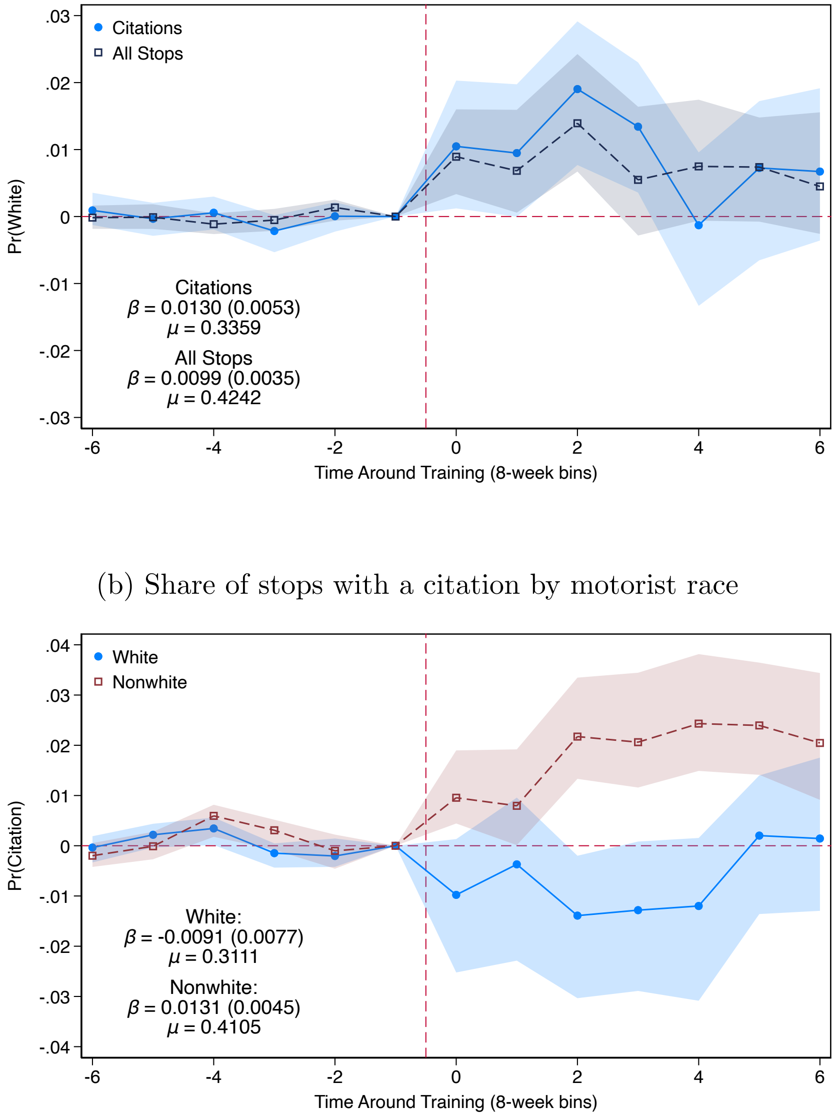
Figure A-5: Effect of training on stop composition, by training type



*N*otes: This figure plots event study estimates around the time of training for cultural diversity training (same as our main estimates) and all other trainings.

Figure A-6: Fraction of stops resulting in a citation (versus formal warning)

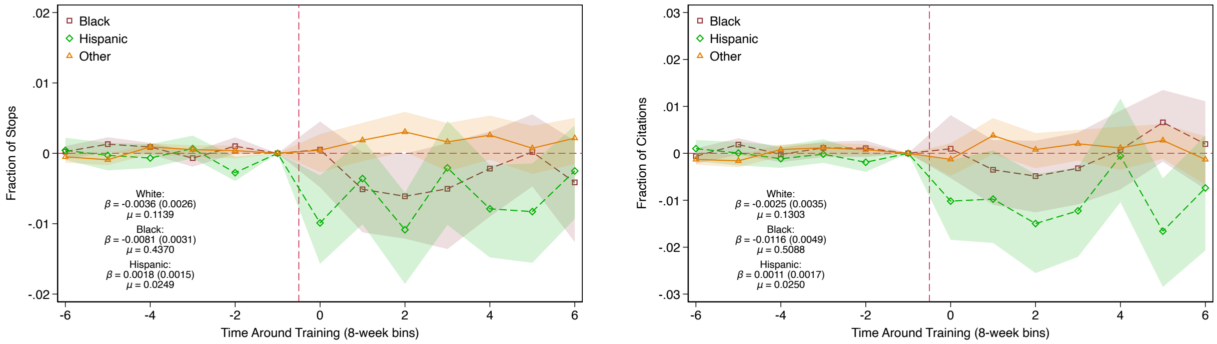
(a) Racial composition of citations versus stops



(b) Share of stops with a citation by motorist race

Notes: Panel (a) reproduces figure 1 and then also presents identical estimates using all traffic stops in our data instead of all stops with a citation. Panel (b) replicates panel (b) of figure 3.

Figure A-7: Racial composition of stops around training by detailed race

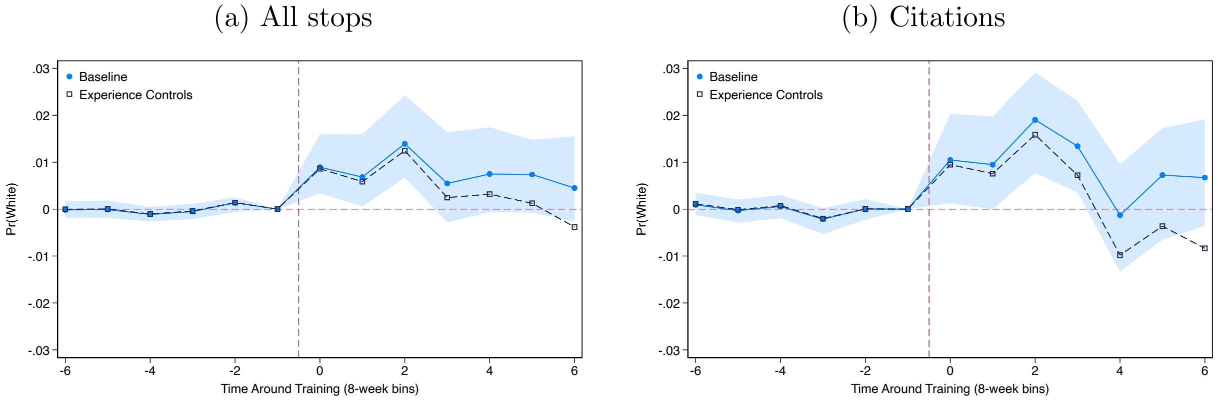


(a) All stops

(b) Citations

*N*otes: Same as figure 1 except that the outcomes are indicators for whether a stopped motorist is Black, Hispanic, or any other race/ethnicity. Panel (a) uses all traffic stops, while panel (b) uses all traffic stops with a citation.

Figure A-8: Robustness: experience controls

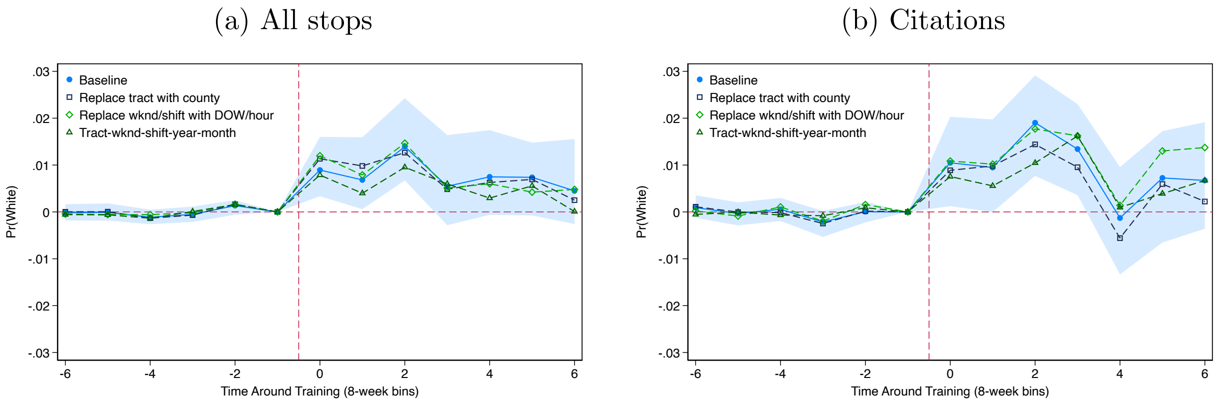


(a) All stops

(b) Citations

*N*otes: Same as figure 1 for all stops (panel a) and all citations (panel b) and adding a quartic in officer experience as a control.

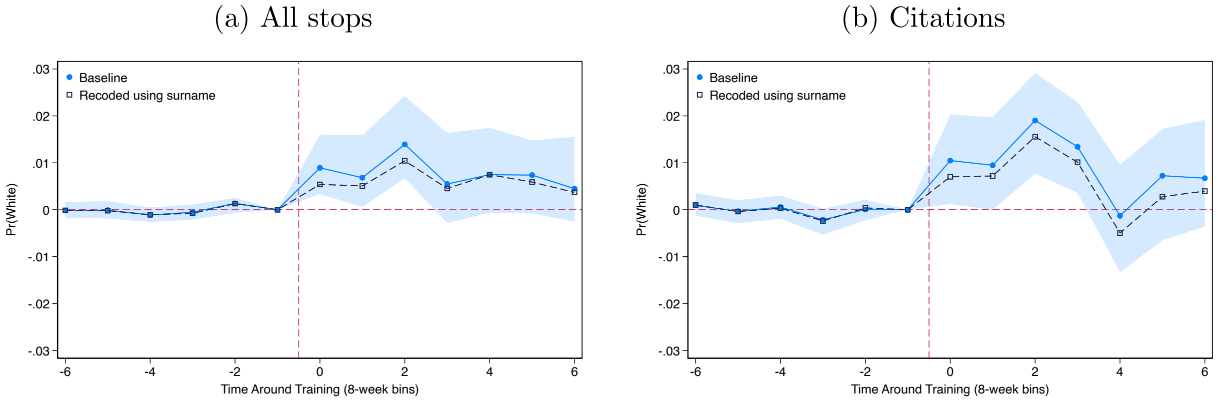
Figure A-9: Robustness: fixed effects


(a) All stops

(b) Citations

*N*otes: Same as figure 1 for all stops (panel a) and all citations (panel b) but varying the fixed effects. Baseline fixed effects are officer, date, tract × **1**[weekend] × shift, and county × year × month. The second specification replaces tract with county. The third specification returns to tracts but replaces weekend and shift with day of week and hour of day. The fourth combines the assignment and time effects together by including tract × **1**[weekend] × shift × year × month fixed effects.

Figure A-10: Robustness: recoding Hispanic status

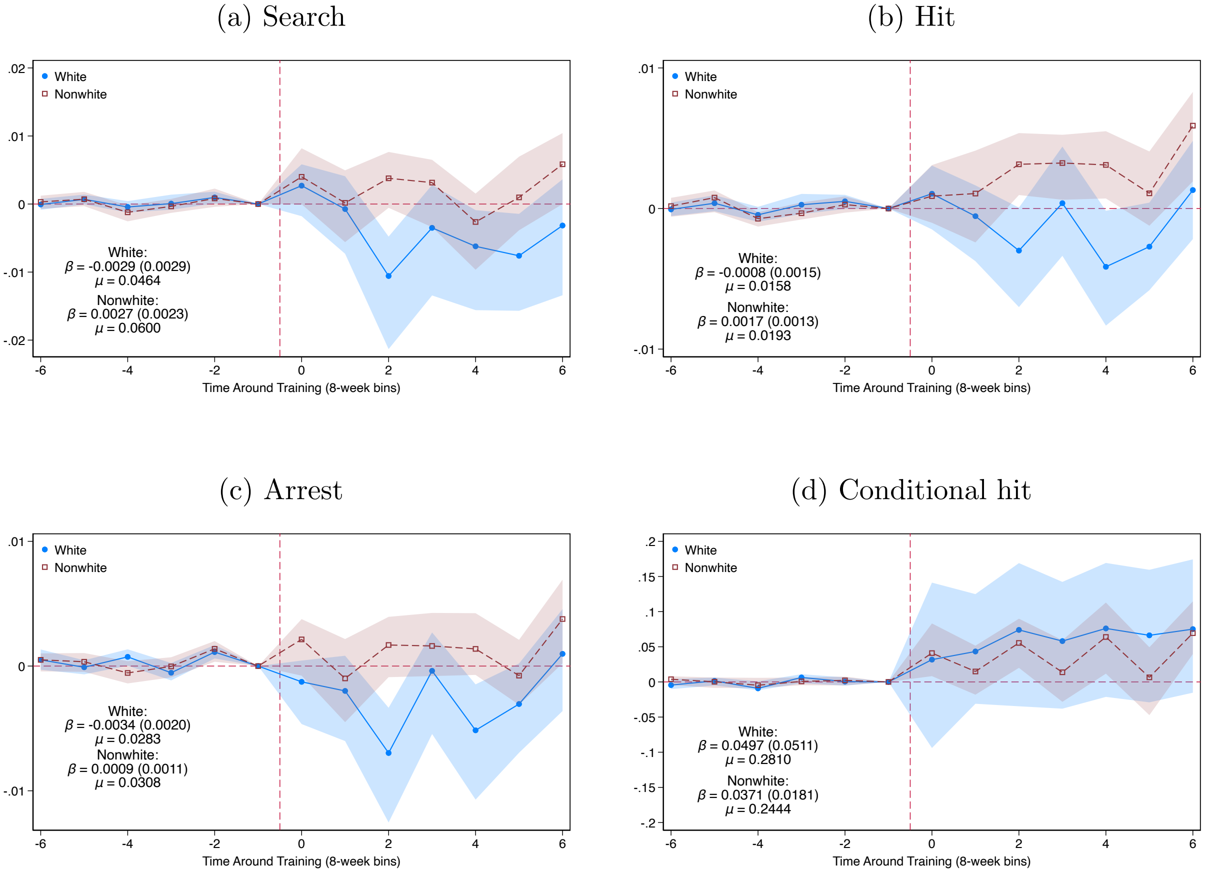


(a) All stops

(b) Citations

*N*otes: Same as figure 1 for all stops (panel a) and all citations (panel b) but recoding individuals as Hispanic if their surname is associated with an 80 percent (or greater) likelihood of Hispanic status in the Census, following Goncalves and Mello (2021).

Figure A-11: Post-stop outcomes by motorist race


(a) Search

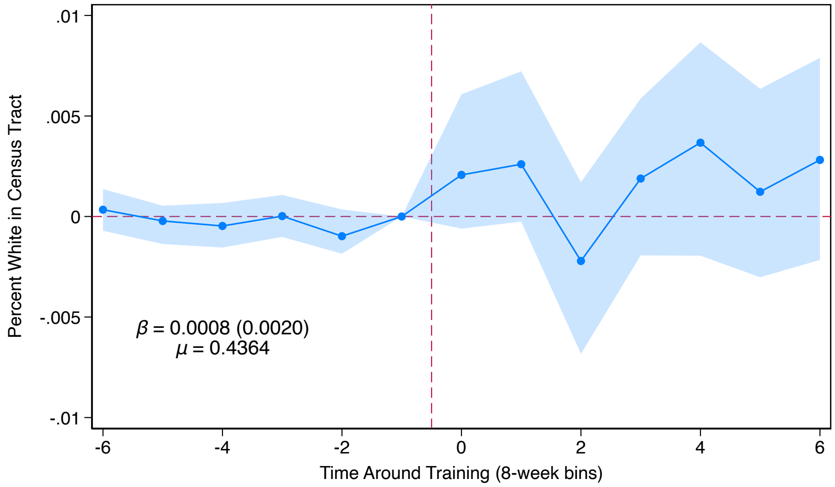
(b) Hit

(c) Arrest

(d) Conditional hit

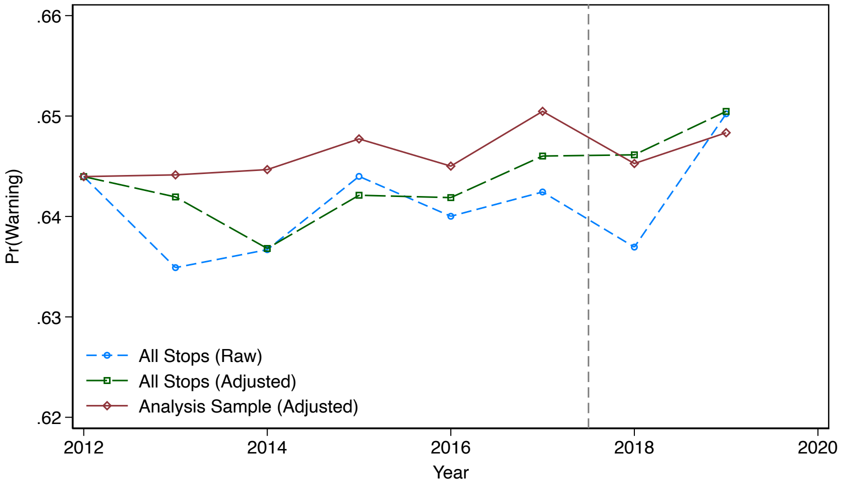
*N*otes: Same as figure 4 except using all traffic stops instead of citations.

Figure A-12: Patrol locations around training (within assignments)



*N*otes: This figure plots event study estimates around the time of cultural diversity training estimated at the traffic-stop level where the outcome of interest is share of the population that is white in the census tract where the stop is made, constructed by geocoding traffic stops to census tracts. Estimates condition on county × weekend × shift and county × year × month fixed effects (as well as trooper and week fixed effects).

Figure A-13: Share of stops resulting in a formal warning over time



*N*otes: This figure plots the fraction of stops in our administrative stop data the result in a formal warning (as opposed to a citation) over time. Blue circles indicate annual averages using all traffic stops. Green squares indicate annual averages using all traffic stops, adjusted for census tract × shift × weekend fixed effects and month fixed effects. Red diamonds indicate annual averages using only untreated observations from our analysis sample, again adjusting for for census tract × shift × weekend fixed effects and month fixed effects. Vertical dashed line indicates the timing of the law change prohibiting the use of informal warnings.

B Data appendix

B-1 Data sources

We rely on four datasets obtained via public information requests to the Texas Department of Public Safety (DPS) and the Texas Commission on Law Enforcement (TCOLE). Each of these datasets contains varying amounts of information about the officer. A brief description of each dataset and the associated officer information includes:

- DPS Traffic Stop Data: These data contain detailed information pertaining to traffic stops made by the Texas Highway Patrol from 2006-19. Each traffic stop is associated with a unique badge number. In addition to the badge number, the 2009-15 data also contain the officer's first initial and full last name from 2009-15 In the 2016-19 data, we observe the officer's full first/last name and middle initial.
- DPS Demographics: These data contain each officer's badge number, full first/middle/last name, demographics (race/ethnicity, sex, age), and hire date. The sample includes only officers employed by DPS as of April 2019.
- Comptroller Demographics: These data are organized into job position spells and contain each officer's full first/last name, demographics (race/ethnicity, sex, age), hire date, and termination date. The sample includes any officer employed by DPS from January 2006 to April 2019.
- TCOLE Training Rosters: These data contain historical course-level rosters for anyone employed by DPS from January 2006 to December 2019. These data contain a unique officer id (not linkable to other datasets) and each officer's full first/last name and middle initial.

In order to create an analytical sample consisting of traffic stops linked to individual officers who are characterized by their demographics and their prior training history, we sequentially merge each of these datasets.

The matching procedure occurs in the following sequence:

1. Based on the last date we observe a given officer in the DPS Traffic Stop data, we are able to associate a badge number with 4,982 officers with either a full first/last name and middle initial (82.41%) or a first initial and full last name (3.34%).
2. DPS Demographic data were merged to the DPS Traffic Stop data based on badge number. We match 3,154 (54.94%) of officers in the full sample. For the matched officers, we now have complete name information including a full middle name as well as associated demographic information.
3. For the 2,587 (45.06%) of officers in the DPS Traffic Stop data but not in the DPS Demographic data, we link to the DPS Comptroller data based on the officers' name.

We match 1,769 (68.38%) of officers whom we already had full or partial name information from the DPS Traffic Stop data and an additional 20 (0.7%) of officers whom we had no information from the DPS Traffic Stop data.[12]

4. Next, we match our data to the TCOLE Training Rosters based on the best available information on an officer's full first/last name and middle initial. After steps 1-3, we were left with 4,982 officers where we have such information from either the DPS Traffic Stop or Demographic files. Of these officers, we can match 4,626 (92.85%) to the TCOLE Training Roster data.[13]

At the end of the matching procedure, we can obtain demographic information (from either the DPS Demographic or Comptroller Demographic datasets) for 4,857 (97.49%) of the total 4,982 officers who made at least one traffic stop after 2009 and have name information available. We arrive at our final analytical sample by excluding 2,471 officers who make their first traffic stop in 2009 or later. We drop an additional 510 officers who were employed by DPS prior to 2009. Next, we drop an additional 61 officers that we're unable to match

[12] A total of 361 of these matches are made with a deterministic link between the first and last name. We drop illogical matches where the first traffic stop is before the hire date. We break ties using the mean squared error of the difference between the first traffic stop and hire date. An additional 929 matches were then made with a deterministic link on the last name only. As before, we first drop illogical matches where the first traffic stop is before the hire date. Next, we break ties by keeping only the potential match with the lowest Levenshtein distance between the first name in both datasets. Finally, we break any remaining ties using the mean squared error of the difference between the first traffic stop and hire date. The remaining 499 matches were made based on a fuzzy match to the last name where we only keep matches with a similarity score (based on relative Levenshtein distance) if 80 percent or higher. First, we break ties by keeping only the potential match with the lowest Levenshtein distance between the first name in both datasets. Next, we break any remaining ties using the mean squared error of the difference between the first traffic stop and hire date. At the end of every stage, we drop any observations with more than one potential match.

[13] A total of 4,255 of these matches are made with a deterministic link between the first and last name. We drop illogical matches where the first traffic stop is before the hire date. First, we break ties conditioning on whether the middle initial matched between the datasets. Next, we break ties using the mean squared error of the difference between the first traffic stop and hire date. An additional 101 matches were then made with a deterministic link on the last name only. As before, we drop illogical matches where the first traffic stop is before the hire date. First, we break ties conditioning on whether the middle initial matched between the datasets. Next, we break ties by keeping only the potential match with the lowest Levenshtein distance between the first name in both datasets. Finally, we break any remaining ties using the mean squared error of the difference between the first traffic stop and hire date. The remaining 157 matches were made based on a fuzzy match to the last name where we only keep matches with a similarity score (based on relative Levenshtein distance) if 80 percent or higher. First, we break ties conditioning on whether the middle initial matched between the datasets. Next, we break ties by keeping only the potential match with the lowest Levenshtein distance between the first name in both datasets. Finally, we break remaining ties using the mean squared error of the difference between the first traffic stop and hire date. At the end of every stage, we drop any observations with more than one potential match.

to the TCOLE Training Rosters. We also drop an additional 216 officers who we observe taking TCOLE courses prior to 2009, typically because of employment as a municipal officer, corrections officer, or 911 dispatcher. Our final sample of 1,724 officers has complete coverage in terms of demographic data.

B-2 Patrol assignments for panel data approach

While we can directly control for a variety of stop characteristics when estimating our main regressions at the stop-level, the unit of observation in our panel data analysis is an officer $j \times$ week t. Hence, for our panel data analyses, we estimate patrol assignments for each officer × week using the observed distribution of an officer's stops in that week. Specifically, for each $j \times t$, we compute the share of stops made by time of day, made on weekends or weekdays, and made across geographic locations.

For time of day, we use a simple partition of the day that accords well with a typical policing schedule: 6AM–2PM; 2PM–10PM; 10PM–6AM. In 20 percent of all officer-weeks, all stops are made in one of these partitions, while stops are made in all three times of day in 27 percent of officer-weeks. We assign each officer-week to the time of day in which they make the majority of their stops. 68 percent of stops are made in the time of day partition to which we assign officers.

For weekends, we compute the share of stops made during weekends as opposed to weekdays, and then code that officer × week as a weekday or weekend officer if more than 40 percent of their stops in that week were made on weekends. In the stops data, 67 percent of all stops occurring on weekends are made by officers that we designate as weekend officers.

We combine the time-of-day and day-of-week assignments into a single assignment measure, which we call the "shift." Shifts can take six values (weekend v. weekday × three times of day). For example, in a given week, one officer will be coded as working overnight during weekdays, while another will be codes as working the morning shift on weekends.

We perform the same exercise for geographic locations. In about 65 percent of officer-weeks, all stops are made in a single county, while in 91 percent of officer-weeks, all stops are made in one or two counties. In the stops data, 91.5 percent of all stops are made in the county in which that officer is assigned to for that week.

B-3 Additional institutional details

Cultural diversity requirements: House Bill 2881 amends peace officer continuing education requirements in September 2001 so cultural diversity is to be taken once every 48 months (link). House Bill 3389 amends 1701.352 to require officers holding only a basic proficiency certificate, to complete cultural diversity training as part of the continuing education requirements for the 2009-2013 training cycle and amends 1701.402 to require completion of cultural diversity for an intermediate certificate (link). The relevant four-year training cycles during our study period are 09/01/2005–08/31/2009; 09/01/2009–08/31/2013; 09/01/2013–08/31/2017; and 09/01/2017–08/31/2021. Officers who begin their careers more than two years into an ongoing training cycle are given until the end of the following training cycle to complete the required in-service trainings.

Reaching intermediate proficiency: Service time and in-service training requirements for reaching intermediate proficiency depend on an officer's educational background and prior military service. The majority of peace officers employed by the Texas Highway Patrol have either a bachelor's degree or four years of military service. Officers with an associates degree or two years of military service are required to complete four years of THP service to reach intermediate proficiency. Officers without a college degree or military service are required to complete either two years of THP service and 2,400 hours of in-service training, four years of THP service and 1,200 hours of in-service training, six years of service and 800 hours of in-service training, or eight years of service and 400 hours of in-service training.

All troopers must complete seventeen required in-service training courses to achieve an intermediate proficiency certificate. These courses are: Child Abuse Prevention and Investigation; Crime Scene Investigation; Use of Force; Arrest, Search and Seizure; Spanish for Law Enforcement; Identity Theft; Asset Forfeiture; Racial Profiling; Human Trafficking; Crisis Intervention Training; Interacting with Drivers whe are Deaf/Hard of Hearing; De-escalation Techniques; Missing and Exploited Children; Child Safety Check Alert List; Canine Encounters; Cultural Diversity; and Special Investigative Topics.

C Technical appendix

C-1 Imputation estimator from Borusyak et al. (2022)

Consider a standard panel data setup with units indexed by i and time indexed by t. Each unit receives treatment at time g_i, with $g_i = \infty$ for never-treated units. Let $D_{it} = \mathbf{1}[t \geq D_{it}]$ denote whether a unit has been treated as of time t. The imputation estimator proceeds in two steps. First, the outcome is regressed on controls X, unit fixed effects α, and time fixed effects δ using only untreated observations ($D_{it} = 0$):

$$Y_{it} = \gamma X_{it} + \alpha_i + \delta_t + u_{it}$$

Estimated coefficients from this regression are then used to construct estimates of untreated potential outcomes for each unit × time:

$$\hat{Y}(0)_{it} = \hat{\gamma} X_{it} + \hat{\alpha}_i + \hat{\delta}_t$$

Event study estimates are then constructed for each $\tau = t - g$ by averaging the difference between the observed and predicted outcomes at each event time τ:[14]

$$\hat{\theta}_\tau = E(\tilde{Y}_{it}|\tau) = E(Y_{it} - \hat{Y}(0)_{it}|\tau)$$

From our perspective, this solution to the well-documented issues with canonical TWFE estimation of event studies is particular appealing for several reasons. First, this approach does not require a conventional panel data structure, as we have in our analyses of the data at the traffic-stops level. Second, this approach easily accommodates a more complex fixed effects structure than simple TWFE, which is necessary in our setting to address the fact that troopers patrol, for example, diverse geographic areas.

Our analysis based on a panel dataset at the officer × week level, used primarily to examine the impact of training on the number of traffic stops, closely mirrors the standard panel data setting with TWFE with two exceptions. The first is that we typically also condition on assignment fixed effects. In most specifications, we include trooper and week fixed effects in addition to a detailed assignment fixed effects (county × shift), as well as more aggregated time effects that are allowed to vary by geography (county × year × month). The second is that we aggregate our event-time estimates at a level higher than the time dimension of the panel. In other words, while our panel data are weekly, we report event time coefficient for 8-week groups instead of for individual weeks, primarily to increase precision.

[14] Note that while Borusyak et al. (2022) and Gardner (2021) propose identical imputation-based estimates for event study coefficients in the post-treatment period ($\tau \geq 0$), Borusyak et al. (2022) advocate a regression-based approach to computing the pre-treatment coefficients, whereas Gardner (2021) suggests the same procedure for computing pre- and post-treatment estimates. We use the Gardner (2021) approach to compute the pre-treatment coefficients but report the pretrend diagnostic test suggested by Borusyak et al. (2022). This pretrends test entails regressing the outcome on a set of pre-treatment event time indicators s using only not-yet-treated observations and then conducting a joint significance test of the event time indicators.

In practice, our approach is identical to that described above except that in the second stage, we take averages over 8 week bins instead of for each individual week relative to treatment.

Our analysis based on data at the traffic-stop level further diverges from the canonical TWFE setup, but the imputation procedure accommodates this data structure as well. We can still estimate a "first stage" using only not-yet-treated observations, construct predicted values from this regression, and then average differences between observed and predicted outcomes for different time periods relative to treatment. In our event study models, we drop each trooper's exact week of training, as well as all trooper-weeks after the trooper has received their second iteration of diversity training.

C-2 Inference procedure

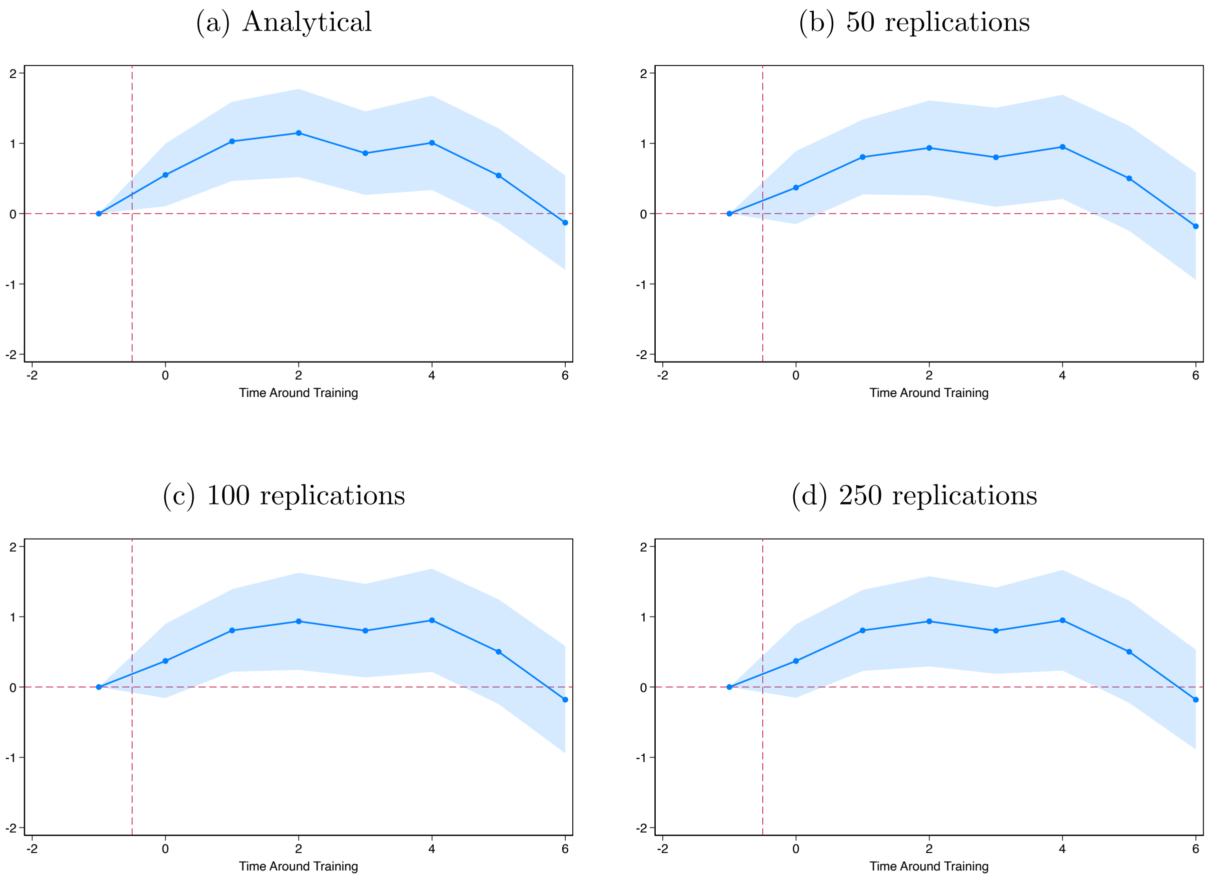
The key inference challenge associated with the imputation estimator is that the residuals $\tilde{Y}_{it} = Y_{it} - \hat{Y}(0)_{it}$, which are averaged in the second step to estimate parameters of interest, are constructed from regression estimates in the first step. Hence, the standard errors of the conditional averages $E(\tilde{Y}_{it}|\tau)$ will be biased downwards because first-stage estimation error is unaccounted for. Both Borusyak et al. (2022) and Gardner (2021) derive analytical standard errors for the imputation event study estimator. However, these analytical standard errors are too computationally intensive for our setting, particularly when estimating regressions at the traffic stop-level, due to both the large N and the large number of treated cohorts (i.e., many different treatment timings).

Instead, we compute standard errors using the Bayesian bootstrap of Rubin (1981), clustering at the trooper-level. The Bayesian bootstrap approach is identical to a classical bootstrap approach except that, instead of random resampling with replacement, random weights are drawn and then applied in each iteration.[15] An important advantage of the Bayesian bootstrap approach is that it preserves the support of all relevant fixed effects in each bootstrap iteration.

Specifically, we draw random Dirichlet weights for each officer in each bootstrap replication, estimate event study parameters weighting by those weights (where the weights are applied in both the first and second stages of the imputation estimator). We then compute the standard deviation, 5th percentile, and 95th percentile of the bootstrapped parameter estimates as our estimates of the standard error and lower and upper 95 percent confidence bounds. Throughout, we use 100 bootstrap iterations for inference. Figure C-1 below illustrates that our Bayesian bootstrap procedure generates similar to confidence intervals to the analytical confidence intervals from Borusyak et al. (2022) when estimating using panel data (analytical standard errors for regressions using stops data could not be calculated due to computing constraints, as mentioned above).

[15] One can think of the standard bootstrap as a special case of the Bayesian bootstrap, where the weights are integers. See, e.g., twitter thread from Peter Hull, January 2022.

Figure C-1: Confidence intervals


(a) Analytical

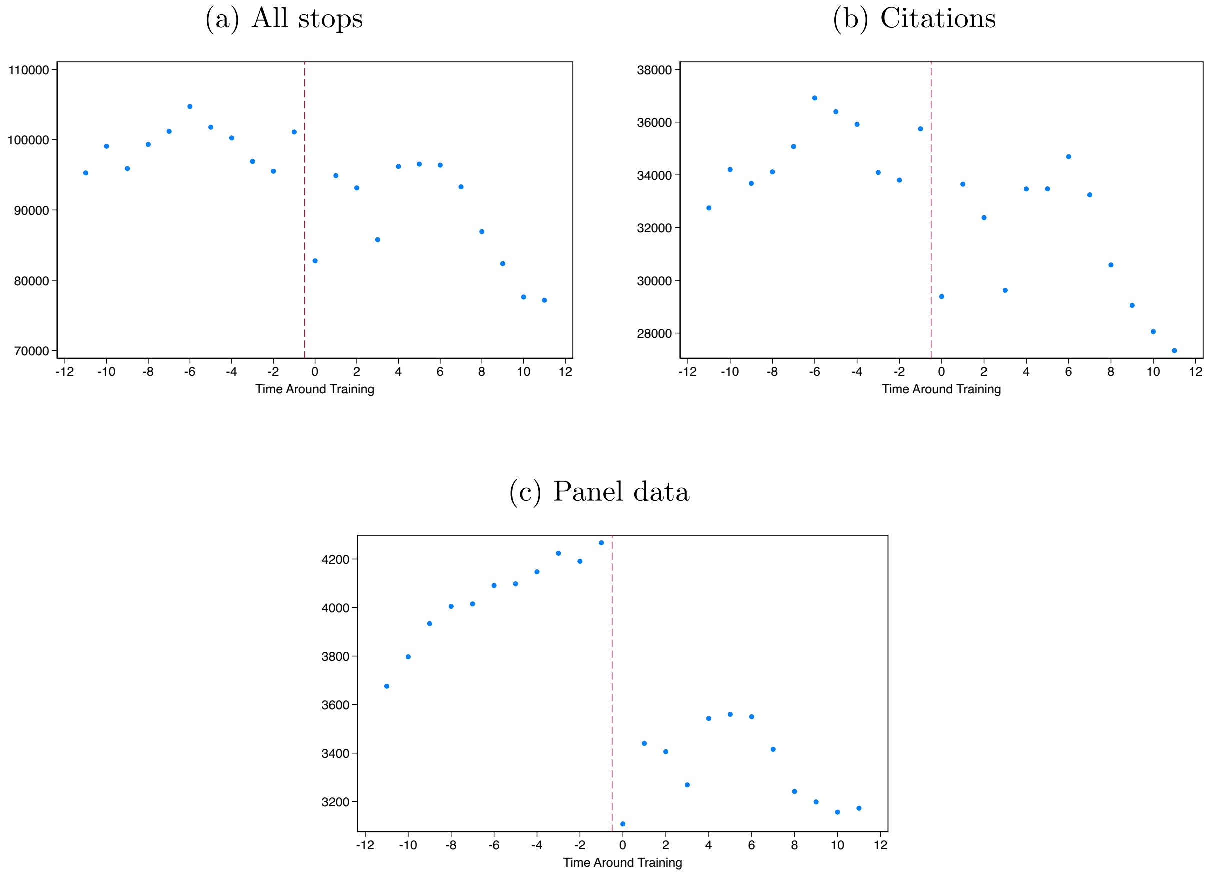
(b) 50 replications

(c) 100 replications

(d) 250 replications

*N*otes: Each panel plots event study estimates and 95 percent confidence bands from imputation event study estimates based on panel data where the outcome is number of stops of white motorists and the conditioning fixed effects are trooper, week, county × year × month and shift × year × month. In panel (a), confidence bands are computed using the analytical standard errors from Borusyak et al. (2022). In panels (b)-(d), confidence bands are computed using a Bayesian clustered bootstrap, varying the number of bootstrap replications. Note that slight differences between the point estimates in panel (a) and panels (b)-(d) are due to the different methods of aggregation. The STATA command accompanying Borusyak et al. (2022) only allows for computation of event study parameters at the level of the panel (in this case, weeks). Hence, to use their package to construct comparable estimates, we compute estimates for each week after training and then average the weekly estimates into 8-week bins, computing standard errors using the postestimation `lincom` command.

Figure C-2: Effective number of observations for event study estimates



(a) All stops

(b) Citations

(c) Panel data

*N*otes: This figure illustrates the number of observations used to compute event study estimates for each period relative to training (8-week bins) when estimating regressions at the traffic stop-level (panel a), when restricting to stops with a citation (panel b), and when using the officer × week panel data approach (panel c). Note that the excess drop in the effective number of observations at $\tau = 0$ is due to the fact that we drop the exact week of training; hence, period zero is based on one fewer week of data than other periods.

D *Veil of darkness* test

To benchmark our estimated effects of cultural diversity training on stop behavior, we use the so-called *veil of darkness* test of Grogger and Ridgeway (2006). This test compares the racial composition of stops made during day and night hours, with the key premise being that motorist race is observable *prior* to the stop during the day but not in darkness. Hence, a decline in the minority share of stops during darkness suggests that excess stops of minorities are being made during daylight hours due to racial profiling.

Although this test has been criticized by Horrace and Rohlin (2016), who argue that day versus night only crudely captures the observability of race to officers, particularly in urban environments with streetlights, and by Ross et al. (2023), who argue that minorities may endogeneously change their driving behavior in response to the perceived risk of racial profiling, we nonetheless argue that the veil of darkness test represents a straightforward means of benchmarking our estimated magnitudes.

In terms of operationalizing the veil of darkness test in our setting, we follow the procedure of Ross et al. (2023) and focus on the "intertwilight" window, or the period of the day when the sunset varies throughout the year. We also use only the not-yet-treated subset of our stops data to avoid contamination due to treatment effects from training. Using this subset of the data, we regress an indicator for whether a motorist is white on an indicator for daylight (as opposed to dark), conditioning on hour fixed effects. A negative coefficient indicates that white motorists comprise a lower share of stops during daylight than during darkness, hence suggesting racial profiling bias against minorities.

Table D-1 below presents our veil of darkness estimates. Column 1 presents estimates conditioning only on hour and day of week, which each column progressively adding additional fixed effects. Focusing on column (4), which includes census tract and trooper fixed effects, we estimate that a stopped motorist is two percentage points less likely to be white during daylight than during darkness.

Table D-1: Veil of darkness estimates

	(1) White	(2) White	(3) White	(4) White
Daylight	-0.0530	-0.0374	-0.0367	-0.0211
	(0.00313)	(0.00193)	(0.00184)	(0.00150)
Mean	0.483	0.483	0.483	0.483
Hour FE	Yes	Yes	Yes	Yes
DOW FE	Yes	Yes	Yes	Yes
County FE	No	Yes	No	No
Tract FE	No	No	Yes	Yes
Officer FE	No	No	No	Yes
Officers	1476	1476	1476	1476
Observations	875051	875050	873419	873419