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

Steven Mello Matthew B. Ross Stephen L. Ross Yuling Han

July 10, 2025

Abstract

We study the effects of cultural diversity training on employee behavior in the context of policing. Leveraging variation in the timing of a mandated diversity training, we find that Texas highway patrol officers respond to training by adjusting their racial composition of stops. The probability that a stopped motorist is white increases by 1.3 percentage points over the two years after training, and officers achieve this by stopping additional white motorists rather than reducing their stops of minorities. We find evidence that officers stop less "guilty" white motorists after training, suggesting that the training reduces troopers' lenience towards white drivers.

JEL Codes: M53, J15, K42

[*] Mello: Dartmouth College and NBER. Ross, M: Northeastern University. Ross, S: University of Connecticut and NBER. Han, Northeastern University. We are grateful to Hunter Johnson who served as research assistant and collaborator on an earlier version of this paper. We also thank Leah Boustan, Felipe Goncalves, Erzo Luttmer, Emily Owens, Emma Rackstraw, Doug Staiger, and Cody Tuttle, as well as seminar participants at the Online Economics of Crime Seminar, the Western Economics Association, and Montana State University for helpful comments. M. Ross thanks the Russell Sage Foundation (G-2011-30016) for financial support. Jackyln Pi and Zeyu Wang provided excellent research assistance. 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 found positive effects on employee attitudes towards various groups (e.g., Chang et al. 2019; Bezrukova et al. 2012), others have documented a "backlash" effect, with attitudes worsening 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.

The emerging literature on this topic has also overlooked the possibility that the effects of diversity training may depend on the mechanisms underlying disparate behavior towards different groups. If disparities arise due to racial animus, as posited by Becker (1957), effective training would be expected to foster more favorable treatment of minority groups. On the other hand, if disparities are driven by stereotyping behavior (Bordalo et al., 2016), with individuals over-estimating the distribution of positive characteristics among the majority group, diversity training might be expected to produce less favorable treatment for that majority group, rather than better outcomes for minority groups.[2]

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

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

[2] Notably, Hull (2021) provides several examples where empirical tests that reject accurate statistical discrimination are also inconsistent with prejudiced-based discrimination. Hull (2021) shows that the marginal treatment effect should be increasing in the likelihood of treatment when agents practice either accurate statistical or prejudiced based discrimination, suggesting that racial differences in bail release documented by Arnold et al. (2018) are due to inaccurate beliefs as opposed to prejudice. Similarly, Feigenberg and Miller (2020) provide evidence that police searches appear to be random conditional on observables with minority drivers searched more, and Hull (2021) notes such behavior implies a flat relationship between treatment intensity and marginal effects, again inconsistent with prejudice based discrimination.

on officers' 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 educators (e.g., Tumen et al. 2022).

After completing the police academy, Texas Highway Patrol (THP) 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 the 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 officers using an event study approach, leveraging the staggered timing of training across 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). An advantage of this approach is that it easily accommodates a more complicated fixed-effects structure than two-way fixed effects. Throughout our analyses, we condition on officer effects, as well as time effects that vary by an officer's assigned patrol district (to address the fact officers patrol diverse regions) and flexible controls for officer experience (because we are focused on early-career officers).

Our identifying assumption is that absent training, a given officer's enforcement behavior would have followed the same trend as similarly experienced officers patrolling the same areas but who receive training at a later date. While all officers in our sample are eventually required to receive diversity training, the timing variation we exploit in practice is "natural" variation driven to some extent by officer choices of when to take up training. However, we note that the timing and location of course offerings are at the discretion of a decentralized system of service providers and we show that trooper characteristics do not systematically differ by treatment timing, that trooper patrol assignments do not change systematically around the timing of training, that officer behavior does not change systematically in the weeks leading up to diversity training, and that results are robust to controlling for when officers complete other mandated trainings, all of which increase confidence in the validity of our empirical approach.

In the weeks following diversity training, officers respond by adjusting the racial and ethnic composition of their traffic stops. We find that an officer's share of stops that are of white motorists increases by 1.3 percentage points ($se = 0.007$), or about three percent relative to the mean. This increase emerges immediately following training and attenuates in the medium term, but appears generally persistent over the two years following training. The change in stop composition is driven entirely by an increase in the number of stops of white motorists, rather than a decline in stops of minority drivers. In particular, we find that training is associated with 0.9 ($se = 0.298$) additional stops of white motorists per week. On the other hand, we find essentially no evidence of changes in either the number or composition of stops of minority motorists after training. Our baseline findings are robust to a variety of alternate specifications, including those which flexibly control for exposure to other contemporaneous in-service trainings taken by officers.[3]

To benchmark the magnitude of our estimates, we compare the treatment effect of training on the racial composition of stops with results from the "veil of darkness" (VOD) test (Grogger and Ridgeway, 2006). This common methodology for detecting police discrimination in the literature compares the racial composition of stops during daylight and darkness, leveraging the principle that race should be unobservable to officers prior to a stop in the latter condition. Our estimate of the impact of training on the white share of stops is about 65 percent as large as the extent of discrimination we estimate when applying the VOD test to pre-treatment stops in our data. In other words, our estimates suggest that diversity training erodes a significant share of the estimated level of discrimination in police stops. Consistent with this view, we also cannot reject the null of no discrimination when applying the same VOD test to stops made by officers immediately following diversity training.

Given our central finding that officers respond to training by increasing their stops of white motorists, we explore the hypothesis that training induces officers to reduce their lenience towards white motorists on the stopping margin. Consistent with this view, we find that the impacts of training are driven by the responses of officers who stop a lower share of white motorists pre-training and by the responses of non-white officers, who we posit may have been more deferential to white motorists prior to training. We find that the increase in stops of white motorists is largely attributable to additional stops for speeding infractions and that, following training, the average speeds of white motorists stopped for speeding declines. In other words, training appears to induce additional stops of less "guilty" whites whom officers were letting pass prior to training. We also find a slight increase in pretextual stops of white motorists following training, potentially implying an overall heightened suspicion of white drivers more generally and suggestive of the the view that pre-training disparities are explained by stereotyping behavior on the part of officers (Bordalo et al., 2016).

Overall, our findings clearly suggest the potential for cultural diversity training to change employee behavior but also come with several caveats. First, the evidence we present suggests that officers respond to training by reducing their lenience towards white motorists, rather

[3] Perhaps unsurprisingly, we find that adjusting our estimates for exposure to a racial profiling training, which is distinct from the cultural diversity training we study in terms of focus, slightly attenuates our estimates.

than changing their behavior towards minorities. In many settings, this outcome may not align with the desired objectives of a diversity training initiative. Modern policy proposals in policing, for example, often advocate for reducing the number of interactions between minorities and officers given the potential risks of escalation (Woods, 2021).

Second, the features of our setting limit our ability to draw strong conclusions about the longer-term impacts of diversity training, the effectiveness or repeated versus one-off trainings, or the impacts of training in environments with minimal other on-the-job trainings. Although Texas patrol officers undergo recurring mandatory training, our empirical framework and institutional setting preclude us from assessing the effectiveness of repeated diversity trainings, often required in employment settings. While we find that results are not substantively changed by controlling for exposure to simultaneous training, we find some evidence that racial profiling training also impacts officer behavior and we cannot rule out interaction effects of the various trainings taken during an officer's career.

In addition to providing novel evidence on the effects of cultural diversity training, our paper contributes to a broad literature in economics on worker training programs. 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 police training (Ipsos, 2021). Nonetheless, evidence of the effects of police training on enforcement behavior is both limited and mixed. Randomized control trials have yielded some evidence that procedural justice training trainings 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 imporance of field-training on officer outcomes. We document the effects of a training aimed specifically at officer behavior towards minority groups.

In a similar 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 of 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). A second exception is the emerging literature on legislative and judicial interventions, including federal consent decrees (e.g., Fagan and Geller 2020; Long 2019; MacDonald and Braga 2019; Campbell 2023; Devi and Fryer 2021; Heaton 2010; Naddeo and Pulvino 2024; Matsuzawa 2024) as well as targeted department-level cooperative interventions (e.g., Parker et al. 2024). 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, recruits to the 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 which is taken during the third phase (in-service training).

All licensed peace officers in Texas, including THP officers, have differing semi-annual in-service training requirements based on their "proficiency" status. After completing academy training, officers are granted a basic proficiency certificate and are then required to work towards achieving intermediate, advanced, and masters proficiency certificates.[4] While the specific incentives vary across agencies, peace officers throughout Texas generally receive pay increases for achieving higher proficiency levels. To advance from basic to intermediate proficiency, THP officers must reach a minimum of two years of service and complete a set of 17 courses on a variety of topics including cultural diversity.[5]

[4] THP officers are assigned to 24 weeks of field training after the academy where they are supervised by a senior field training officers. In the majority of our estimates, we drop this period of supervised patrol.

[5] The majority of peace officers employed by THP have either a bachelor's degree or four years of military service. Officers with an associate's degree or two years of military service are required to complete four years rather than two years of service. Officers without a college degree or military service are required to complete either two years of service and 2,400 credit hours of in-service training, four years of service and 1,200 credit hours of in-service training, six years of service and 800 credit hours of in-service training, or eight years of service and 400 credit hours of in-service training. The seventeen required courses to advance to intermediate proficiency include: Child Abuse Prevention and Investigation; Crime Scene Investigation; Use of Force; Arrest,

There are four courses (cultural diversity, crisis intervention, de-escalation, special investigation topics) which are required to be taken once in each four-year training cycle until a officer achieves intermediate proficiency.[6] The focus of our analysis is an officer's first cultural diversity training, taken prior to reaching immediate proficiency. Depending on exact unit assignments, most THP officers are required to take several other in-service training courses. We discuss the potential complications associated with identifying effects of diversity training while other trainings are also occuring below in section 3.

Note that the training received by THP officers is broadly representative of police training throughout the country. According to a 2018 survey of 681 state and local law enforcement agencies, police recruits across the United States complete an average of 833 hours of basic academy training and 508 hours of field training (Bureau of Justice Statistics, 2018). Among the 14,731 of 15,323 policing agencies in the U.S. that require in-service training, officers complete an average of 39 hours of in-service training annually (Bureau of Justice Statistics, 2020). Relative to these national averages, officers with the THP take fewer hours of field training (31 percent less) and more hours of traditional classroom or simulation-based training (26 percent more basic academy training and 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 training.

Cultural diversity training for law enforcement officers is largely geared towards making officers more effective at their job, a focus that may differ from similarly-named public or private sector trainings aimed primarily at changing workplace culture. While cultural diversity training in law enforcement has roots dating back to the 1960s, its modern incarnation emerged during the 1990s (Hennessy, 2001). The stated goal of this contemporary form of cultural diversity training is to equip officers with a better understanding of their evolving communities and to enhance their ability to communicate more effectively with populations from diverse backgrounds, which is viewed as crucial for building trust, ensuring fairness,

Search and Seizure; Spanish for Law Enforcement; Identity Theft; Asset Forfeiture; Racial Profiling; Human Trafficking; Crisis Intervention Training; Interacting with Drivers Deaf/Hard of Hearing; De-escalation Techniques; Missing and Exploited Children; Child Safety Check Alert List; Canine Encounters; Cultural Diversity; and Special Investigative Topics.

[6] 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 09/01/2009–08/31/2013; 09/01/2013–08/31/2017; and 09/01/2017–08/31/2021.

and fostering effective communication within the communities they serve.

Rather than overtly aiming to mitigate implicit or explicit bias, the training pragmatically emphasizes the critical need for officers to understand and be sensitive to cultural differences among the populations they serve. The course is presented to officers as providing a crucial set of skills designed to foster community trust, enhance de-escalation in volatile situations, and ensure equitable and dignified treatment for all individuals. This approach prioritizes practical application over theoretical explorations of bias, focusing on observable behaviors and interactions. The training specifically aims to enhance officers' self-awareness and interpersonal skills.

In Texas, cultural diversity training, as overseen by the Texas Commission on Law Enforcement (TCOLE), is structured into two sequential four-hour blocks, for a total of eight hours of instruction. The initial block comprises two mandatory modules for all officers: "introduction to diversity" and "cultural diversity". The "introduction to diversity" module lays out foundational concepts, identifying key elements surrounding culture and diversity, including various salient dimensions of diversity such as race, ethnicity, gender, socio-economic status, and age. The "cultural diversity" module then delves into the nuances of various cultural backgrounds as relevant to policing in Texas, emphasizing how biases can lead to discriminatory actions if left unchecked. It also covers how recognizing and addressing bias can foster trust and confidence in the communities officers serve. The second block then broadens the scope, offering modules such as "generational diversity," "workplace diversity," "gender diversity," and "law enforcement as a diverse culture." These later modules address how diversity impacts internal police dynamics and interactions within the police workforce, rather than emphasizing only engagement with the community.

The pedagogical approach emphasizes active learning. These trainings are specifically designed to be "hands on, interactive, and scenario based," incorporating role-playing exercises that directly relate to officers' day-to-day experiences both on and off duty. Scenario-based learning allows officers to practice communication techniques and problem-solving strategies in simulated diverse encounters, improving their decision-making and conflict resolution skills in high-pressure situations. Assessment opportunities include oral or written testing, interaction with instructors and students, case studies, and scenarios. A potentially significant distinction from other similar trainings, which are often led by academics or social workers, is that this course is typically taught by a senior police officer with substantial on-the-job experience. This instructional approach, delivered by a peer with practical insights, aims to enhance the credibility and applicability of the training content for its participants. The curriculum is designed by subject matter experts who are nationally recognized and licensed instructors, many of whom testify in law enforcement-related defense cases.

All in-service training, including cultural diversity, is offered at a state-level academy in Austin and in over 50 other locations throughout the state, including many community college campuses and a few dedicated academies collocated with large, municipal agencies.[7] In our sample of early-career officers, about 60 percent take the training at the state academy in Austin.[8] Officers schedule their in-service trainings somewhat at their own discretion and with coordination with their commanding officer, typically several months in advance. Given that enrollment in courses at the main Austin academy or its satellite locations are capped, officers typically have imperfect control over the exact date and location of their in-service trainings.

2.2 Data

Our analysis relies on administrative records of all traffic stops made by the THP over the period 2010–2019, provided by the Texas Department of Public Safety (TDPS). These records include officer badge numbers and detailed stop information—such as date, time, and GPS coordinates—which we then map to counties and census tracts. The dataset also includes motorist demographics (race, age, gender) and stop outcomes, such as whether a citation was issued (and for what violation), if a search was conducted, whether contraband was found, and if an arrest was made.[9] We match officers in the THP traffic stop data with the universe of historical in-service training rosters for all peace officers certified by the Texas Commission on Law Enforcement (TCOLE), the state agency administering civil service requirements and certifications. We use these records to identify the date at which each THP officer completes various trainings, including both academy and in-service trainings. Supplementary information on officer 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 officers in our sample.

Our analysis sample is comprised of 1,662 officers who we observe starting their careers in 2010 or after and who can be matched to both the demographics and TCOLE datasets. Table 1 reports summary statistics for this analysis sample of officers. Officers are around 30 years of age at career start. The vast majority of officers are male (90 percent) and either white (52 percent) or Hispanic (38 percent). As shown in table 2, officers in our sample typically make about 25 traffic stops per week prior to training and 23 stops per week after training. Officers write approximately eight citations per week before and after training.

[7] See https://www.tcole.texas.gov/law-enforcement-academies for a map of academy locations.

[8] In our analysis sample, nearly all of the recruits participated in live in-person training. In more recent years, many of the required in-service training courses have been made available online.

[9] Luh (2022) presents evidence that THP officers systematically misreport Hispanic motorists as white in order to mask their racial profiling behavior. In our main analysis we use a definition of Hispanic status based on surnames of the stopped individual (e.g., Goncalves and Mello 2021) but note that our findings are qualitatively similar when we use officer-reported ethnicity. We find similar results regardless of whether we implement this correction or not.

3 Empirical approach

We rely on an event study approach to examine changes in officer behavior around the timing of diversity training, relative to comparable officers patrolling similar areas and with similar levels of experience but who take training at a later date. Throughout our analysis, we condition on THP region by time fixed effects to isolate comparisons only between officers with the same patrol assignments, as well as fixed effects for time employed by the THP to flexibly account for changes in patrol behavior as officers gain experience.

Our identification strategy thus relies on a parallel trends assumption – had they not faced training at a particular date, the enforcement behavior of trained officers would have trended similar to that of officers of the same experience and working in the same region who take training later. While not required for parallel trends to hold, we view variation in the timing of training, holding constant officer cohort and patrol assignment, as approximately random. As discussed above, all officers must complete diversity training before progressing to intermediate proficiency status, typically occurring after two-six years of service depending on prior educational attainment and military service. The exact timing of a particular officer's training is typically influenced by the accessibility and availability of courses, seats in which can be scarce due the uniform training requirements applied to Texas peace officers.

Consistent with this view, appendix figure C-1 demonstrates that the timing of an officer's training cannot be predicted by officer characteristics other than experience, which we control for in our main specification (and is, to some extent, mechanically related to the timing of training given the institutional environment). In our discussion of results below, we present additional evidence suggesting the quasi-randomness of the timing of training. Specifically, we show that the timing of training cannot be predicted by changes in enforcement behavior (i.e., there are no "pretrends") and that training does not coincide with changes in, for example, officer assignments.

For our analyses, we collapse the stops data into cells at the level of an officer i by calendar week t. Each officer × week is also indexed by the time since diversity training, τ, weeks of THP experience as of that week e, and the assigned region of the state j. Using this officer by week panel dataset, we estimate event study models of the form:

$$Y_{ijte} = \sum_{\tau} \theta_\tau + \alpha_i + \delta_{jt} + \psi_e + u_{ijte} \tag{1}$$

where Y_{ijte} is the outcome of interest, θ_τ is an event time indicator, and α_i, δ_{jt} and ψ_e are fixed effects for officer, region by calendar week and officer weeks of experience.[10] Our primary outcomes of interest are the number of stops of motorists of a given race in a given week and the racial composition of stops, which we typically parameterize as the share of

[10] We obtain similar results including calendar week or district by calendar week (a smaller geography) instead of region by calendar week fixed effects.

stopped motorists who are non-Hispanic whites. Note that we do not weight regressions by stop volume because stop volume is an officer choice which may itself respond to treatment.

Recent advances in the econometrics literature have documented the various empirical issues associated with estimating event studies with two-way fixed effects via ordinary least squares (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 effects created by undesirable comparisons between currently treated and previously treated units as well as underidentification issues in fully dynamic specifications. To address these concerns, we estimate our event studies using the two-step imputation estimator proposed by Borusyak et al. (2022) and Gardner (2021).

In the first step, the fixed effect parameters are estimated by regressing the outcome on the fixed effects using only the not-yet-treated observations, i.e. calendar weeks prior to the officer participating in cultural diversity training:

$$Y_{ijte} = \alpha_i + \delta_{jt} + \psi_e + \tilde{u}_{ijte} \text{ for all } \tau_{it} < 0 \tag{2}$$

Estimated coefficients from this regression are then used to obtain estimates of the expected outcomes (if untreated) for each officer by calendar week observation:

$$\hat{Y}_{ijte}(0) = \hat{\alpha}_i + \hat{\delta}_{jt} + \hat{\psi}_e \tag{3}$$

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. To improve precision, we aggregate up from weekly estimates into sets of eight weeks over the two years before and after cultural diversity training:

$$\hat{\theta}_{\tau^k} = E(Y_{ijte} - \hat{Y}(0)_{ijte} | \tau^k <= \tau_{it} < \tau^k + 7) \tag{4}$$

We omit the week of training itself, so we estimate parameters $\hat{\theta}_1$, $\hat{\theta}_9$, $\hat{\theta}_{17}$, etc. extending forward in time for two years and $\hat{\theta}_{-8}$, $\hat{\theta}_{-16}$, etc. extending back two years prior to training.[11]

This imputation approach is particularly well-suited to our setting for three important reasons. First, officers vary considerably terms of when they obtain training and we observe many cohorts of officers entering service and eventually receiving training. As a result, many of the alternative approaches for addressing bias in "staggered roll-out" designs, which rely

[11] 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 also report the Borusyak et al. (2022) test for parallel pre-event trends.

on constructing cohort-by-cohort comparisons and then aggregating, are computationally problematic given the large number of cohorts (treatment timing groups) in our setting. Second, this imputation estimator easily accommodates more complex fixed effect structures than simple two-way fixed effects, which is potentially important in our setting given our interest in flexibly controlling for officer experience and for heterogeneity across officer patrol assignments. Finally, the imputation estimator allows for straightforward aggregation of event study estimates by taking averages over groups of event times in the second step. As noted above, we use a panel at the officer by week level to preserve the week-level variation in the timing of treatment, but estimate effects on eight period windows around training because the week-specific event study estimates tend to be imprecisely estimated.

Because our second stage estimates are conditional on the estimated fixed effects from the first stage, we compute standard errors using a Bayesian bootstrap (Rubin, 1981), clustering at the trooper-level.[12] The Bayesian bootstrap approach is identical to a classical bootstrap except that random weights are applied to each cluster in each iteration, rather than resampling clusters with replacement. An important advantage of this approach is that it preserves the support of all fixed effects in each replication. We draw random weights from a Dirichlet distribution for each officer in each bootstrap replication, and following Rubin (1981) we normalize the total weight for each officer to one over all bootstrap iterations. Our standard errors are simply the standard deviation of the estimates of the weighted, two stage event study parameters. Throughout, we use 100 bootstrap iterations for inference.

We also conduct the pretrend diagnostic test suggested by Borusyak et al. (2022). This pretrend test entails regressing the outcome on a set of pre-treatment event time indicators τ^{κ} using only not-yet-treated observations, i.e. for $\kappa < 0$, and then computing a joint significance test of the treatment lead indicators.

4 Results

4.1 Main results

Figure 1 examines the racial composition of officer stops around the timing of cultural diversity training. Specifically, panel (a) presents event study estimates for the stop volume of non-Hispanic white motorists and the combined stop volume of black and Hispanic motorists, and panel (b) presents the event study estimates for the share of stops that are of non-Hispanic white motorists. The horizontal axis presents weeks relative to training for two years before and after training. The event study estimates, which are estimated for 8-week

[12] Both Borusyak et al. (2022) and Gardner (2021) derive analytic standard errors for the imputation event study estimator. However, these analytical standard errors are too computationally burdensome in our setting given the large number of treatment timing groups. 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.

bins, are presented centered on the average of the weeks included in an estimate, so for example the estimate $\hat{\theta}_{\tau^1}$ is located at 4.5 weeks. In panel (a), the number of non-Hispanic white stops is identified by solid dots, while the number of stops of black and Hispanic motorists is identified by a hollow square. In both panels, shaded regions represent 95 percent confidence intervals.

First, we note that there is no evidence of differential changes in officer behavior in either the number of non-Hispanic white, the number of black and Hispanic stops, or the white share of stops during the period leading up to training. We also generally pass the test of pre-trends prescribed by Borusyak et al. (2022) over a two year pre-period for white stops ($F = 1.034$, $p = 0.414$), black and Hispanic stops ($F = 1.177$, $p = 0.294$), and share of stops that are white ($F = 1.761$, $p = 0.050$). While the null of no significant deviations from zero is rejected at the ten percent level for white share stops over a two year window, panel (b) shows minimal visual evidence of a pre-training trend. Further, when extending the pre-period window to four and six years prior to training, these pre-trend tests yield p-values of 0.136 and 0.069, respectively. In other words, conclusions about pre-training trends are not meaningfully different as we look back further in time.

Immediately following training, we observe an increase in the number of stops of white motorists which persists over the post-training period. Of the 13 eight-week estimates, 10 are statistically significant at 95 percent confidence level. The overall difference-in-differences estimate implies an increase of 0.9 stops of white motorists per week ($se = 0.3$), or about a 9 percent increase relative to the mean. On the other hand, the estimated impact of training on minority stops is quantitatively small, with event study estimates generally not statistically distinguishable from zero. The aggregated difference-in-differences estimate for minority stops implies an increase of 0.3 stops per week, which represents just a 2 percent increase relative to the mean and is not statistically significant at conventional levels.

Turning to the racial composition of stops, presented in panel (b), we find a corresponding increase in the probability that stopped motorist is white immediately following training. Mirroring the pattern observed in panel (a), which shows a modest increase in minority stops during the middle of the post-training period, event study estimates for the white share of stops attenuate over the period 30-60 weeks out from training but then revert back to their initial levels, suggesting some persistence in the treatment effects. The aggregated difference-in-difference estimate implies an estimated increase of 1.3 percentage points ($se =$ 0.7 percentage points), or about a 3 percent increase relative to the benchmark mean. Taking panels (a) and (b) together, our findings suggest that troopers reduce their minority fraction of stops following cultural diversity training but achieve this by making additional stops of white motorists, rather than reducing their stops of black or Hispanic drivers.

A natural question is whether the effects that we document are large or small in terms of magnitude. On the one hand, a 1.3 percentage point increase in the share of stops which are of white motorists (or a 9 percent increase in non-Hispanic white stops) may seem quite small in absolute terms, *prima facie*. On the other hand, if the practical effect of cultural diversity training is to mitigate discriminatory behavior by officers, then the magnitude of our treatment effect estimates should be bounded by the extent of discrimination among officers in our sample.

To benchmark our estimated magnitudes, we compute 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, the idea being that an over-representation of minorities during daylight suggests racial profiling by officers when race is observable *prior* to making a stop. Using only the pre-training data, our benchmark VOD estimate implies that a stopped motorist is about 2 percentage points ($se = 0.4$ percentage points) less likely to be non-Hispanic white during daylight. Hence, our estimate of the impact of cultural diversity training on the fraction of white stops is 65 percent as large as an estimate of share of white stops attributable to discrimination or racial profiling behavior by officers using a well-established method in the literature.

In appendix table D-1, we also explore how the results of our VOD test change following diversity training. While we find evidence of racial profiling by the troopers in our sample using the VOD test prior to cultural diversity training, the same test yields no evidence of discrimination in the post-training period, consistent with our main findings.

4.2 Validation and Robustness Tests

A salient threat to the validity of our main findings is the possibility that troopers' patrol assignments change systematically around the time of cultural diversity training. For example, our main finding that an officer's racial composition of stops changes following training could be driven by changes in patrol areas rather than behavioral changes by officers.

THP troopers are typically assigned to one large region of the state (recall that we control for time by assigned region fixed effects in our baseline specification) and then do the vast majority of the patrolling with a specific district, consisting of several counties, within that region. While a trooper's geographic assignment is not recorded in our data, we can empirically assign officers into districts in a given week based on where we observe officers making the majority of stops (as described in further detail in appendix B-2).

To verify that changes in patrol assignments do not coincide with the timing of cultural diversity training, we repeat our baseline event study estimates where the outcome of interest is the racial composition of residents of the officer's assigned patrol district in each week (taken from the 2019 census). As shown in figure 2, we find no evidence of systematic changes in assigned districts leading up to training or following training. The overall difference-in-difference estimate suggests that, following training, officers patrol districts which are 0.3 percentage points less white ($se = 0.8$ percentage points).

Another salient concern for our results arises from the fact that, during this period of their careers, troopers are taking other trainings in addition to the cultural diversity training that is the focus of our analyses. Some of these trainings, such as a dedicated training on the topic of racial profiling, might also be expected to change an officer's stopping patterns.

Noting that this is only a concern if the timing of other trainings tends to coincide with cultural diversity, we first report event study estimates where the outcome of interest is an indicator for whether a trooper took another training in appendix figure C-5. Although the estimated magnitude is small (about 2 percentage points), we indeed find evidence that some officers take other trainings, especially trainings on arrest protocol and racial profiling, around the same time as they take cultural diversity training.

To assess the importance of this concern for our empirical conclusions, we re-estimate versions of our baseline event studies which control for the potential influence of the other four primary training modules: racial profiling, arrest, traffic stop and de-escalation training. Specifically, we add as controls to our event study specification interactions between calendar week effects and the time each officer took these alternative trainings. Hence, these specifications identify the impact of cultural diversity training by comparing among troopers who took the alternative training of interest at the same time (i.e., the identifying variation is variation in the timing of cultural diversity training, holding fixed the timing of, say, the racial profiling or arrest training).

Focusing on the weekly number of stops of white motorists, these results are presented in figure 3. We present the corresponding estimates for the number stops of Black and Hispanic motorists, as well as for the white share of stops, in appendix figures A-6, A-7, and B-1 respectively. As shown in panels (b) through (d), Controlling for exposure to arrest, traffic, and de-escalation training has minimal impact on our conclusions; if anything, our difference-in-difference estimates of the impact of cultural diversity training are slightly larger when conditioning on these other trainings. Controlling for exposure to racial profiling training, on the other hand, attenuates our overall DiD estimate by roughly one third.

While this suggests that our baseline estimates may be biased slightly upward by the confounding impact of contemporaneous training focused on racial profiling, also worth highlighting is the fact that we cannot statistically reject the equality of the overall DiD estimates with and without controls for the timing of racial profiling training. Moreover, estimates where the outcome of interest is the white share of stopped motorists (shown in figure B-1) are very similar with and without conditioning on the racial profiling training.

4.3 Mechanisms

In figure 4, we examine which types of officers are most responsive to diversity training in terms of their enforcement behavior. Panels (a) and (b) present event study estimates for the number and fraction of stops of white motorists, respectively, by officer race. Perhaps surprisingly, the baseline impacts we document are driven primarily by the responses of nonwhite officers. Following training, Black and Hispanic officers make an additional 1.3 weekly stops of white motorists ($se = 0.42$), while the comparable figure for white officers is just 0.37 ($se = 0.43$) for white officers. This difference is particularly stark in light of disparate counterfactual means by officer race; these estimates imply about a 20 percent increase for nonwhite officers and just a two percent increase for white officers. Panel (b) similarly illustrates that the overall increase in the white share of stops is driven by nonwhite officers. However, we also note that the impact for nonwhite officers is just below standard thresholds for statistical significance.

Panels (c) and (d) of figure 4 report estimates when splitting officers based on their pre-training stop patterns. As shown in panel (a), we estimate similar impacts of training on the number of weekly stops of white motorists for officers with above and below median white share of stops prior training. However, panel (d) reveals that the overall impact of training on the white share of stops is wholly explained by officers with a below median white share of stops prior to diversity training, whose share of stops attributable to white motorists increases by 0.38 percentage points, or about 14 percent relative to their counterfactual mean. Note that the apparently disparate conclusions offered by panels (c) and (d) imply that officers with an above-median white share of stops prior to training also increase their number of minority stops, so there seems to be a sense in which the training encourages additional stops of whichever group an officer makes disproportionately fewer stops.

A natural hypothesis arising from our findings that cultural diversity training increases the number of white stops, particularly for those with a lower white share of stops prior to training, is that the training induces officers to make additional stops of white motorists guilty of more "marginal" traffic violations. In figure 5, we explore this theory by estimating event studies where the outcomes of interest are the number of speeding stops, the number of stops we classify as pretextual[13], and the average speed among those stopped for speeding, focusing only on white motorists in all cases. As shown in panels (a) and (b), we find statistically significant increases in both speeding and pretext stops of white motorists following diversity training; the total increase in these two stop types is about 1.1 additional stops per week, similar to our baseline estimate of an additional 0.9 weekly stops of white motorists. While the overall increase in white stops appears primarily attributable to speeding stops, we also note that the increase in pretextual stops is sizeable as a fraction of the counterfactual mean (about 15 percent).

Given the result that diversity training increases the number of speeding stops of white

[13] We use this term to refer to stops possibly made in order to observe the motorist more closely due to a suspicion of criminal activity, rather than stops due to a specific traffic violation, *per se* (e.g., Feigenberg and Miller 2023). See the appendix for additional details on how we classify pretextual stops in our sample.

motorists, we next explore the impact on the severity of these speeding stops in panel (c).[14] We find that the average speed (relative to the posted limit) among white motorists cited for speeding declines by about 0.5 miles per hour following training, with the overall average estimate statistically significant at conventional levels. We interpret this finding as evidence that the increase in the number of speeding stops is attributable to additional stops of white motorists committing less severe speeding infractions. In other words, the impact of training appears to be a reduction in lenience towards white speeders, with trained troopers making stops of less "guilty" white motorists they were previously letting pass without stopping.[15]

Another mechanism we consider is whether diversity training induces officers to shift their patrol areas (within their assigned patrol region). Specifically, we assign each stop to a county and estimate event studies where the outcome of interest is the racial composition of the county where the stop occurred (parameterized as the white population share from the 2019 ACS), rather than the race of the stopped motorist. As in our baseline specification, we condition on the assigned patrol region, so any impacts on this margin are attributable to officers changing the location of their enforcement activities *within* their assigned regions. We report these event study estimates in panel (a) of figure 6. Although the estimates are somewhat imprecise, we do find suggestive evidence of an increase in enforcement in areas with larger white populations following diversity training.

Panel (b) of figure 6 then asks to what extent this potential change in patrol location can explain our overall findings. Specifically, we repeat our baseline analysis examining the number of white stops, adding as a control the white population share of an officer's patrol area. Consistent with panel (a), we find some attenuation in the specification with controls indicating that some of the observed effect of training comes from officers making more stops in predominately white areas. However, our baseline conclusions are not altered meaningfully in either a quantitative or substantive sense in this alternative specification. In other words, changes in patrol locations (within assignments) cannot explain our overall estimated impact of diversity training.

[14] Note that, while evidence of manipulation in the speed distribution (i.e., bunching below fine increases) has been documented in other settings, we find no evidence of such manipulation for speeding citations issued by the Texas Highway Patrol.

[15] In figure A-8, we show that there is no comparable increase in speeding or pretextual stops for minority motorists. In figure A-9, we show that the rate at which troopers issue citations versus warnings to stopped white motorists is not affected by diversity training, suggesting that officers also issue citations to these marginal white speeders. We also find no changes in the probability that stop leads to a search or arrest following training and find no impact of training on these outcomes.

5 Conclusion

In this paper, we study the effects of cultural diversity training on the enforcement behavior of earlier-career highway patrol officers in Texas. 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 traffic stops changes systematically following diversity training. Immediately following training, the share of a trooper's stops of white motorists increases by around two percentage points. While this estimate attenuates slightly in the medium term, we find an overall increase in the white share of stops of around 1.3 percentage points over the two years after training.

Benchmarked against a standard estimate of the extent of discrimination in stopping decisions from the literature, the veil of darkness (VOD) test, applied to our untreated data, our estimates suggest that diversity training erodes over half the discrimination practiced by the average officer. Moreover, this standard VOD test detects no evidence of disparate treatment when applied specifically to officers following diversity training.

Trained troopers achieve this change in the racial composition of stops by stopping additional white motorists, rather than reducing their stops of minority drivers. Specifically, we estimate that, following training, officers make an additional 0.9 weekly stops of white motorists. We find suggestive evidence that these additional stops are of less "guilty" drivers. Specifically, we document that the increase in stops of white motorists is largely attributable to speeding stops and then show that, following training, the average speeds of white motorists stopped for speeding declines. In other words, we find that diversity training induces additional stops of white drivers committing less severe infractions. This aligns well with the theory that diversity training prompts officers to stop more 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).

On the one hand, our analyses suggest that cultural diversity training can reduce racial disparities in enforcement behavior among highway patrol officers. Our findings suggest that a key mechanism behind this result is causing officers to systematically reduce lenience in their behavior towards white drivers, improving the overall fairness of traffic enforcement. On the other hand, we also note that more stringent enforcement applied to white motorists, coupled with no change in behavior towards minority civilians, 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 with minorities (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.

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 lternative test of racial prejudice in motor mehicle searches: Theory and evidence. *American Economic Review 96*(1), 127–151.

Arnold, D., W. Dobbie, and C. S. Yang (2018). Racial bias in bail decisions. *Quarterly Journal of Economics 133*, 1885–1932.

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. S. (1957). The economics of discrimination. *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.

Bordalo, P., K. Coffman, N. Gennaioli, and A. Shleifer (2016). Stereotypes. *The Quarterly Journal of Economics 131*, 1753–1794.

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.

Campbell, R. (2023). What does federal oversight do to policing and public safety? evidence from seattle. *Working Paper*. (Working Paper published in 2023, cited as 2024 in text).

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*.

Devi, T. and R. G. Fryer, Jr. (2021). Policing the police: The impact of 'pattern-or-practice' investigations on crime. *Working Paper*.

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*.

Fagan, J. A. and A. Geller (2020). Profiling and consent: Stops, searches, and seizures after soto. *Virginia Journal of Social Policy and the Law 16*.

Feigenberg, B. and C. Miller (2020). Racial disparities in motor vehicle searches cannot be justified by efficiency. *NBER Working Paper 27761*.

Feigenberg, B. and C. Miller (2023). Class disparities and discrimination in traffic stops and searches. *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.

Heaton, P. (2010). Understanding the effects of antiprofiling policies. *The Journal of Law and Economics 53*(1), 29–64.

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.

Hull, P. (2021). What marginal outcome tests can tell us about racially biased decision-making. *Working Paper*.

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.

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.

Long, W. (2019). How does oversight affect police? evidence from the police misconduct reform. *Journal of Economic Behavior Organization 168*, 94–118.

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

MacDonald, J. and A. Braga (2019). Did post-floyd et al. reforms reduce racial disparities in nypd stop, question, and frisk practices? an exploratory analysis using external and internal benchmarks. *Justice Quarterly 36*(5), 954–983.

Matsuzawa, K. (2024). Are pretextual stops inefficient inequitable? *Working Paper*.

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.

Naddeo, J. and R. Pulvino (2024). The effects of reducing pretextual stops: Evidence from saint paul, minnesota. *Working Paper*.

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.

Parker, S. T., M. B. Ross, and S. L. Ross (2024). Driving change: Evaluating connecticut's collaborative approach to reducing racial disparities in policing. *NBER Working Paper*.

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. (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*.

Table 1: Summary statistics for officers in analysis sample

	All (1)
Age	30.18
Male	0.903
Race = White	0.517
Race = Black	0.077
Race = Hispanic	0.381
Experience at Training	40.78
Troopers	1662

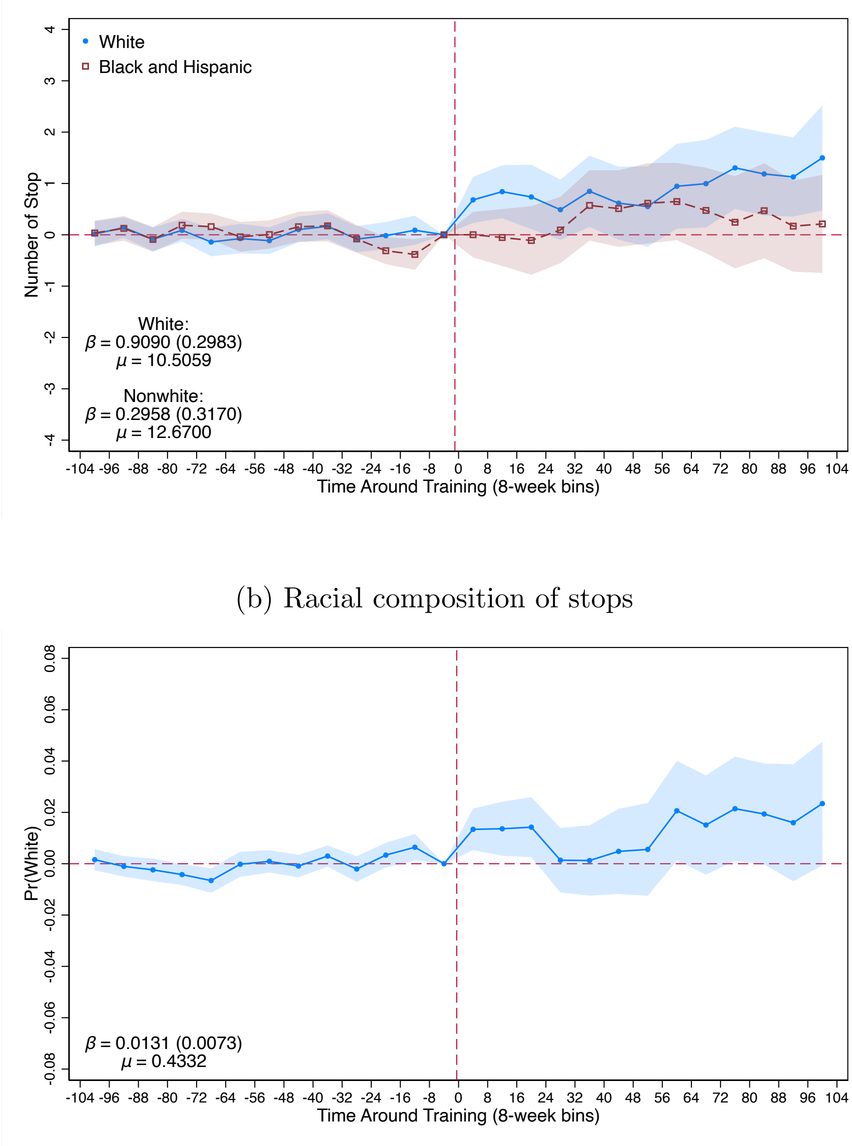
*N*otes: This table reports summary statistics for officers in the analysis sample. Age is measured as of the officer's hire date. Experience at training is an officer's months of writing citations prior to their cultural diversity training.

Table 2: Summary statistics for stops in analysis sample

	Pre Train (1)	Post Train (2)
Weekly Stops	25.04	23.02
Share White	0.426	0.439
Weekly Citations	8.33	8.11
Share White	0.398	0.374

*N*otes: This table reports summary statistics for stops in the analysis sample. Column (1) reports means for stops before training and column (2) reports means for stops after training.

Figure 1: Effects of cultural diversity training on officer stops

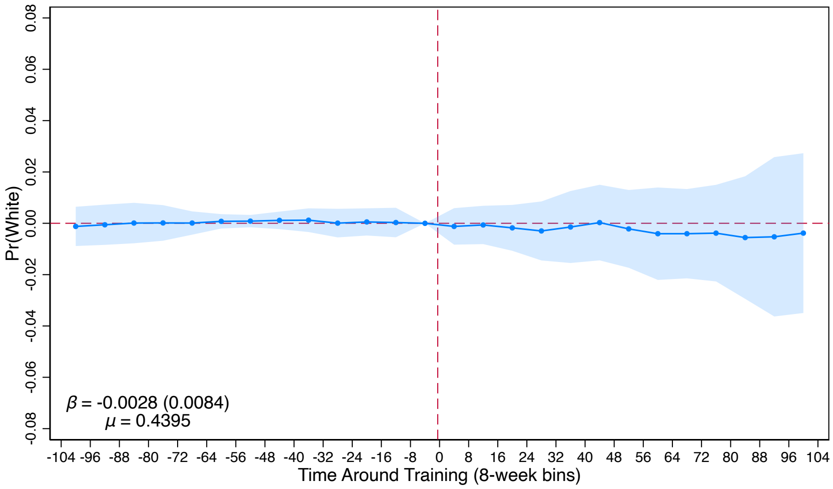

(a) Stop volume by motorist race

(b) Racial composition of stops

*N*otes: Panel (a) plots imputation-based aggregated 8-week event study estimates using an officer × week panel where the outcome is an officer's number of stops in a given week, separately by motorist race. Panel (b) plots imputation-based aggregated 8-week event study estimates where the outcome is the share of officer's stops in a given week that are of white motorists. The *p*-values from the Borusyak et al. (2022) pretrends test are 0.66, 0.22, and 0.06.

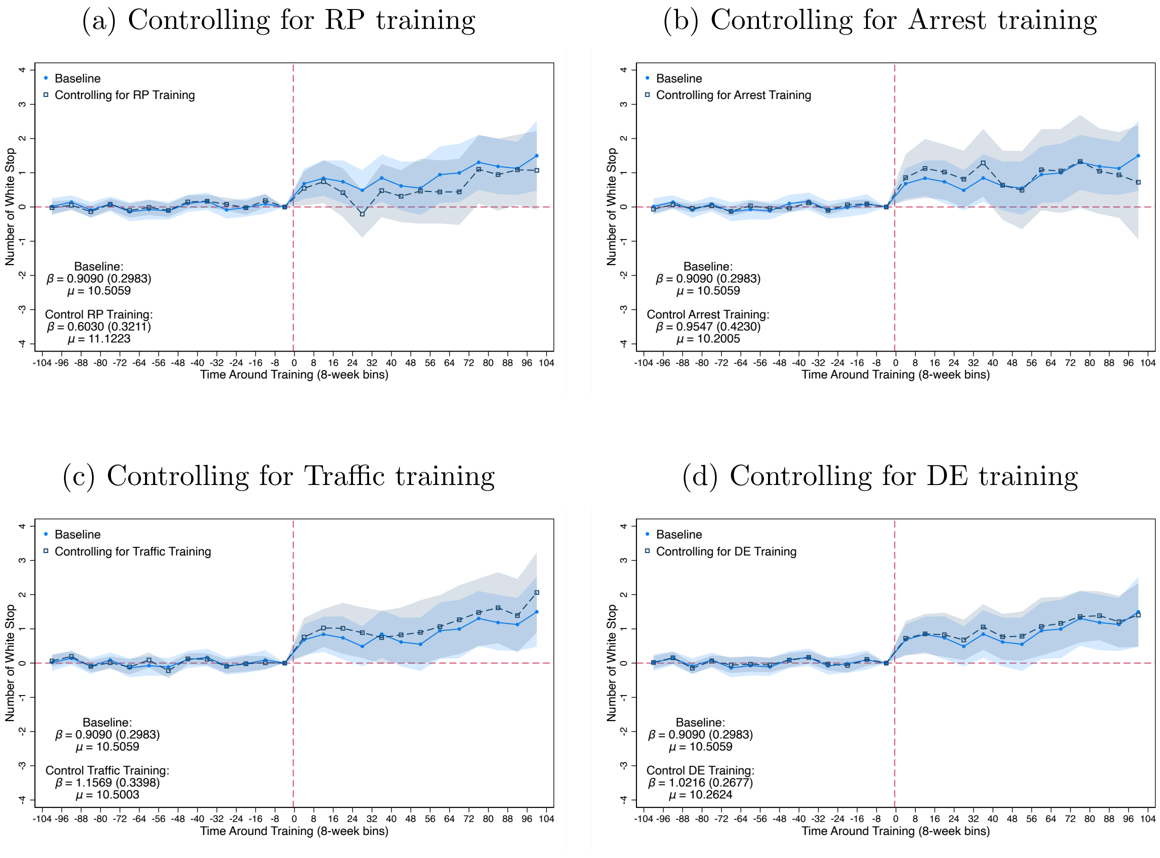
Figure 2: Cultural diversity training and assigned patrol locations

(a) Racial composition at district level



Notes: Similar to figure 1, this figure reports event study estimates using an officer × week panel and officer, region by calendar week and officer weeks of experience fixed effects, except that the outcome of interest is the white share of residents in an officer's assigned patrol district in each week (based on the 2019 ACS).

Figure 3: Effects of cultural diversity training controlling for the timing of other trainings



(a) Controlling for RP training

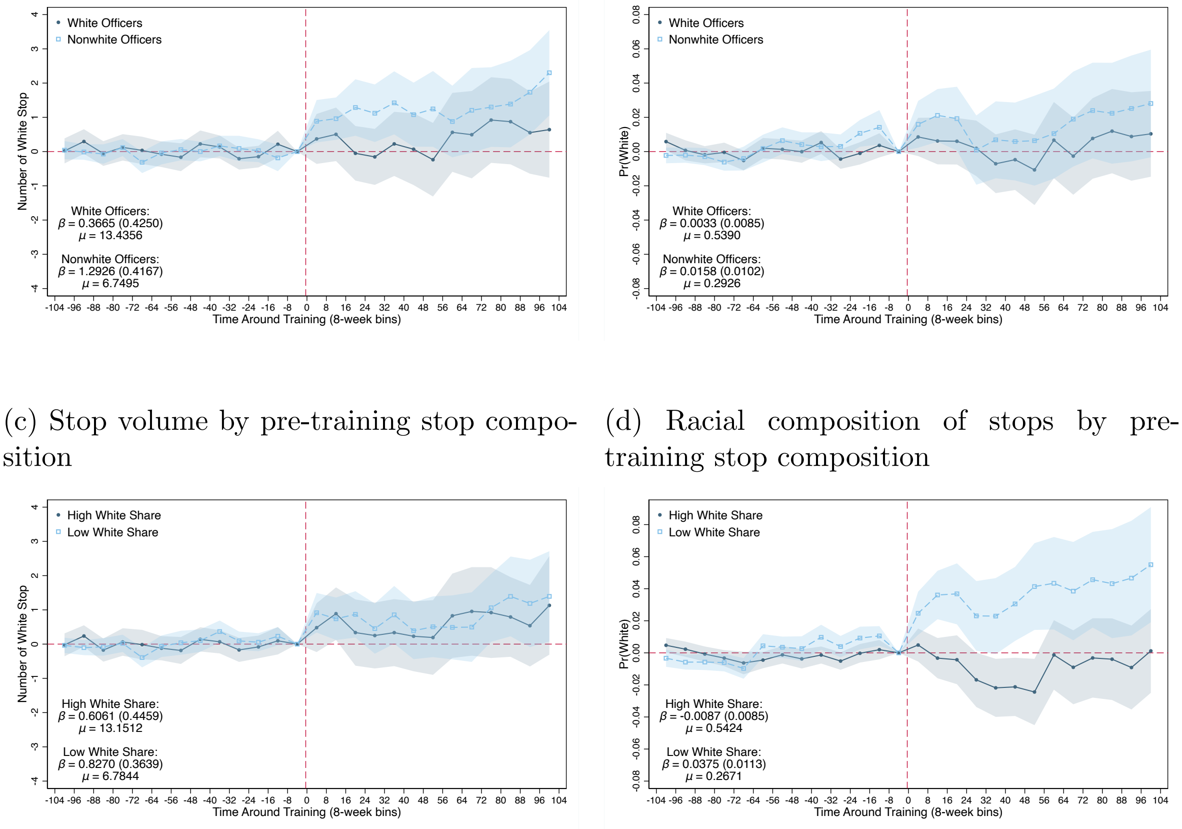
(b) Controlling for Arrest training

(c) Controlling for Traffic training

(d) Controlling for DE training

*N*otes: This figure reports imputation-based event study estimates where the outcome of interest is the weekly number of stops of white motorists. In each panel, we report estimates from our baseline specification as well as a specification which flexibly controls for exposure to other trainings (racial profiling, arrest, traffic, and de-escalation) by including week fixed effects which are interacted with the timing of the (denoted) other training. For example, in panel (a), hollow squares report estimates which also include week fixed effects interacted with the week in which a trooper took racial profiling training. Appendix figures A-6 and A-7 report the identical estimates for the number of stops of Black and Hispanic motorists, and appendix figure B-1 reports the corresponding estimates where the outcome is the white fraction of stops.

Figure 4: Heterogeneous effects by officer characteristics



(a) Stop volume by officer race

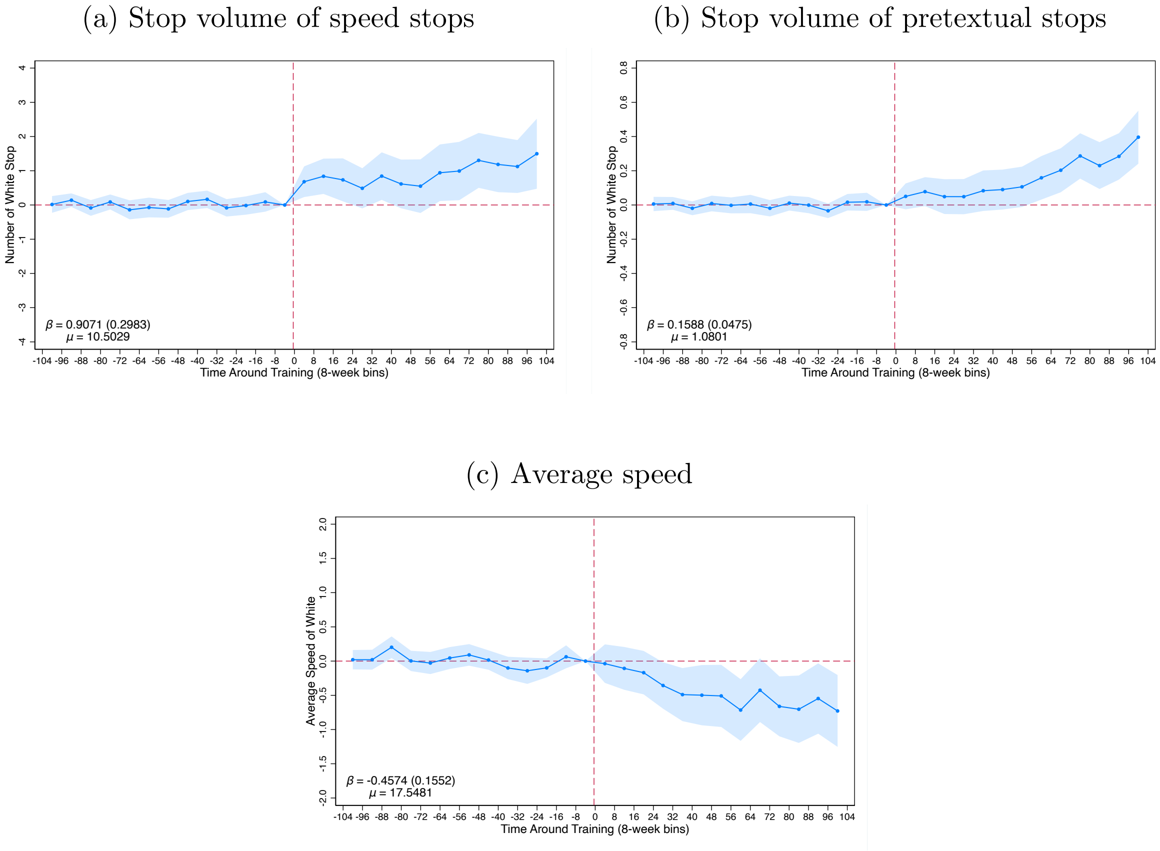
(b) Racial composition of stops by officer race

(c) Stop volume by pre-training stop composition

(d) Racial composition of stops by pre-training stop composition

*N*otes: This figure reports imputation-based event study estimates where the outcome of interest is the weekly number of stops of white motorists. Panels (a) and (b) are the same as figure 1 except that we show estimates separately by officer race. Panels (c) and (d) are the same as figure 1 except that we show estimates separately by troopers' pre-training racial composition of stops. Specifically, we estimate each officer's location-adjusted white share of stops in the pre-training period and then split officers into those with above- and below-median pre-training white shares. Figures A-3 and A-4 report the same estimates for stops of Black and Hispanic motorists, respectively.

Figure 5: Effects of cultural diversity training on stop composition



(a) Stop volume of speed stops

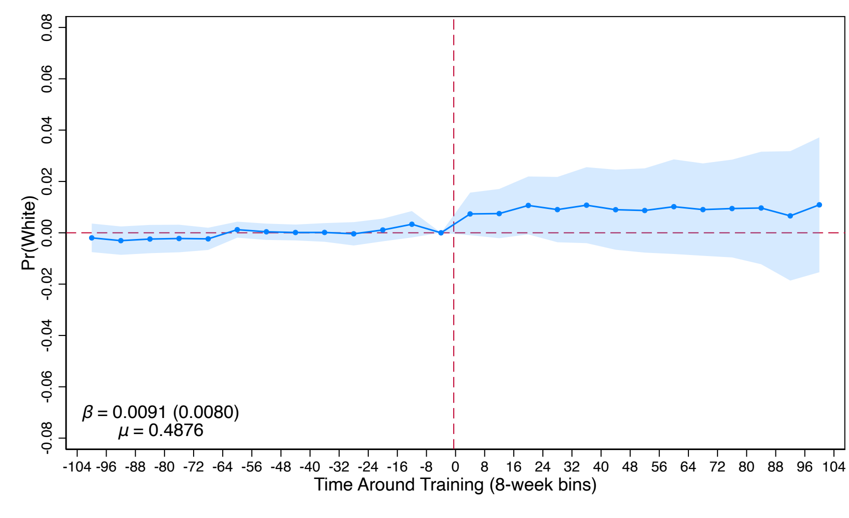
(b) Stop volume of pretextual stops

(c) Average speed

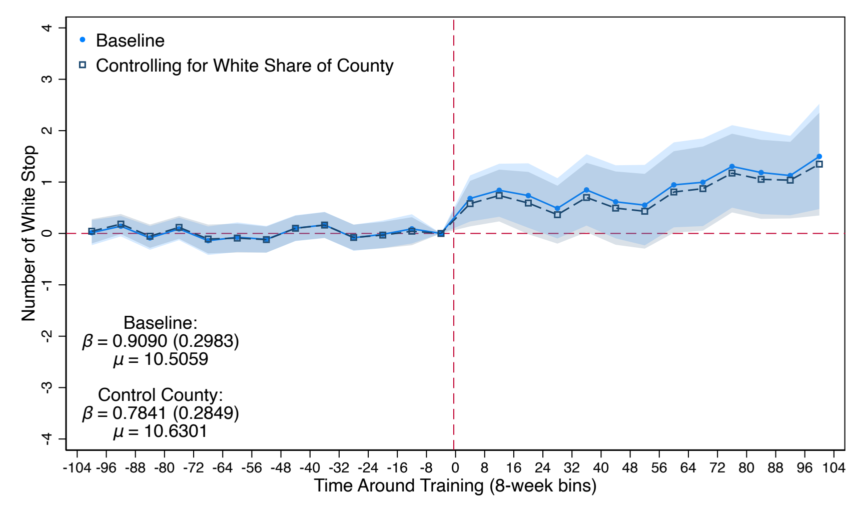
*N*otes: Panels (a) and (b) report imputation-based event study estimates where the outcome of interest is the weekly number of stops of white motorists for specific types of stops. In panel (a), the outcome of interest is the number of stops for speeding infractions. In panel (b), the outcome is the number of stops we classify as *pretextual*, or stops made with the intent of detecting more serious crime (Feigenberg and Miller, 2023). In panel (c), we report imputation-based event study estimates where the outcome is the average speed (relative to the posted speed limit) among stopped (white) speeders in a given week.Figure A-8 reports identical estimates for Black and Hispanic motorists.

Figure 6: Effects of cultural diversity training on stop locations

(a) Racial composition at county level



(b) Stop volume : controlling for county census



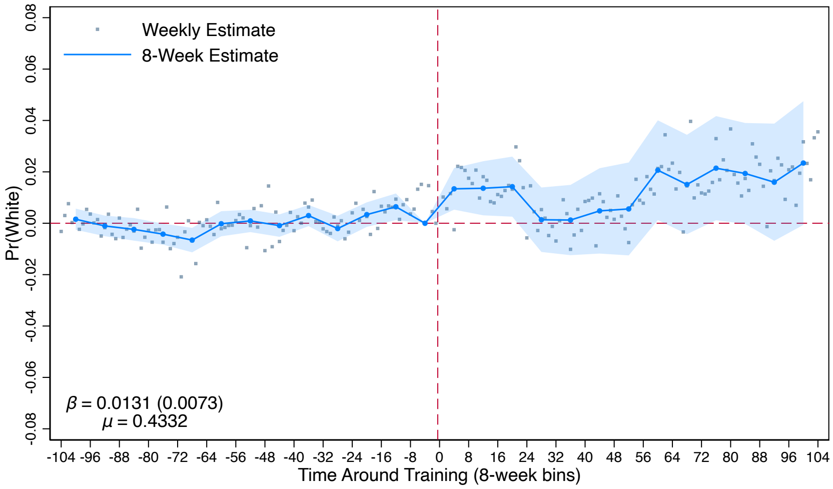
*N*otes: Panel (a) is same as panel (b) of figure 1 except that the outcome is the white share of residents in the county where the stop occurred (based on the 2019 ACS). Panel (b) is the same as panel (a) of figure 1 except that we also report estimates from an additional specification which we add the share of white residents in the officer's assigned patrol county as a control variable.

FOR ONLINE PUBLICATION: APPENDICES

A Supplementary results

Figure A-1: Effects of cultural diversity training on officer stops

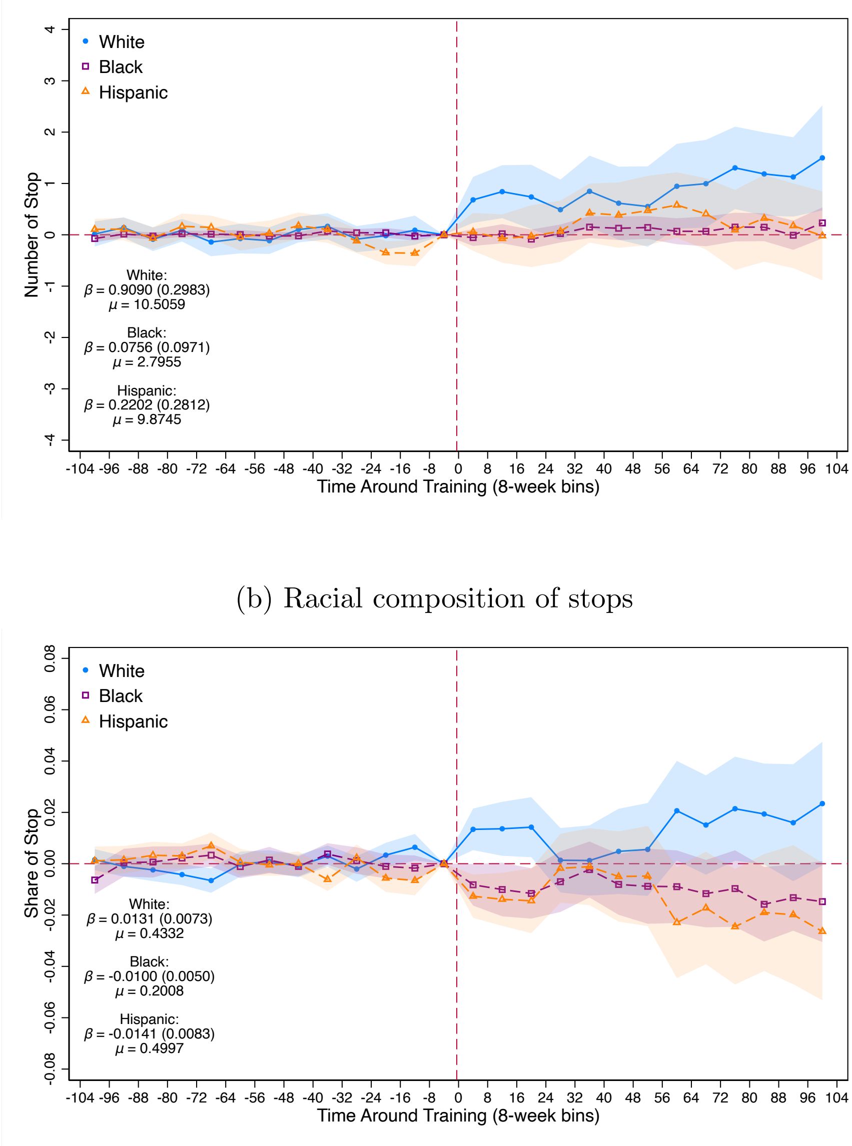
(a) Racial composition of stops



*N*otes: Same as figure 1 panel (b) except that we also reports imputation-based weekly (open circles) event study estimates.

Figure A-2: Effects of cultural diversity training on officer stops

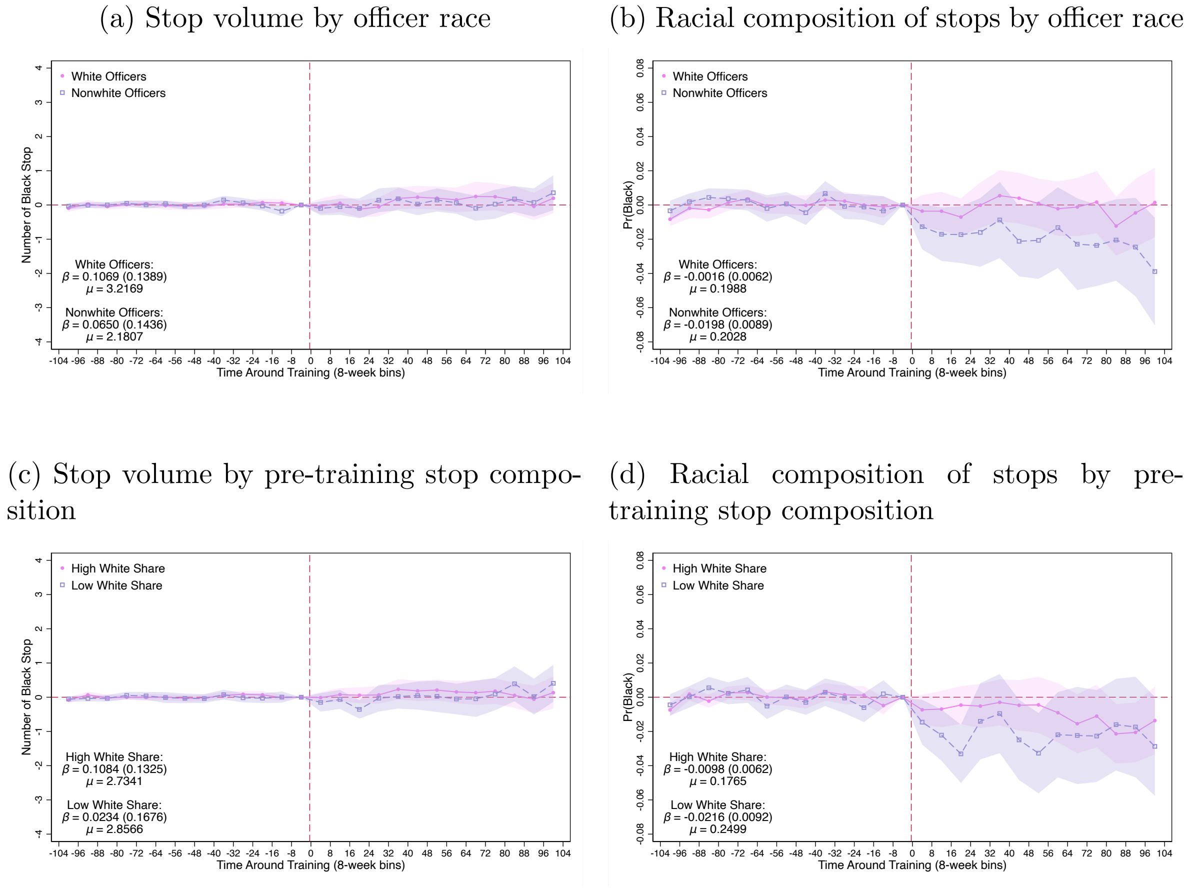
(a) Stop volume by motorist race



(b) Racial composition of stops

*N*otes: Same as figure 1 except that the outcome of interest is shown separately by motorist race.

Figure A-3: Heterogeneity effects by officer characteristics (stops of Black motorists)


(a) Stop volume by officer race

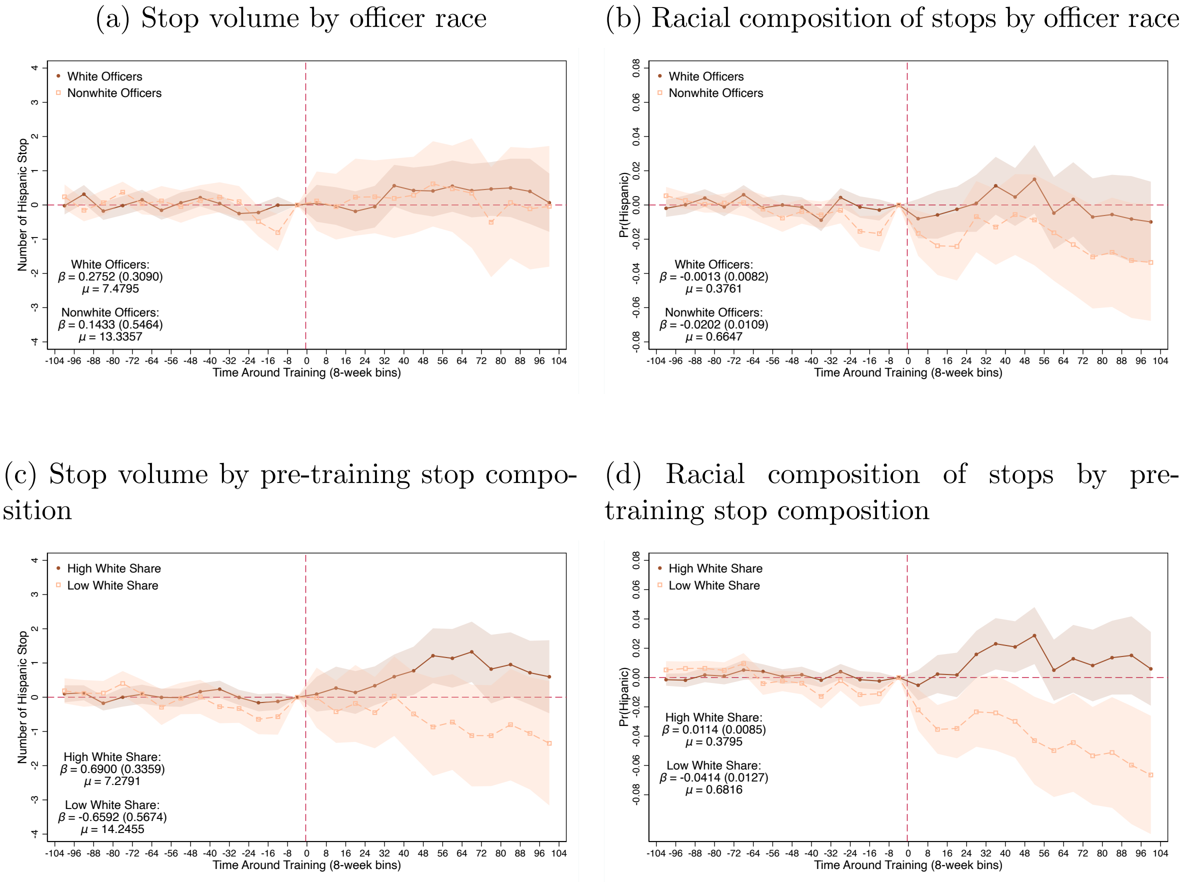
(b) Racial composition of stops by officer race

(c) Stop volume by pre-training stop composition

(d) Racial composition of stops by pre-training stop composition

*N*otes: Same as figure 4 except that the outcome of interest is an officer's number/share of stops of Black motorists in a given week.

Figure A-4: Heterogeneous effects by officer characteristics (stops of Hispanic motorists)



(a) Stop volume by officer race

(b) Racial composition of stops by officer race

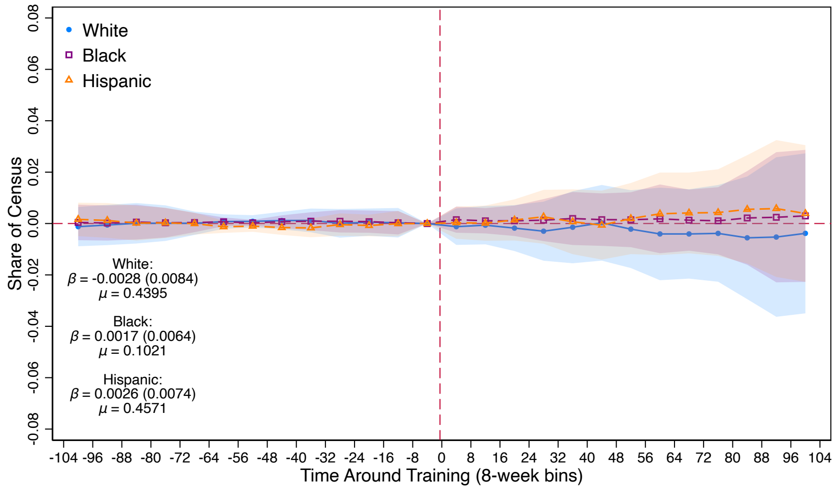
(c) Stop volume by pre-training stop composition

(d) Racial composition of stops by pre-training stop composition

*N*otes: Same as figure 4 except that the outcome of interest is an officer's number/share of stops of Hispanic motorists in a given week.

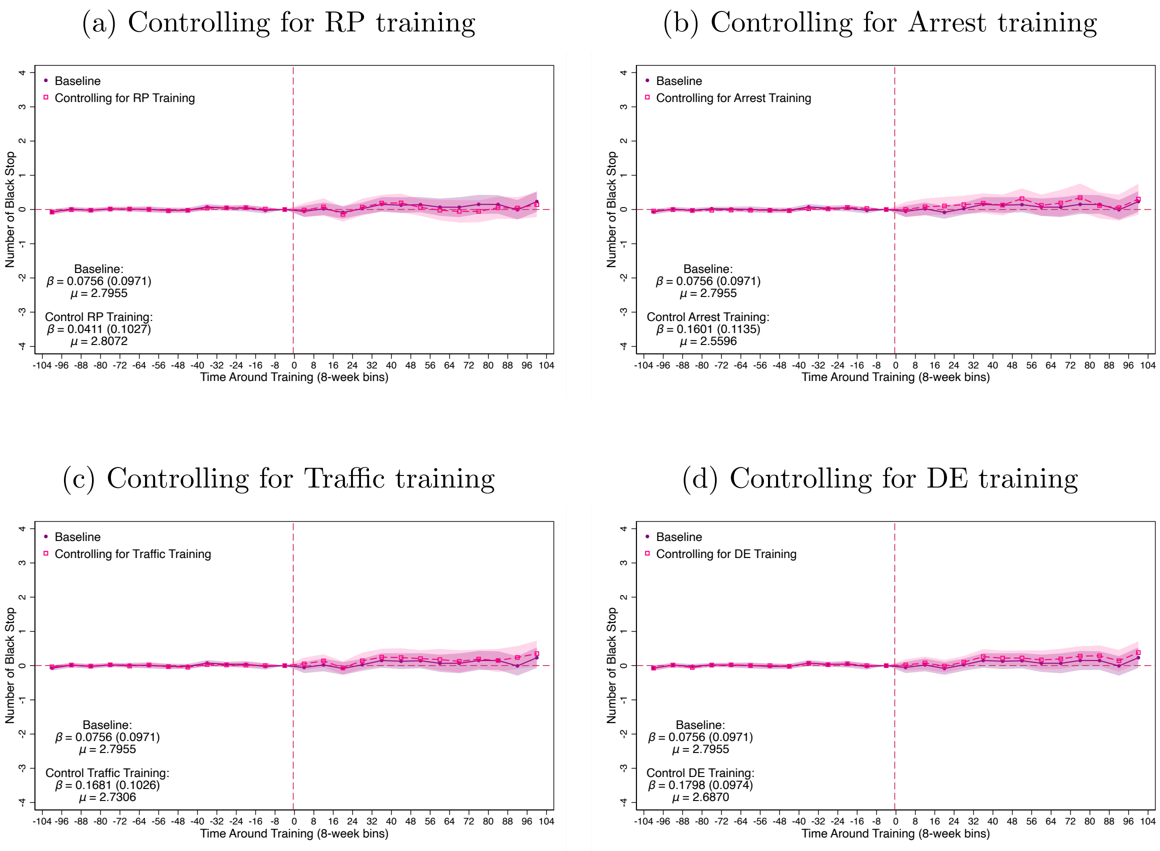
Figure A-5: Cultural diversity training and characteristics of patrol locations

(a) Racial composition at district level



*N*otes: Same as figure 2 except that we also show results for Black and Hispanic resident shares (from the 2019 ACS) for officers' assigned districts.

Figure A-6: Effects of cultural diversity training on stops of Black motorists, controlling for the timing of other trainings



(a) Controlling for RP training

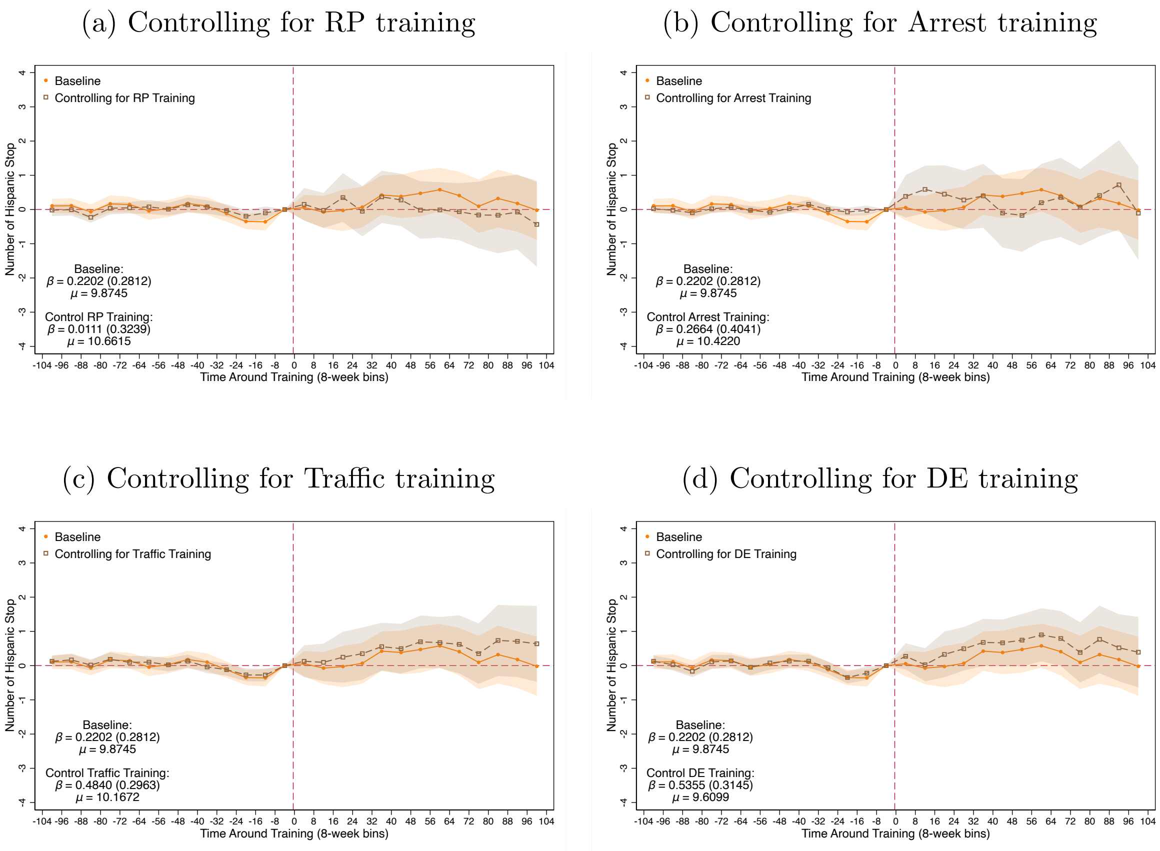
(b) Controlling for Arrest training

(c) Controlling for Traffic training

(d) Controlling for DE training

*N*otes: Same as figure 3 except that the outcome of interest is an officer's number of stops of Black motorists in a given week.

Figure A-7: Effects of cultural diversity training on stops of Hispanic motorists, controlling for the timing of other


(a) Controlling for RP training

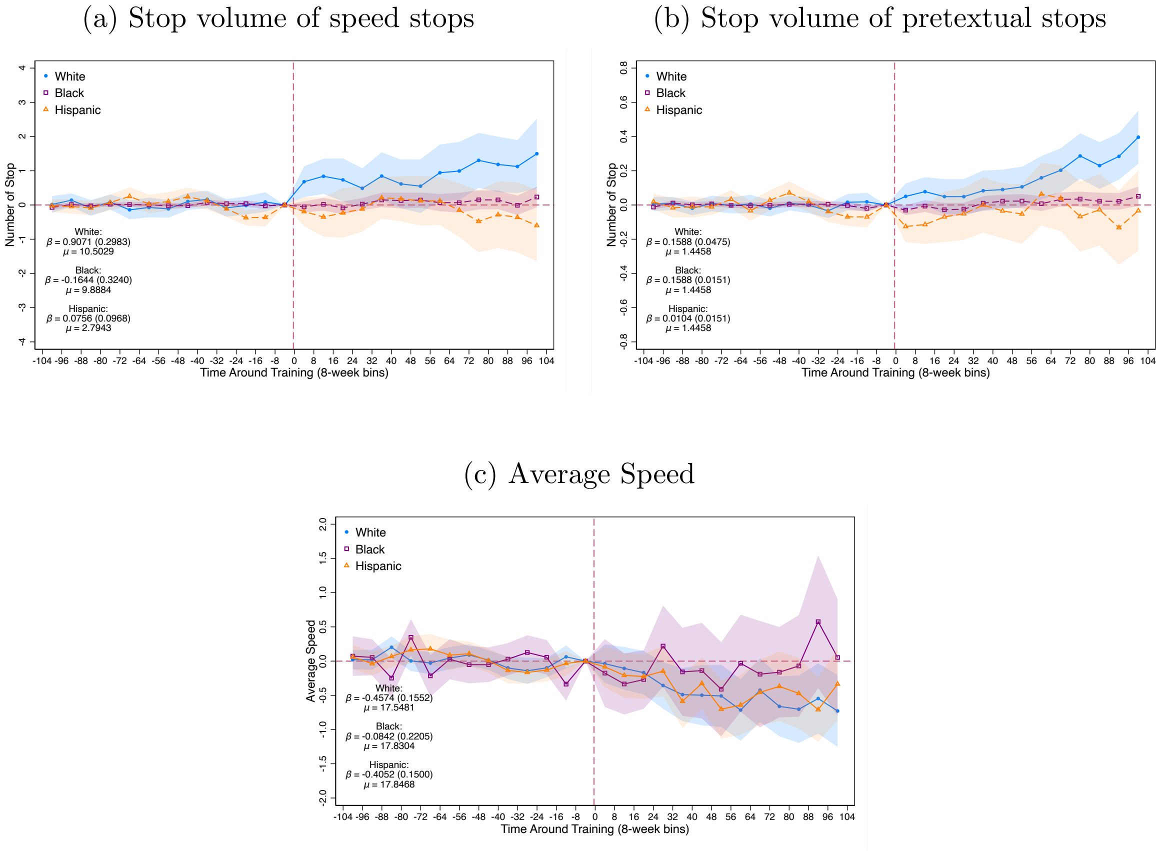
(b) Controlling for Arrest training

(c) Controlling for Traffic training

(d) Controlling for DE training

*N*otes: Same as figure 3 except that the outcome of interest is an officer's number of stops of Hispanic motorists in a given week.

Figure A-8: Effects of cultural diversity training on stop composition



(a) Stop volume of speed stops

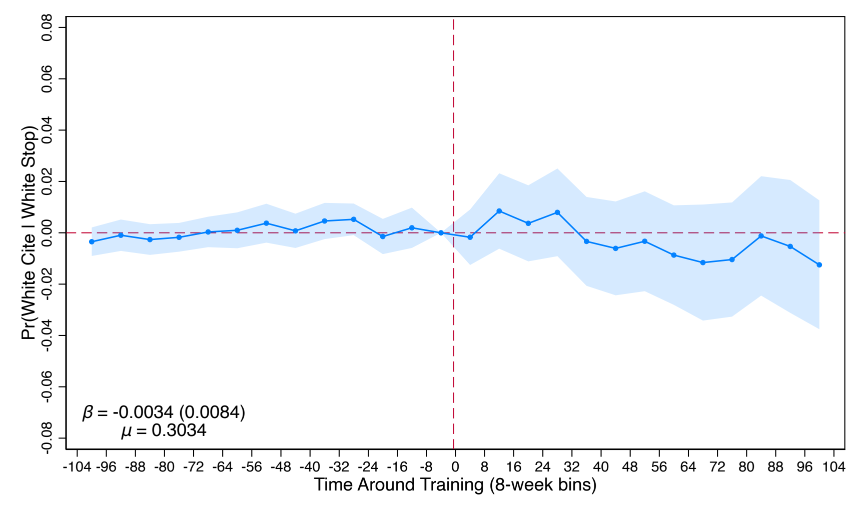
(b) Stop volume of pretextual stops

(c) Average Speed

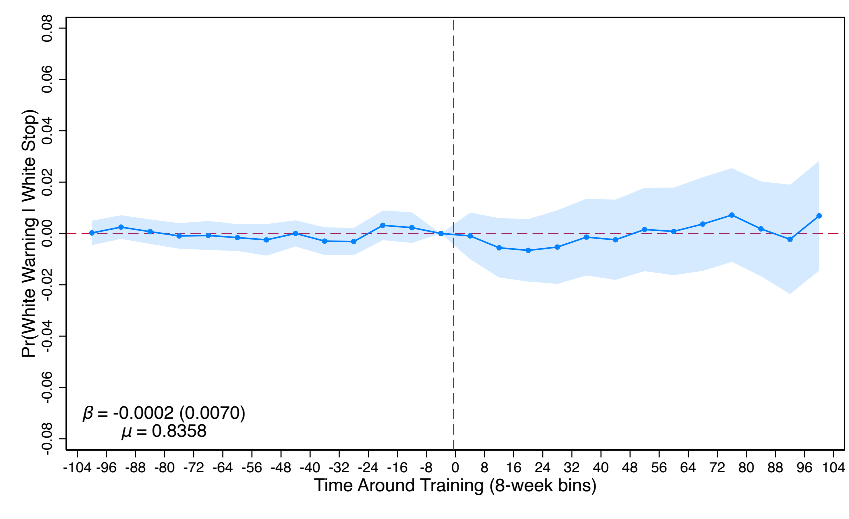
*N*otes: Same as figure 5 except that the outcome of interest is shown separately by motorist race.

Figure A-9: Effects of cultural diversity training on stop dispositions

(a) Citation issued

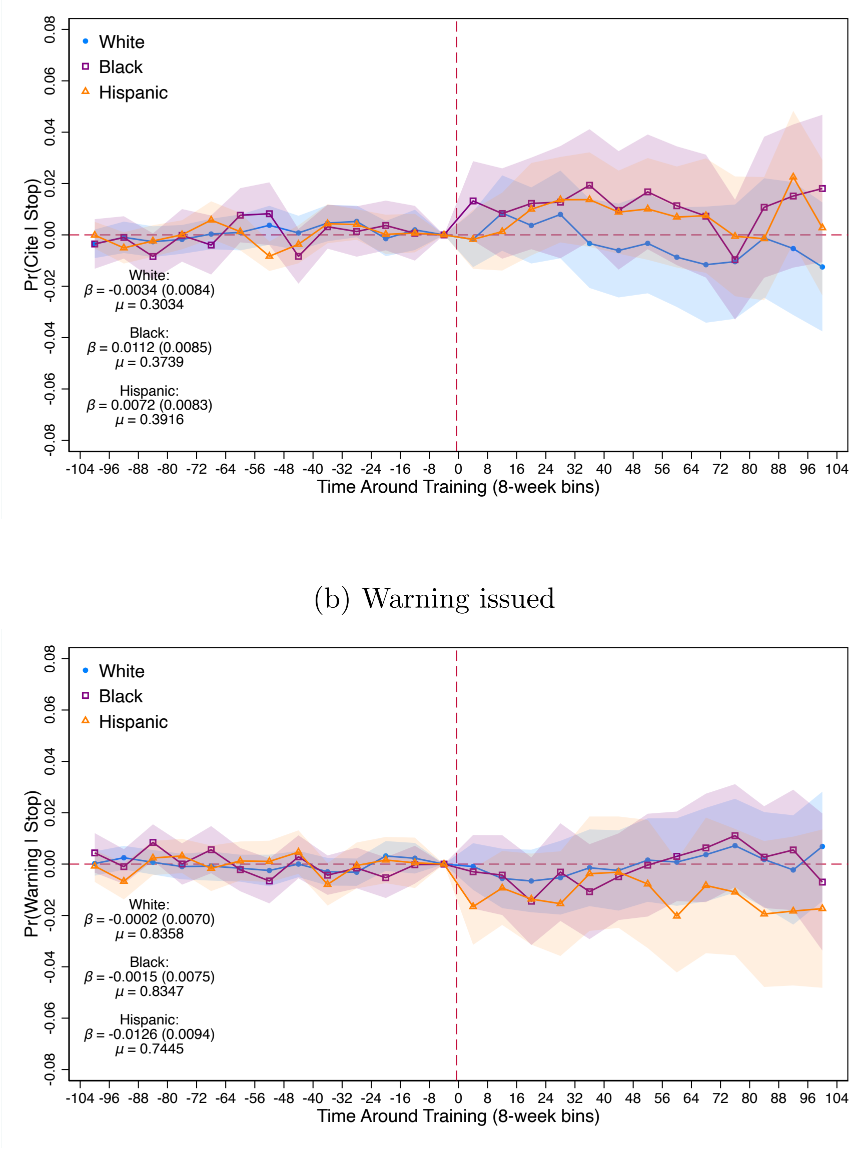


(b) Warning issued



*N*otes: Same as figure 1 except using on the different dispositions of stops and the outcome of interest is the weekly number of stops of white motorists. Panel (a) uses the rate of citation conditional on a stop of white motorists and panel (b) uses the rate of warnings conditional on a stop of white motorists.

Figure A-10: Effects of cultural diversity training on stop dispositions

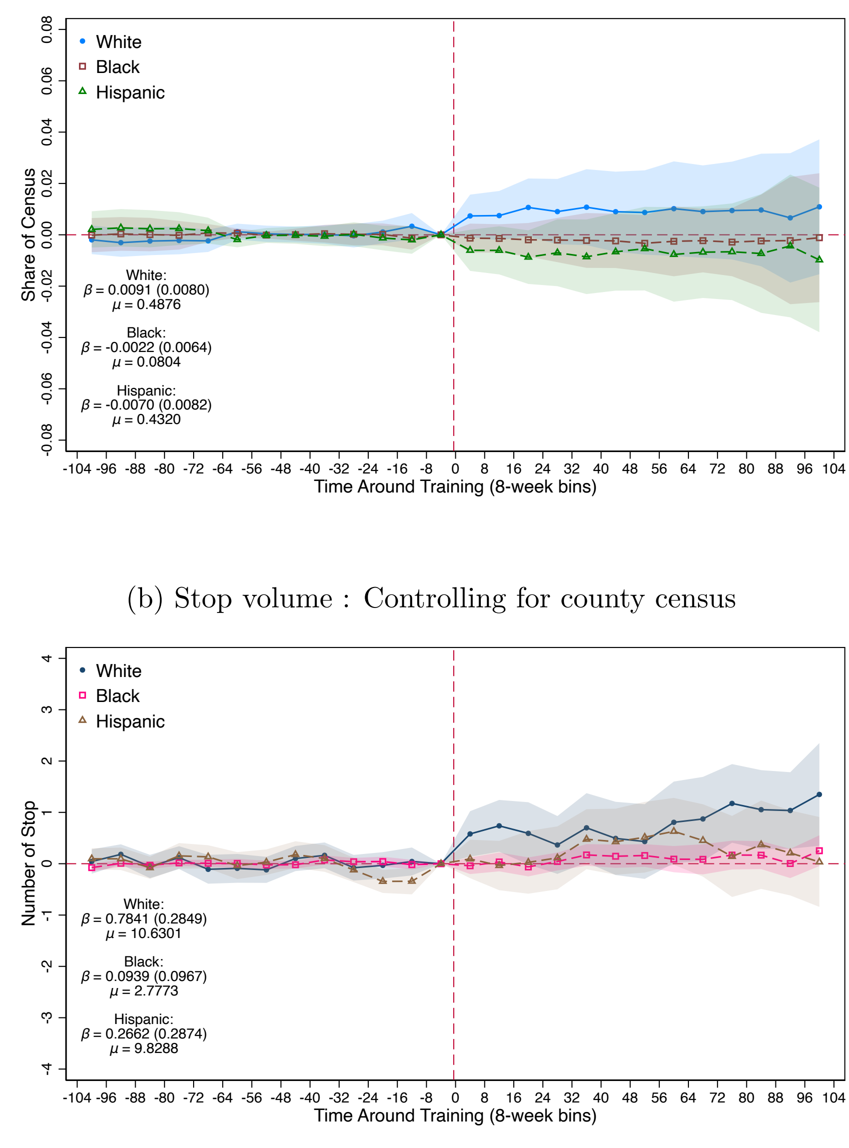


(a) Citation issued

(b) Warning issued

*N*otes: Same as figure A-9 except that the outcome of interest is shown separately by motorist race.

Figure A-11: Effects of cultural diversity training on stop locations

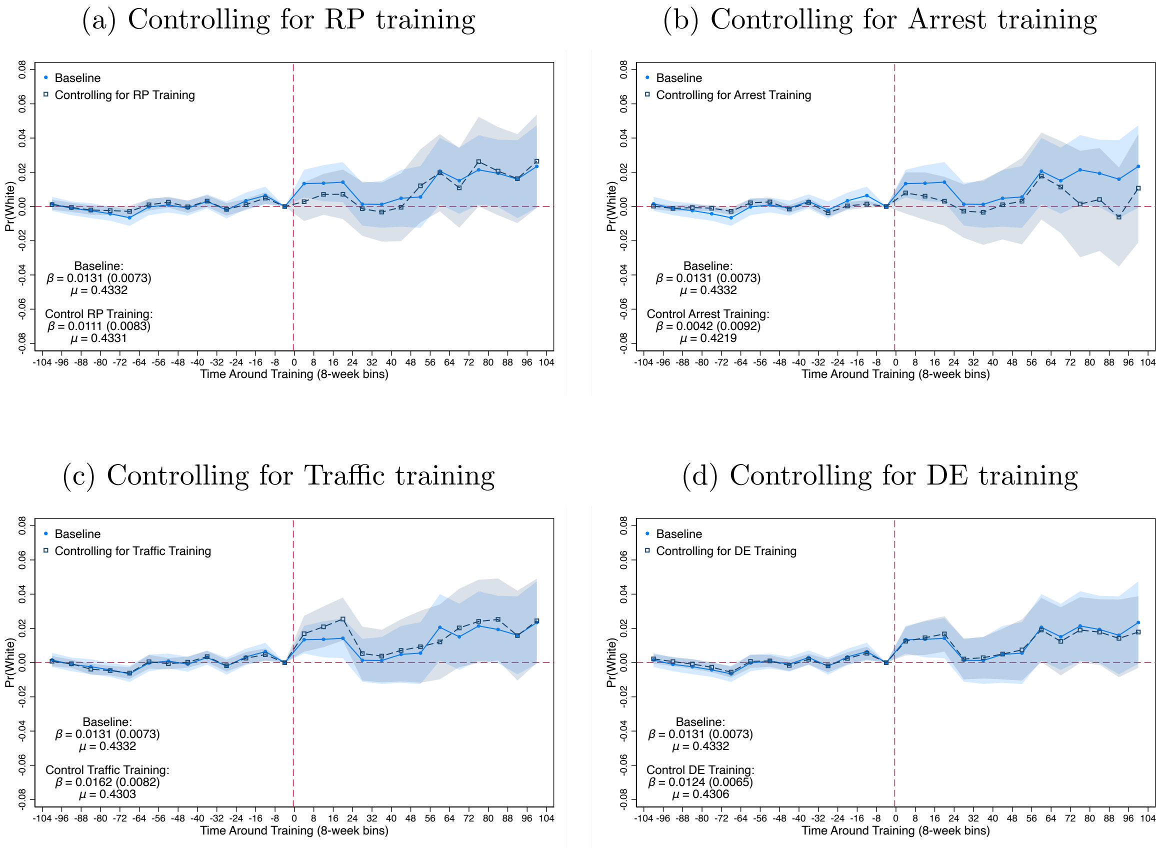


(a) Racial composition at county level

(b) Stop volume : Controlling for county census

*N*otes: Same as figure 6 except that the outcome of interest is shown separately by motorist race.

Figure B-1: Effects of cultural diversity training on the racial composition of stops (White), controlling for the timing of other trainings


(a) Controlling for RP training

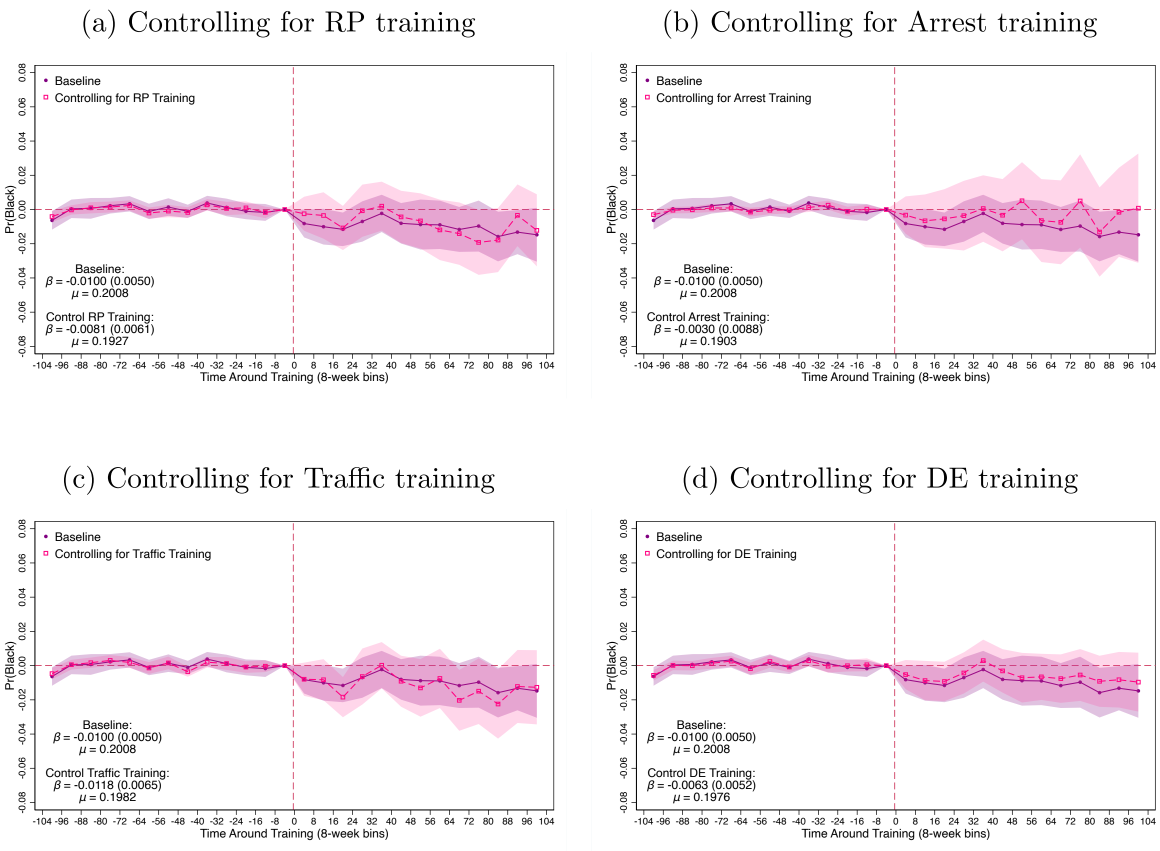
(b) Controlling for Arrest training

(c) Controlling for Traffic training

(d) Controlling for DE training

*N*otes: Same as figure 3 except that the outcome of interest is the share of officer's stops in a given week that are of white motorists.

Figure B-2: Effects of cultural diversity training on the racial composition of stops (Black), controlling for the timing of other trainings



(a) Controlling for RP training

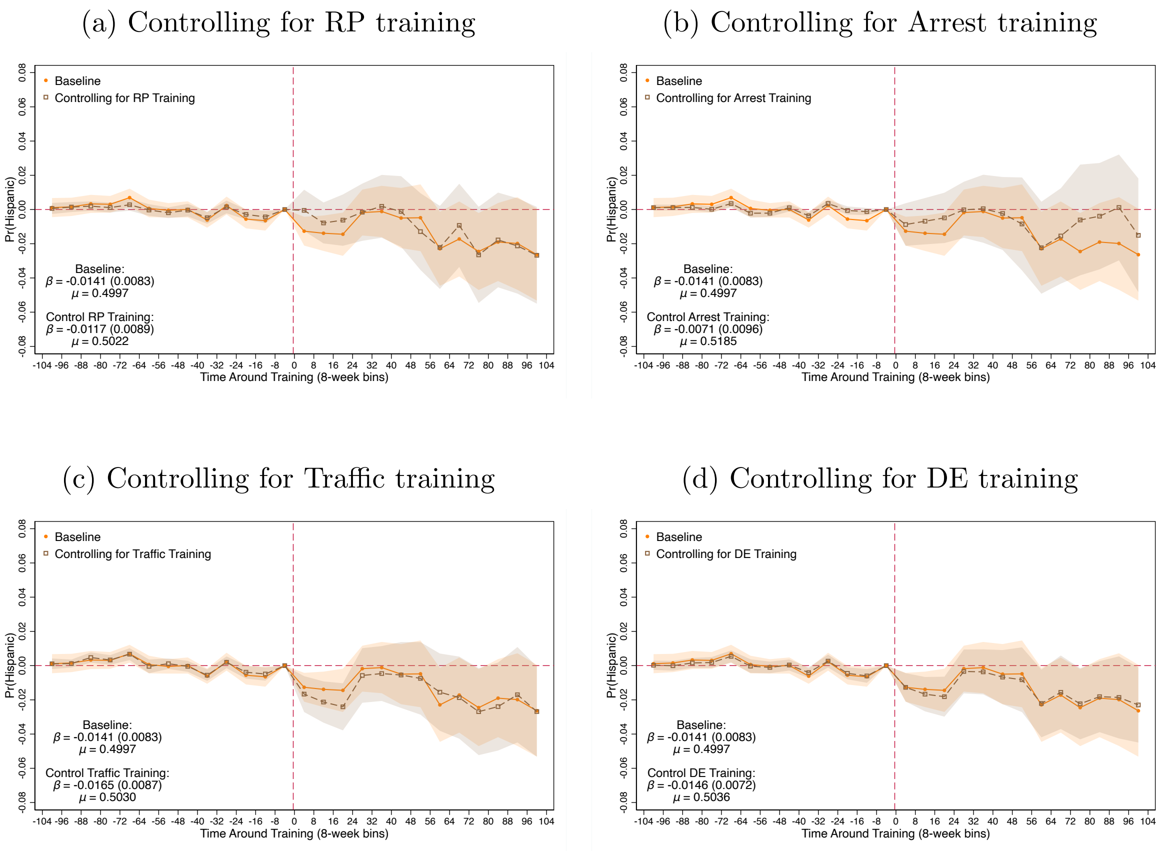
(b) Controlling for Arrest training

(c) Controlling for Traffic training

(d) Controlling for DE training

*N*otes: Same as figure 3 except that the outcome of interest is the share of officer's stops in a given week that are of Black motorists.

Figure B-3: Effects of cultural diversity training on the racial composition of stops (Hispanic), controlling for the timing of other trainings



(a) Controlling for RP training

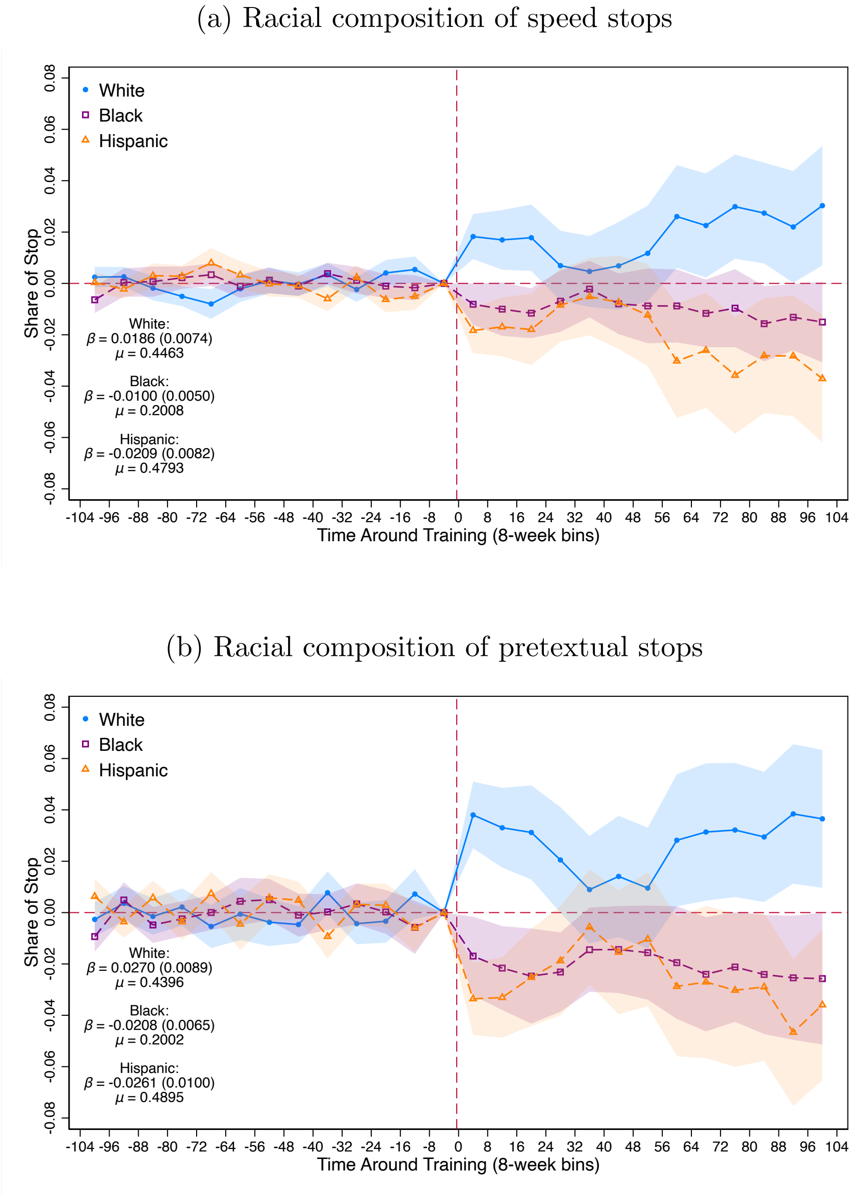
(b) Controlling for Arrest training

(c) Controlling for Traffic training

(d) Controlling for DE training

*N*otes: Same as figure 3 except that the outcome of interest is the share of officer's stops in a given week that are of Hispanic motorists.

Figure B-4: Effects of cultural diversity on racial composition of stop subtypes



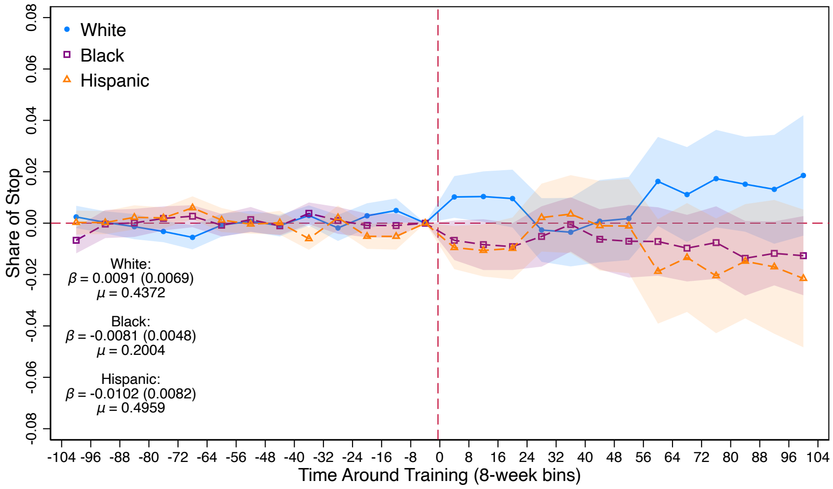
(a) Racial composition of speed stops

(b) Racial composition of pretextual stops

*N*otes: Same as figure 5 except that the outcome of interest is the race share of an officer's pretextual or speeding stops in a given week, separately by motorist race.

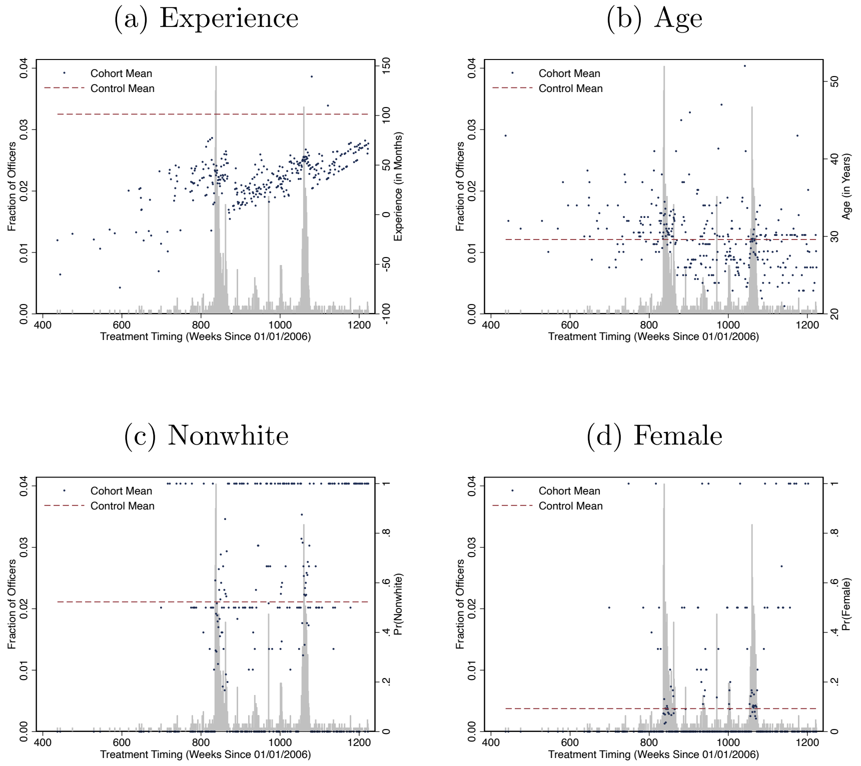
Figure B-5: Effects of cultural diversity training on stop locations

(a) Racial composition of stop : Controlling for White share of county census



*N*otes: Same as figure 6 except that the outcome of interest is the share of officer's stops in a given week, separately by motorist race.

Figure C-1: Event study cohorts


(a) Experience

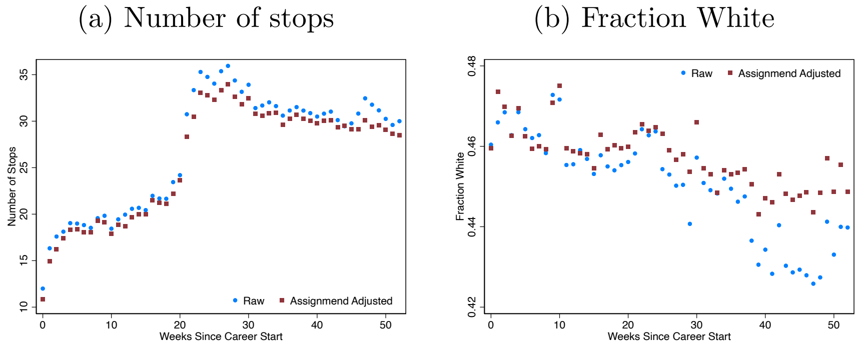
(b) Age

(c) Nonwhite

(d) Female

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

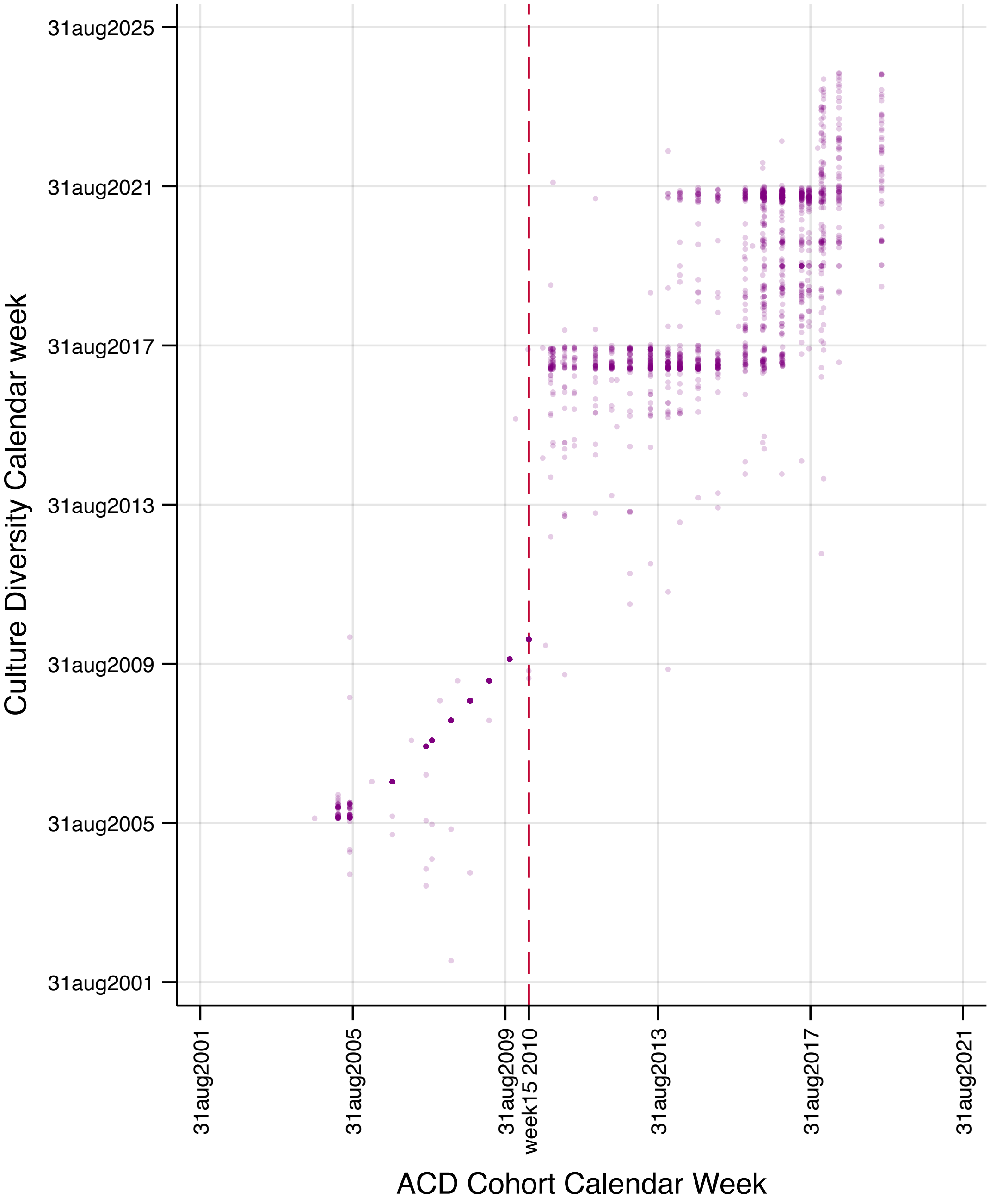
Figure C-2: Experience profiles for early-career officers


(a) Number of stops

(b) Fraction White

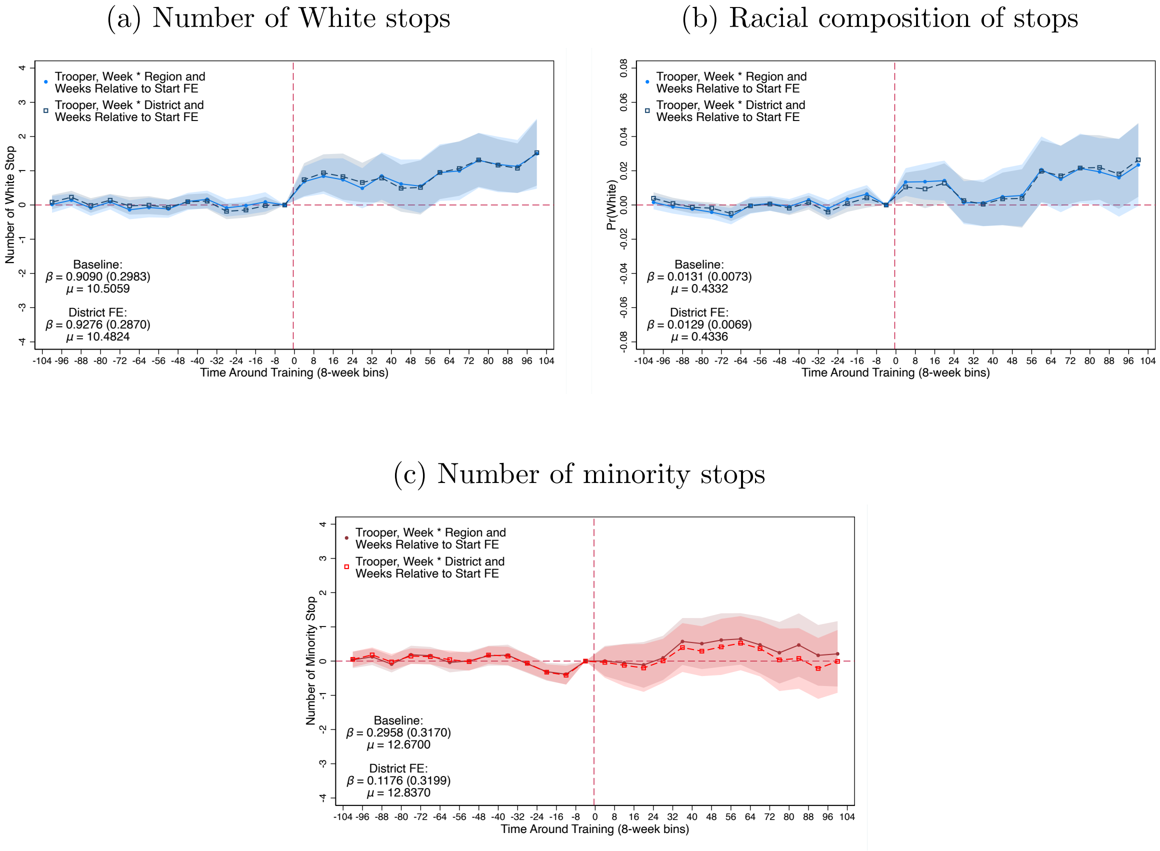
*N*otes: This figure reports the average the number of stops each week and the average share of stops that are of white motorists each week by week since career start using our analysis sample of early-career officers. In each panel, blue circles report raw averages, while red squares report averages which are adjusted for officer assignments.

Figure C-3: Officer cohort



*N*otes: This figure reports a scatter plot of culture diversity training timing and academy cohort timing. Each purple dot represents each officer. Dot color intensity reflects the number of officers overlapping in cohort.

Figure C-4: Robustness: varying fixed effects


(a) Number of White stops

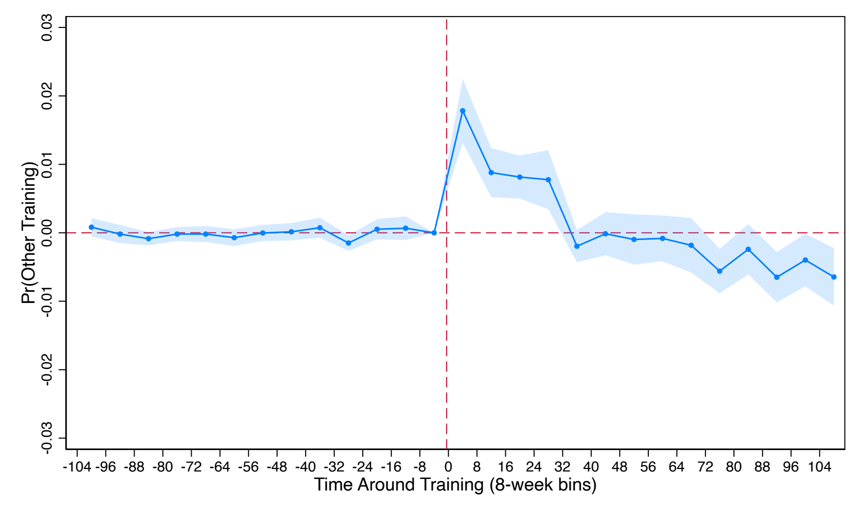
(b) Racial composition of stops

(c) Number of minority stops

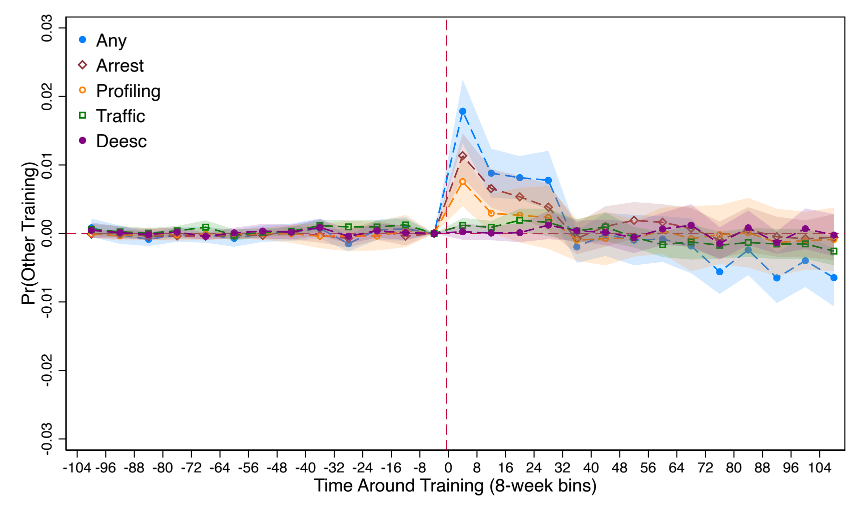
*N*otes: This figure reports imputation-based event study estimates from our baseline specification as well as a specification using different sets of fixed effects. The outcome of interest are an officer's number of stops of white in a given week (panel (a)), the share of officer's stops in a given week that are of white motorists (panel (b)) and an officer's number of stops of Black and Hispanic in a given week (panel (c)).

Figure C-5: Relationship between timing of cultural diversity training and other trainings

(a) Any other training

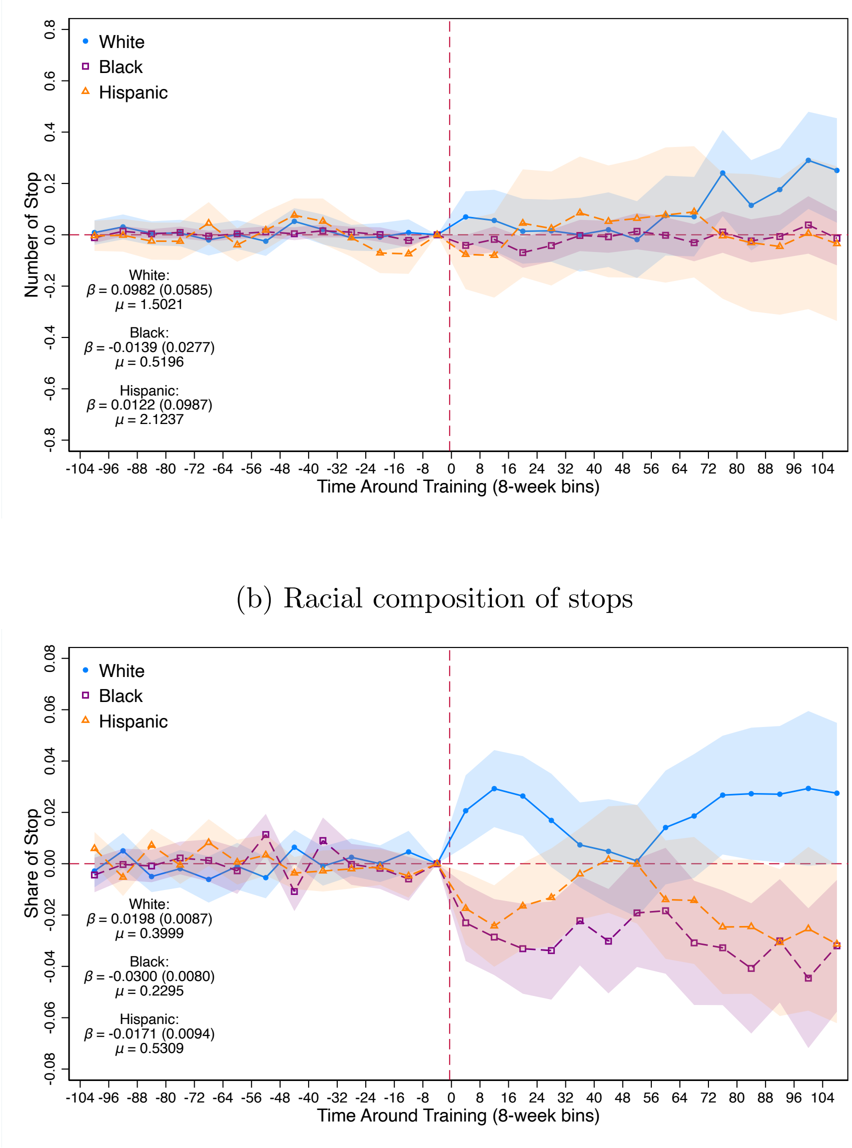


(b) By training type



*N*otes: This figure presents estimates from our baseline event study approach where the outcome of interest is whether took another training in a given week. Panel (a) reports estimates for any other training and panel (b)is shown separately by the type of training.

Figure C-6: Effects of cultural diversity training on equipment stops

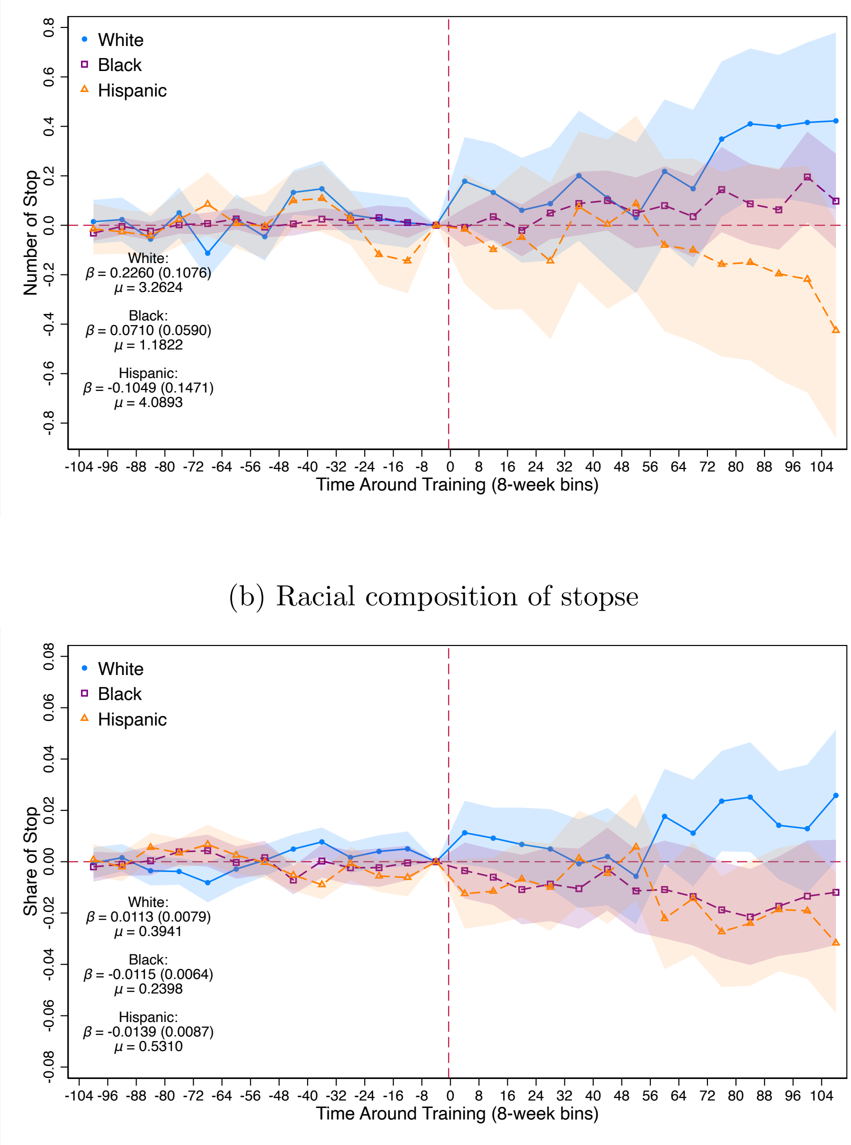


(a) Stop volume by motorist race

(b) Racial composition of stops

*N*otes: Same as figure 1 except using on the subset of stops for equipment violations, separately by motorist race.

Figure C-7: Effects of cultural diversity training on administrative stops

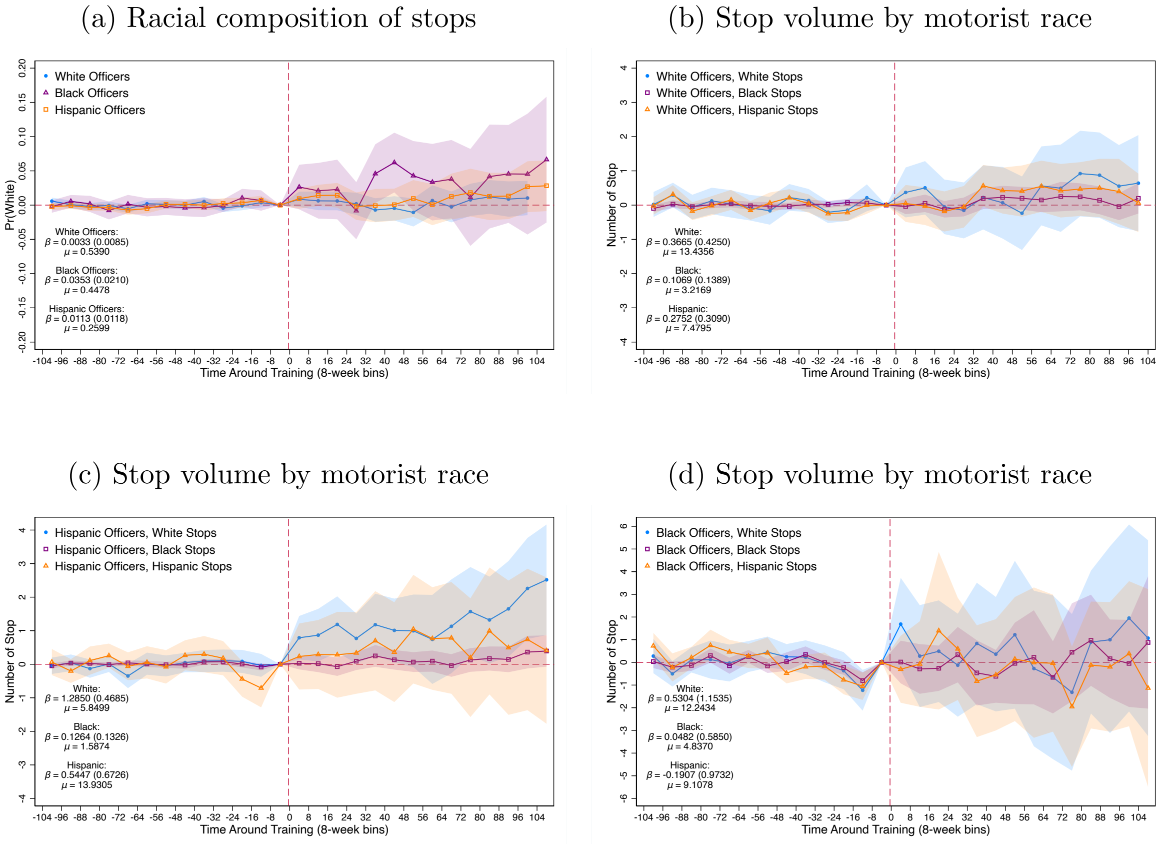


(a) Stop volume by motorist race

(b) Racial composition of stopse

*N*otes: Same as figure 1 except using on the subset of stops for administrative (i.e., paperwork) violations, separately by motorist race.

Figure C-8: Heterogeneous effects by officer race


(a) Racial composition of stops

(b) Stop volume by motorist race

(c) Stop volume by motorist race

(d) Stop volume by motorist race

*N*otes: Panel (a) is the same as figure 4 panel (b) except that we show estimates separately by officer race. Panels (b), (c) and (d) is the same as panel (a) except the outcome of interest is an officer's number of stops in a given week, separately by motorist race.

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 THP 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.[16]

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.[17]

At the end of the matching procedure, we can obtain demographic information for 4,933 officers. We arrive at our final analytical sample by excluding officers who make their first

[16] 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.

[17] 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.

traffic stop before 2004 and their last traffic after 2019. After dropping officers with academy after first drop, dropping officers who take academy more than one year before starting and keeping officers who survive to BPO and need intermediate, we can keep 2,329 officers. Based on an institutional change of when DPS officers recieved training, we further restrict our sample to officers who complete the academy after week 15 of 2010. Our final sample of 1,662 officers which 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 × 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.

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:[18]

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

From our perspective, this solution to the well-documented issues with canonical two-way fixed effects estimation of event studies is particular appealing because it easily accommodates a more complex fixed effects structure than simple two-way fixed effects estimation, which is necessary in our setting to address the fact that officers 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 two-way fixed effects with two exceptions. The first is that we typically also condition on assignment fixed effects. In most specifications, we include officer 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 (district × 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

[18] 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.

event time coefficient for 8-week groups instead of for individual weeks, primarily to increase precision. 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. Note that this aggregation is *not* the same as aggregating the data further into an 8-week panel and estimating event studies using such a panel, because our approach preserves and leverages the variation in the timing of treatment at the week-level.

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 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 officer-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.[19] 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 as our estimates of the standard error. Throughout, we use 100 bootstrap iterations for inference.

[19] 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.

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. All columns present estimates conditioning on hour × year, day of week × year and county fixed effects, and the standard error is clustered at hour × year, county × year level.

Table D-1: Veil of darkness estimates

	(1) Black	(2) Hispanic	(3) Any Minority
daylight_x_pretrain	0.0167***	0.0156**	0.0198***
	(0.00374)	(0.00515)	(0.00445)
daylight_x_postrain	0.00579	-0.00367	-0.00251
	(0.00559)	(0.00638)	(0.00674)
_cons	0.175***	0.456***	0.534***
	(0.00191)	(0.00330)	(0.00308)
N	588696	893265	1049155
r2_a	0.139	0.343	0.218
F	10.06	8.315	13.82
p	0.000354	0.00111	0.0000376

Standard errors in parentheses

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