Proud to Belong: The Impact of Ethics Training on Police Officers in Ghana*

Donna Harris† Oana Borcan Danila Serra
Henry Telli Bruno Schettini Stefan Dercon

June 9, 2025

Abstract

We study the impact of an ethics and integrity training program for police officers in Ghana using a randomized field experiment. The training, grounded in theories of identity and motivation, aimed to re-activate intrinsic motivations and foster a shared identity as "Agents of Change." Survey data collected 20 months post-training show lasting improvements in officers' values, attitudes toward citizens, along with reduced willingness to engage in unethical behavior, as measured by an incentivized cheating game. District-level administrative data on filed cases, convictions and fines, available for a subsample of districts, show improvements in field behavior, but only in the short-run. Our findings suggest that identity-based ethics training can be a valuable tool for improving integrity among police officers, but periodic reinforcement and organizational support are likely needed to sustain long-term impact.

Keywords: Ethics training, traffic police, experiment.

JEL classification codes: K42, M53, D73, H76

*We thank the Ghana Police Service, and in particular, Superintendents Samuel Sasu-Mensah and Jerome Kanyog, for their continuous support. We thank Paul Collier for facilitating the fieldwork, and Michelle Craske and Alan Stein for their help with the design of the training modules. We are grateful to Chris Blattman, Alex Coutts, Jennifer Doleac, Marcel Fafchamps, Andrew Foster, Emily Owens, Imran Rasul, Abu Siddique, and participants in the NBER Summer Institute, the BREAD/IZA conference, and various seminars, for useful comments. We thank Daniel Gomez-Vasquez for excellent research assistance. We are indebted to the late Samuel Kweku Yamoah, our brilliant field manager and friend. We acknowledge financial support from the Economic Research on Identity, Norms, and Narratives Network (ERINN). The project received ethical clearance from the University of Oxford, and is registered in the AEA Registry: AEARCTR-0003631.

†Harris: University of Oxford, donhatai.harris@conted.ox.ac.uk. Borcan: Department of Economics, University of East Anglia, o.borcan@uea.ac.uk. Serra: Corresponding author, Department of Economics, Texas A&M University, dserra@tamu.edu. Telli: International Growth Center, Ghana, henry.telli@theigc.or. Schettini: Ministry of the Economy, Federal District, Brazil. Dercon: Blavatnik School of Government and Department of Economics, University of Oxford, stefan.dercon@bsg.ox.ac.uk.

1 Introduction

A fundamental challenge in many countries is the persistent underperformance of government organizations in delivering public services. Recent research has examined how recruitment, monitoring, and incentives affect public sector performance (e.g., Ashraf et al., 2014; Banuri and Keefer, 2016; Cassar and Armouti-Hansen, 2020; Ferraz and Finan, 2008; Fisman and Miguel, 2007; Hanna and Wang, 2017).[1] Other studies highlight the role of management practices such as goal-setting and supervision (Bloom et al., 2020; Dunsch et al., 2017; Rasul and Rogger, 2018). These approaches are often difficult to apply to the police, as officers tend to operate with broad discretion and limited oversight. In addition, compared to other public sectors, police outcomes are harder to observe and quantify, making it challenging to design effective interventions and evaluate their impact on performance.

This paper introduces a novel approach to improving police officers' performance by targeting identity and intrinsic motivations. We design and experimentally evaluate a training program for police officers in Ghana that aims to realign officers' sense of professional identity with the public service mission of their organization. Unlike most prior work, which focuses on selection into mission-driven jobs or external performance incentives, our intervention seeks to shift intrinsic motivations among incumbent public servants operating in a highly corrupt environment.

The setting of our study is both policy-relevant and empirically challenging. The police, in Ghana and in most of the world, are perceived as highly corrupt by the public.[2] Officers often operate within a highly hierarchical structure, which leaves lower-ranking officers powerless against prevailing norms. Accountability systems that rely on citizen complaints tend to be ineffective due to officers' discretionary power, which can result in the imprisonment or physical harm of complaining citizens. As a result, traditional mechanisms to motivate officers by aligning their behaviors with the original public service mission (e.g., monitoring, supervision, and financial rewards) often fail. Similarly, managerial interventions, e.g., giving officers more autonomy, are unlikely to succeed.[3]

Given these challenges, we developed a two-day "Proud to Belong" training program rooted in theories of identity economics (e.g., Akerlof and Kranton, 2005, 2008; Bénabou and Tirole, 2011; Chen and Li, 2009), social psychology (Tajfel, 1974; Hogg et al., 2017), and

[1] For a review, see Finan et al. (2017). See also Ali et al. (2021); Avis et al. (2018); Banerjee et al. (2015); Borcan et al. (2014, 2017); Callen et al. (2020); Deserranno (2019); Dal Bó et al. (2013); Dhaliwal and Hanna (2017); Duflo et al. (2012); Khan et al. (2019); Linos (2018); Olken (2007).

[2] Transparency International's 2019 Africa Corruption Barometer, a survey of over 47,000 citizens in 35 African countries, identifies the police as the sector of the government perceived by the public as the most corrupt. See `https://images.transparencycdn.org/images/2019_GCB_Africa3.pdf`.

[3] For a comprehensive review of the economics of policing, see Owens and Ba (2021).

intrinsic motivation (Besley and Ghatak, 2005; Bursztyn and Jensen, 2017). The program encouraged officers to reflect on their original motivations for joining the police, define the traits of an ideal officer, brainstorm on field challenges and possible solutions, and generate a new collective identity as "Agents of Change." The intervention was reinforced through symbolic awards and social recognition (during an award ceremony held by the Police 9-month post training), but critically, it did not rely on performance-based rewards or sanctions.

We conduct our study in Ghana, a country where the police is consistently ranked as the most corrupt sector of the government (Ghana Integrity Initiative, 2017) and where about 60 percent of citizens think that most or all police officers are corrupt, and about one third of those who come into contact with the police have to pay a bribe (Pring and Vrushi, 2019). Traffic officers, in particular, are known to either purposely create roadblocks to extort bribes from law-abiding drivers, or to reduce the number and/or the frequency of checks on smugglers/traffickers in exchange for various sums of money (Foltz and Opoku-Agyemang, 2015; Norman et al., 2017; Tankebe, 2010).[4]

We implemented the intervention among traffic police officers in the Greater Accra region, covering all 33 police districts with a Motor Traffic and Transport Department (MTTD) in the region. This included the Central (Headquarters) MTTD district, which is significantly larger than the other districts, employing over 200 officers who typically attend official ceremonies and are sometimes deployed across the region, rather than being stationed in a specific geographical area. To minimize spillover effects, we excluded the Central district from the intervention but included its officers in the data collection process. The remaining 32 districts employ approximately 600 officers. We randomly assigned 21 of the 32 police districts to treatment. About half of the traffic police officers operating in the treatment districts received training between April and May 2019. Participation was based on officers' availability during the training weeks, which was determined by pre-existing MTTD duty rosters that were assigned quasi-randomly through a rota system.

The training program, which we called "Proud to Belong," required participation in two full days of training and was delivered by a research team member who had been a police officer in a different country, where he had led training programs.[5] The training was developed and implemented as a collaborative endeavor with the Police MTTD, which required officers'

[4]Even though the sum of money involved is usually small, this type of corruption is worrisome for two reasons. First, due to the frequency of potential encounters between traffic officers and citizens, this is the sector of the government with the highest utilization rate and therefore the greatest number of potential "corruption victims." Second, along with the judiciary, the police force is critical to law enforcement. Consequently, experiences of corruption in the police are likely to compromise the legitimacy of law and order institutions and promote a climate of distrust toward government officials and the state more generally.

[5]Bruno Schettini is a former Brazilian Police Inspector who came to the team with extensive expertise in police training.

participation. It required active in-person participation one day per week over two consecutive weeks. The first training day combined presentations and interactive activities, including role-playing exercises. Officers reflected on their motivations for joining the police, envisioned the qualities of an ideal officer, and collectively identified core professional values—such as honesty, respect, and cooperation. The session also provided communication tools and encouraged critical reflection on current behaviors and attitudes. The second day focused on team-building exercises to foster a sense of unity, shift beliefs about peer willingness to improve organizational culture, and promote a shared identity as "Agents of Change." To reinforce these messages, trained officers participated in follow-up WhatsApp groups and were formally recognized in an MTTD award ceremony in December 2019. Each officer received a certificate of completion, a letter of commendation from the Inspector General of Police, and a symbolic "Agent of Change" lapel pin.[6]

We collaborated with the MTTD also for the purposes of data collection. We were able to access police stations in October and November 2018 to survey all available officers, i.e., all officers present during the assigned week of data collection. We successfully collected data on nearly 500 officers, i.e., about 80 percent of all traffic police officers employed in the Greater Accra region.[7] The endline data collection was initially scheduled for April 2020 – one year after the training implementation – but was delayed because of the COVID-19 pandemic. Ultimately, we conducted the endline survey by phone in December 2020, approximately 20 months after the training, 12 months after the MTTD award ceremony, and over 2 years after the baseline data collection. We were able to re-survey 412 of the officers surveyed at baseline, with no differential attrition by treatment status. We also re-surveyed 133 randomly selected officers from the Central police district (i.e., about 50 percent of all Central officers).

Our pre-specified primary survey-generated outcomes are measures of officers' attitudes towards citizen, perceived corruption, tolerance for rule-breaking, and willingness to report misconduct by colleagues. In addition, our primary objective was to assess officers' propensity to engage in unethical behavior. Measuring such propensities through self-reported survey questions is inherently difficult, particularly in sensitive settings like policing. To address this challenge, we implemented an incentivized behavioral game, modeled after Abeler et al. (2014), designed to elicit dishonest behavior in a controlled and anonymous manner. Specifically, at the completion of the survey, while still on the phone, each officer was asked to flip a coin four times and report the total number of tails, knowing that each reported tail

[6]This component relates to interventions using social recognition to motivate public servants (e.g., Ashraf et al., 2014; Fracchia et al., 2023). However, unlike performance-based awards, this symbolic recognition was tied solely to participation in the training and aimed to reinforce the shared identity and commitment to integrity.

[7]We also surveyed almost all officers employed in the Central district.

would earn them 10 Ghanaian Cedi (GHS), for a maximum of 40 GHS if they reported four tails. This created an incentive to misreport the outcome of the coin flips.[8] Incentivized cheating games have become a widely used method to capture preferences for honesty and law-abiding behavior. These games have been applied in diverse settings, including among Rwandan citizens with varying historical exposure to centralized state authority (Heldring, 2021), and descendants of individuals subject to different pre-colonial institutions in Central Africa (Lowes et al., 2017). Prior research has shown that behavior in such games is predictive of real-world unethical actions, such as health worker absenteeism in India (Hanna and Wang, 2017), illicit drug possession among inmates in Switzerland (Cohn et al., 2015), academic dishonesty by students, also in Switzerland (Cohn and Maréchal, 2018), fare evasion on public transit in France (Dai et al., 2018), and misappropriation of unearned income in The Netherlands (Potters and Stoop, 2016).

Given that only about half of the police officers in the treatment districts participated in the training, we estimate both Intent to Treat (ITT) and Local Average Treatment Effects (LATE). Our data show that, nearly two years after the implementation of the training program, officers in treatment districts score significantly higher in our aggregate survey indexes capturing attitudes toward citizens, and officers' values and identity. We do not see evidence of robust impacts on our survey-generated measures of reporting of unethical behavior, monitoring of subordinates, and perceptions of corruption. Crucially, however, we find that the training program significantly lowered officers' likelihood to engage in unethical behavior, as measured by the incentivized cheating game. We also find no evidence of spillover effects between trained and untrained officers within treatment districts, suggesting that the impact of the training requires direct participation.

In order to examine whether changes in individual attitudes and propensity to act unethically led to changes in field behavior, we had originally aimed to follow and survey mini-bus drivers in the Greater Accra area, with the objective of measuring interactions with traffic police, using a methodology similar to that employed by Lameke et al. (2023) in the DRC. However, this became unfeasible due to the COVID-19 pandemic. As an alternative, we sought to obtain administrative data from the Motor Traffic and Transport Department (MTTD) on observable indicators of officers' field conduct. Ultimately, we secured district-level data covering a 60-month period (January 2017 to December 2021), including the number of cases filed by traffic officers, the number of cases sent to court, the number of convictions, and the total amount of fines imposed. Data on citizen complaints and potential disciplinary actions against officers were not available.[9] While limited, the administrative data we obtained re-

[8] Based on the wage data generated by our survey, the average hourly wage of a police officer in our sample was 12 GHS.

[9] We suspect that this data does not exist, or is not in a shareable format.

mains valuable, as it captures key aspects of officers' on-the-road activities. Positive impacts of the training program on these measures could reflect increased on-the-job effort, potentially driven by greater motivation, or a reduced tendency to substitute law enforcement with bribe-taking. For the latter, the assumption would be that corrupt officers prefer collecting bribes to issuing court orders, leading to lower numbers of filed cases, sentences, convictions and fines.

One limitation of our analysis is that values from multiple districts (including treatment and control districts, and adjacent areas) are often reported together. For this reason, we are able to use only data for a subset of 17 of the original 32 districts, i.e., those that allow us to keep control and treatment areas distinct. Despite the small sample size, and the district-level nature of the available outcomes (compounded by the fact that only half of the officers in treatment districts participated in the training), our analysis generates suggestive evidence of a positive and large impact of the ethics training on officers' behavior. Specifically, we see significant and substantial increases in filed cases and court orders, leading to more convictions and fines. These results are consistent with recent evidence from India by Conover et al. (2023), who find that improved street lighting increased ticketing, potentially due to greater officer effort or a reduction in bribery due to greater visibility (i.e., greater risk of being seen acting unethically).

Notably, the effect of the training on field behavior is concentrated in the first three months following participation. This temporary impact contrasts with the more sustained effects observed on individual attitudes and willingness to engage in unethical behavior, which persist even 20 months after the training. This pattern suggests a disconnect between private moral attitudes and beliefs, and behaviors if the field, which are partly shaped by social and institutional dynamics. One possible explanation is that the effects of the training require ongoing reinforcement, such as continued interaction with fellow trainees or refresher modules, to be sustained. This interpretation aligns with our finding that the first three months post-training, when WhatsApp group chats were most active, showed the strongest effects on field behavior. Additionally, we observed some improvement in field behavior in the two months following the award ceremony, which could have acted as an additional training reminder. A second possibility is that officers who initially applied the training's principles in the field encountered resistance from peers or superiors, especially since the training had a stronger impact on lower-ranked officers. These officers may have faced reprimands or other forms of pushback from higher-rank officers, as documented by Sánchez De La Sierra et al. (2024) in the DCR. Our data provide some support for this explanation: we see that (only) lower-ranked officers in treatment districts had a lower probability of being promoted in the 25 months between baseline and endline, as compared to their counterparts in control

districts. The third, more positive interpretation - which we cannot confirm or reject with our data - is that drivers' behavior may have changed in response to increased law enforcement in the treatment districts in the three months following the training, reducing subsequent opportunities for either law enforcement or corruption (or both) across both treatment and control districts.

Our study contributes to the growing literature on improving the performance of public sector providers, with a particular focus on law enforcement. Recent work has emphasized the role of police officers' behaviors and their responses to incentives, such as oversight, payment, and policies, particularly in the context of traffic policing. For example, monitoring with traffic cameras in India increased on-the-ground enforcement of tickets (Conover et al., 2023), body-worn cameras in Brazil reduced excessive use of force (Barbosa et al., 2021),[10] and doubling police patrols in the city of Bogota' in Colombia failed to reduce crime (Blattman et al., 2021). Our study departs from this literature by focusing on alternative types of incentives, i.e., training programs centered on ethics and identity, rather than financial incentives or increased monitoring.

Several studies have evaluated procedural justice training for police in the U.S., the U.K., Mexico and Colombia, showing improvements in officer-citizen interactions, such as reduced use of force (Owens et al., 2018; Wood et al., 2021), enhanced procedural fairness (Canales et al., 2025) and improved citizen trust in the police (Abril et al., 2023). Related efforts include soft-skills training in India that improved citizens' perceptions of police (Banerjee et al., 2021), and community policing interventions in developing countries, which yielded mixed results (Blair et al., 2021, 2019; Karim, 2020; Peyton et al., 2019). A recent study with Chicago police found that situational decision-making training reduced use of force and discretionary arrests (Dube et al., 2025). We are not aware of any prior evaluations of training programs that focus explicitly on ethics, identity, and a shared organizational mission among police.

The empirical evidence on the effectiveness of ethics training, more generally, is scarce. While a number of studies show suggestive evidence of a positive impact of training (Beeri et al., 2013; Kumasey et al., 2017; Park and Blenkinsopp, 2013; Warren et al., 2014), the observational nature of these studies limits our ability to draw conclusions on the causal effect of ethics training on attitudes and behaviors.[11] Closer to our approach is a growing body of

[10]For a review of studies evaluating the impact of body cameras on police see Williams Jr et al. (2021).

[11]For instance, Beeri et al. (2013) assess the impact of ethics training on public sector employees in Israel, yet the small sample size (N=108) and the lack of a control group, makes it difficult to draw conclusions on the causal impact of the program. Kumasey et al. (2017) face similar identification problems when examining the relationship between the adoption of an ethical code and the attitudes of public sector employees in Ghana. Borcan et al. (2023) study the impact of an integrity training for Ukrainian law students on their propensity to facilitate bribes in an incentivized game, and find that training is only effective when students

research showing that aligning employees' personal sense of purpose with their professional roles can enhance motivation and performance (Ashraf et al., 2024). Related evidence from community health workers finds that training programs emphasizing the mission of the job improve provider behavior and health outcomes in Pakistan (Khan, 2020), while lighter-touch video interventions on task significance in Guinea-Bissau had no effect (Fracchia et al., 2023).[12]

Overall, our findings suggest that well-designed, in-person training focused on intrinsic motivation, shared identity, and beliefs about collective action can meaningfully shift police officers' attitudes and behavior, even in settings marked by pervasive corruption and abuse of power. These results underscore the potential of ethics- and mission-driven approaches to complement traditional accountability tools, which are often difficult to implement in weak institutional environments.[13] At the same time, our findings highlight that sustaining behavioral change may require ongoing reinforcement through peer networks, refresher sessions, or periodic recognition. Moreover, since efforts by lower-ranked officers to implement ethical practices can be constrained by institutional resistance, aligning organizational structures to support reform is crucial for translating individual-level change into systemic improvements. More broadly, our results contribute to a growing literature on the role of identity and intrinsic motivation in shaping values and ethical behavior within organizations.

2 Context

Ghana is one of the fastest growing economies in Africa. In 2019, prior to the COVID-19 pandemic, the International Monetary Fund projected it to be the fastest growing economy globally.[14] In the last two decades, Ghana has also made remarkable achievements in terms of human development. It has, for instance, increased the adult literacy rate from 65 to 76 percent, nearly halved the infant mortality rate (from 60 to 34 per 1,000 live births), and reduced the percentage of people living in poverty ($1.90 a day headcount) from 35 to 13 percent.[15]

---

are informed that a majority of their peers are trained. For a review of issues related to the study of ethics programs or codes in organizations see, respectively, Kaptein (2015) and Kaptein and Schwartz (2008).

[12]However, an alternative intervention aimed at increasing provider social status through symbolic awards proved effective (Fracchia et al., 2023).

[13]While audits and top-down monitoring can reduce corruption (e.g., Ferraz and Finan, 2008; Olken, 2007), they frequently face implementation barriers in contexts with limited transparency and collusion across bureaucratic levels. Bottom-up accountability efforts also often fall short (Banerjee et al., 2010; Olken, 2007). See also Banerjee et al. (2012), Olken and Pande (2012), and Serra and Wantchekon (2012).

[14]Available at: `https://www.imf.org/en/Publications/WEO/Issues/2019/01/11/weo-update-january-2019`

[15]Available at: `https://databank.worldbank.org/source/world-development-indicators`.

One area where progress has stalled in Ghana is the curbing of corruption. The Transparency International's latest Corruption Barometer - a survey of 47,000 citizens in 35 African countries between 2016 and 2018 - shows that 35% of Ghanaians think that most or all government officials are corrupt. The percentage goes up to 59% when referring specifically to police officers. Further, data from the 2019 Afrobarometer, based on a survey of 2400 Ghanaians, reveal that about one third of the people who come into contact with the police have to give a bribe or a gift, or do a favor to police officers. This is confirmed by Tankebe (2010)'s survey of nearly 400 randomly sampled Ghanaian households, which shows that over half of the respondents witnessed an exchanged of money or promises between a police officer and a citizen. In an attempt to reduce bribery and increase overall job satisfaction, in 2010 police officers saw a doubling of their salaries. However, as Foltz and Opoku-Agyemang (2015) documented by following long-haul truck trips between Ghana and Burkina Faso, bribery actually increased as a result.[16]

Our study involves traffic police officers in the Greater Accra area. This is one of 11 distinct Police regions in Ghana, each headed by its own Regional Police Commander. The Accra region consists of 42 police districts operating in geographically contiguous areas, similar to neighbourhoods.[17] A total of 33 out of the 42 police districts have a Motor Traffic and Transport Department (MTTD) unit and, therefore, employ officers whose duties include monitoring traffic, enforcing penalties for traffic violations, advising and interacting with drivers. These are the focus of our study. The 33 MTTD districts include the Central district, which is a uniquely large district, with over 200 officers. In contrast, the number of officers in the other districts range from 2 to 45, with the average being 25, and the total across the 32 districts being around 600.

Officers are typically deployed to man the traffic within their district and are rarely sent to other districts to work, except if they are permanently transferred. They are assigned to specific locations, e.g., traffic lights and road intersections (henceforth duty posts) according to a centrally determined duty roster. They spend one or two weeks in the same location and with the same team of one to three other officers. The officers in the Central district have a broader range of duties, as they assist in official events, such as motorcades, and are often deployed to other districts to fill in for absent officers.

Crucially, in Ghana, every traffic infraction triggers a court order, requiring drivers to appear in court in person, either to settle fines or await trial. Specifically, when a driver is caught committing a traffic offence, say over-speeding, the police officer can decide to

[16]They document that trucks were stopped on average 16 times in the Ghana section of the route, and were asked to pay a bribe 80 percent of the times.

[17]They amount to over 100 police stations. For details see `https://police.gov.gh/en/index.php/accra-region/`.

let the person go with a warning or to take the person through the due process, which is as follows. As a first step, the officer takes the driver to the Motor Traffic and Transport District (MTTD) office (on the same day) to formally file the case. This could result in the car being impounded or the driver's driving licence being withheld by the police. Once at the district office, a final determination is made regarding whether or not the officer wants to take the case to court. If the officer decides to send the case to court, the driver is called to court typically within two weeks from the traffic violation (and often the following day). When the case is called, the judge decides the fine to be paid, unless the infraction is severe and requires a formal trial. This framework naturally introduces incentives for bribery transactions, as officers may accept (solicited or unsolicited) bribes at different stages of the process, i.e., in exchange for not taking the driver to the MTTD office, or for not filing a case, or for not sending the case to court.

We first approached the Ghana Police leadership in the Accra region in October 2017. In early 2018, we established a collaborative relationship with the Inspectorate General of Police (IGP)'s Head of Research. At that time, the Ghana Police was in the process of developing a *Transformational Agenda* aimed at reforming the police, to enhance accountability, strengthen officers' professionalism and improve relationships with the public. Within this context, our proposal to develop a new integrity training program for traffic police officers was well received by the IGP leadership. Our study was made possible by the continuous cooperation of the IGP, who facilitated the collection of the survey data, helped with the recruitment and invitation of selected officers to the training program, and implemented the award ceremony.[18]

3 The "Proud to Belong" training program

3.1 Design

A central challenge in designing our training was how to shift individual attitudes and behaviors in a context where organizational norms may reward corruption and punish integrity. We drew on the literature on identity in organizations and group behavior (Akerlof and Kranton, 2005, 2008; Akerlof, 2016), which highlights the power of identity alignment and collective action. When employees see their values reflected in the organization's mission, performance tends to improve (Besley and Ghatak, 2005; Cassar and Armouti-Hansen, 2020). Yet, in

[18]Building and maintaining the relationship with the police required consistent communication mainly through personal interactions during numerous visits at the Police Headquarter and several follow-up phone calls, text messages, requests for meetings. For an account of the difficulties of doing research with the police in Ghana, see Sowatey and Tankebe (2019).

settings where informal norms contradict official goals, reactivating the aspirational identity officers had upon joining the force—and creating a new shared identity—may be key. Inspired by this framework, we aimed to (re)ignite officers' intrinsic motivations and cultivate a new collective identity: that of the "Agent of Change."

The training was developed in consultation with mental health and policing experts, including Professors Alan Stein (leading psychiatrist at the University of Oxford) and Michelle Craske (Professor of Psychology, Psychiatry and Biobehavioral Sciences, at the University of California, Los Angeles), and former Brazilian Police Inspector Bruno Schettini, who led a large number of training programs for the Bazilian Federal Highway Police. The program was reviewed and endorsed by Superintendent Samuel Sasu-Mensah of Ghana's Motor Traffic and Transport Department (MTTD).

The training consisted of two days, with one week gap in between. The purpose of this one-week gap was to let the trained officers reflect and internalize what they had learned during the first day. The training was delivered by Inspector Schettini to all officers in both days. Day 1 focused on personal identity and values. Officers reflected on their motivations for joining the force, defined the qualities of an ideal officer, and practiced communication strategies through role-plays. Day 2 emphasized collective identity and team-building. Officers collaborated on creative exercises, watched a documentary on reform in another police force, and discussed how to replicate such change. They set personal S.M.A.R.T. goals and concluded the day with a unity dance inspired by the Māori Haka, symbolizing their commitment to change. Further details on the training modules can be found in Appendix B. Details on the implementation of the program are provided in Section 3.2.

Immediately after the training was completed, we set up a WhatsApp group for each training cohort based on the day that they participated in the training (Monday, Tuesday groups, etc) in order to reinforce the intervention. For six months following the completion of the training, approximately once a month, the groups were sent (the same) inspiration messages, quotes and reminders of their values and mission by Inspector Schettini, who moderated all the groups. We do not have data on the WhatsApp communications. The purpose of the chats was to allow officers to communicate freely, absent any external monitoring. For this reason, we decided not to ask for subjects' consent to access their private communications, and not to use this source of data in the empirical analysis.

Finally, about eight months after the conclusion of the program, the MTTD Police held an award ceremony for the trained officers, where they all received a certificate and a symbolic "Agent of Change" pin aimed at further reinforcing their new identity as "Agents of Change." See Figure A1 in the Appendix for a picture of the pin.

3.2 Randomization and implementation

We evaluate the "Proud to Belong" training program through a randomized control trial. In 2018 we obtained the approval from the IGP to survey all MTTD officers in the Greater Accra area. We were later granted the MTTD's full collaboration in randomly selecting police districts that would participate in a new training program aimed at strengthening professionalism and accountability in the police force. We discussed the research design and collaborated with the IGP's research department in creating the baseline survey and planning the implementation of the training. Ethics approval was obtained from the University of Oxford.

Due to the nature of the interactions between police officers in a given district (they often work in a team), any within-district randomization of our treatment would have led to contamination to non-treated officers. For this reason, we decided to implement a clustered randomized trial design, with 32 out of the 33 MTTD police districts as clusters. We excluded the Central district from the randomization (and, therefore, the main study) due to its uniquely large size and its flexibility in deploying officers across other districts.[19] Of the 32 MTTD districts in the sample, we randomly selected 21 to participate in the training program.[20]

In total, at the time of our baseline survey the 21 districts which were selected to participate in the training (i.e., Treatment districts) employed around 414 officers and the 11 districts assigned to the Control group employed around 205 officers.[21]

We decided to over-sample the Treatment districts because we knew that participation in the training would be conditional on officers' availability in the two weeks we had scheduled for the implementation of the program. Given that the training could not cause any disruption to the officers' daily duties, only officers that were not scheduled to be at duty posts during the two weeks of the training could participate. We anticipated that this would be the case for about half of the officers in the Treatment group. Duty rosters are based on a rota system where officers are randomly assigned to duty posts, and remain in their duty posts for an entire period of one or two weeks. The duty posts for the officers in our sample were decided independently from the implementation of our program. Therefore, we consider the selection of the officers to be trained within the Treatment districts as good as random. An extensive set of covariates collected at baseline confirms that trained and untrained officers

[19] We have, however, administered the baseline survey to all Central district officers, and the endline survey to nearly half of them (randomly selected). This allows us to both compare the characteristics of the officers in Central district and those in the Control MTTD districts, and conduct robustness checks where we include the Central district officers in our control group.

[20] The remaining 11 districts were made aware that they may benefit from the program at a later time.

[21] We computed these numbers based on the duty rosters for the month preceding the baseline data collection.

in Treatment districts are not significantly different from each other (Table 1, columns (4)-(6)).[22]

We implemented a baseline survey of police officers in all districts, including the Central district, in October 2018, through face to face interviews at the officers' places of work. Given the officers' assigned duties, teams of four enumerators visited each MTTD district office multiple days during a given week, to make sure that all officers in that district were surveyed. Participation was voluntary but highly encouraged by the Inspector General of the Ghana Police Service. Overall, at baseline we surveyed a total of 479 officers, amounting to about 80% of all officers stationed in such districts.

In April 2019, the officers selected to be trained were contacted and invited to participate in the program, on a given day of the week, directly by the IGP through a letter reading "IGP directs the following MTTD personnel below should attend a training program on *Strengthening professionalism and integrity of the Ghana Police Service*." Nearly half (151) of the officers that we had surveyed at baseline in Treatment districts participated in the training. A further 17 officers from these districts, who were not surveyed at baseline, participated in the training.

The training took place in the two weeks commencing 29th April and 6th May 2019, over four days in each week. Each officer participated in one day of training in week 1 (Monday, Tuesday, Thursday or Friday), and one day in week 2, and was assigned the same day in both weeks, i.e., participants in a given day in week 1 returned on the same day in week 2. This assured that the same officers – between 30 and 45 – coming from multiple districts participated in any given day of training.

Our initial plan was to collect the endline data through face-to-face interviews about one year after the training, in Spring 2020, and to use the same survey instrument as in the baseline data collection. However, in March 2020 we halted the ongoing training for the endline data collection in response to the COVID-19 pandemic. Due to fieldwork restrictions, we then decided to implement a shorter endline survey, which included an incentivized cheating game, in December 2020, nearly 20 months after the training. At endline, officers were surveyed at the time that was most convenient to them. Enumerators first called officers to arrange a day and time for the phone interview. They then followed up on the agreed upon day to conduct the survey. We were able to successfully re-survey 86% (i.e., 412) of the officers that we had surveyed at baseline more than two years earlier. We do not see any significant differences in attrition rates between Treatment and Control districts, as shown in Table A1 in the Appendix.

We provide details on the baseline and endline surveys, as well as the incentivised cheating

[22]We compare trained and untrained officers in detail in Section 5.

games, in Section 4 below.

4 Data and empirical strategy

4.1 Survey data and incentivized cheating games

Both at baseline, in October 2018, and at endline, in December 2020, we collected survey data and implemented incentivized cheating games with traffic police officers. Due to the COVID-19 pandemic, the methods of implementation and the data collection instruments are different. In Appendix C, we provide details on how the pandemic caused us to deviate from our pre-analysis plan (PAP).

We conducted the baseline survey in person, and the endline survey by phone. At baseline, we also conducted an incentivized cheating game relying on the rolling of a dice. At endline, we shortened the survey and conducted a different cheating game, which was more suitable for phone implementation. We describe the baseline survey and games in Section 4.1.1. We focus on the endline data, and describe specifically our survey-generated outcome variables and endline cheating game in Section 4.1.2.

In addition to the survey data and the incentivized game, we were able to obtain district-level administrative data on filed cases, filed sent to court, convictions and fines issued to drivers for a subset of the districts. We describe these data and analysis in Section 6.

4.1.1 Baseline data

The baseline survey relied on a comprehensive set of questions aimed at measuring the officers' demographics, duties and work environment, their satisfaction with resources and different aspects of the job, and their experiences while interacting with each other and with private citizens, with a special focus on corruption and, more generally, own and others' unethical behaviors. The survey was designed partly with the aim of informing the Ghana Police, and partly to generate baseline measures of officers' attitudes, perceptions and experiences of corruption, as well as their relationship with citizens, and their beliefs about values associated with being a police officer and effective ways to fight corruption.

Importantly, we wanted to measure officers' baseline propensities to engage unethical behavior, which is challenging to do through survey questions. Following previous research showing that cheating in incentivized games predicts unethical behavior in the field (e.g., Dai et al., 2018; Hanna and Wang, 2017; Potters and Stoop, 2016), we decided to include an incentivized cheating game. One concern we had at the design stage was that, since we would be implementing the baseline survey at the officers' place of work, participants might

feel observed or monitored by the research teams or their superiors. The literature shows that individuals' preferences for being *seen* as honest are important, with subjects cheating less when the true outcome is observable to others (Abeler et al., 2019). For this reason, we chose to implement Kajackaite and Gneezy (2017)'s Mind Game, rather than a standard dice roll game where subjects roll a dice, report the number and are paid higher amounts of money for higher numbers, as in Fischbacher and Föllmi-Heusi (2013). The game consists in asking study participants to think of a number between 1 and 6 and then roll a die in private. Subjects then have to report whether the number they rolled matches the one in their mind. If they report a match, they earn a monetary payoff; if they do not, they earn nothing. It has been shown that in the Mind Game, subjects' concerns about being identified as a liar are low, and subjects are more likely to cheat than in other dice rolling games, especially when monetary incentives are large (Kajackaite and Gneezy, 2017).

Teams of four enumerators visited each MTTD district office multiple days per week, with the aim of conducting surveys of four officers at a time, at different time slots during the day. Each time slot lasted two hours, starting at 9 am and ending at 5 pm. Participating officers were informed about the survey by the IGP and scheduled to be interviewed on a given day and time slot. A maximum of 12 officers were interviewed in the same office in any given day.[23] Interviews took place in private rooms or outdoor space set up with marquee. At the conclusion of the survey, officers participated in the Mind Game. Each officer was given a die and a cup to be used to roll the die in private. Officers then had to state on a form whether the number they rolled matched the number in their mind. Checking "YES" on the form earned them 40 GHS. Based on the wage data generated by our survey, this corresponds to the average wage paid to officers for three hours of work.[24] During the session, officers were identified through a randomly assigned player number, and were never asked to state or write down their names. They were paid in private through mobile credit sent directly to their phone on the same day.

4.1.2 Endline data and outcome variables

Our endline data are generated from a survey and an incentivized cheating game implemented over the phone. The COVID-19 pandemic forced us to postpone the data collection to December 2020 (from April 2020) and to substantially shorten our survey instruments. In

[23]Each day, and in each district, four officers were scheduled to be interviewed from 9 to 11 am, four officers from 12 to 2pm, and four officers four from 3 to 5 pm.

[24]Subjects also played a second cheating die-rolling game, originally designed by Weisel and Shalvi (2015), which required sequential play between pairs of individuals. However, since officers only played in one of the two roles in the game, and, due to logistical complications caused by the COVID-19 pandemic, we were unable to conduct a similar game as part of the endline survey over the phone, we do not include this game in the analysis.

particular, we edited our baseline questionnaire to contain a restricted number of questions.[25]

As specified in our pre-registration, our primary outcomes of interest are attitudes of officers regarding: 1) unethical/corrupt behaviour by members of the police force, including their own propensity to engage in such behavior; 2) their relationship with the citizens. We further specified that we would generate survey measures of "sentiments towards the citizens, perceptions of corruption, tolerance for bending/breaking rules, propensity to notice and report unethical behavior of colleagues" as well as measures of propensity to engage in illegal behavior through incentivized behavioral games.[26] As a result, we designed the endline survey to allow us to measure five primary outcomes: 1) values, identity and beliefs associated with the job of a police officer, 2) the reporting of unethical conduct, 3) the monitoring of lower-rank officials (if the respondent manages a team of officers), 4) perceptions and experiences of corruption, and 5) attitudes toward citizens.

Since we have numerous outcome variables, and multiple measures of the same outcomes, one concern may be that multiple hypothesis testing could lead to a high rate of false positives. To address this concern, we compute standardized indices for groups of similar outcomes. Specifically, the *Values and Beliefs Index* captures officers' views on professional qualities, belief in changing norms, self-identification as service providers, and support for public education as an anti-corruption strategy. The *Reporting Index* measures knowledge of channels available for the reporting of unethical behavior, past reporting behavior, and frequency of discussions of unethical conduct with colleagues. The *Monitoring Index*, applicable to officers who supervise other officers, includes frequency of monitoring, detecting, and disciplining junior officers for unethical behavior. The *Perceptions Index* reflects views on the prevalence and severity of corruption in the police, agreement with external assessments, and personal exposure to bribery incidents. Lastly, the *Citizens Relationship Index* includes perceptions of police-citizen relations and attitudes toward aggression versus courtesy in policing.

The questions entering each index are displayed in Table A2 in Appendix A. The full questionnaire is available in the Online Appendix. To construct each index, we first demean each component of the index and divide it by the standard deviation of the corresponding control group variable (so that all measures are on a comparable scale) and then we take their weighted average, using the methodology introduced by Anderson (2008). In addition to creating standardized indexes, we correct the p-values associated to individual hypotheses concerning each survey-generated outcome by employing the step-down multiple testing method developed by Romano and Wolf (2005), as further discussed in Section 4.2.

[25]The baseline questionnaire, in its paper form, was 20 page long. The endline questionnaire was reduced to 7 pages only and can be downloaded here.

[26]More detailed references to the Pre-Analysis Plan (PAP), and explanations of when and why we deviated from it, are provided in Appendix C.

The phone implementation of the endline survey made it impossible to implement a cheating game that would require the rolling of a die (since we could not provide subjects with a die). We therefore decided to employ an incentivized game that would instead require the tossing of a coin, as we expected all subjects to be able to find a coin to participate in the game. We based our design on the coin toss game first introduced by Abeler et al. (2014). Subjects are asked to toss a coin four times in private, and receive a fixed payoff for every tail that they report to have obtained. Since there is no way for the research team to verify the actual number of tails a subject obtains, there is an incentive to lie. In our setting, we rewarded each tail with a bonus of 10 GHS, for a maximum of 40 GHS.[27]

Similarly to the baseline mind game, in this game it is impossible to tell whether a subject who reported a large number of tails was lying. We will start our analysis by comparing the empirical distributions of reported tails obtained for Treatment and Control officers, and the corresponding theoretical distribution. We will then examine through regression analysis the reported number of tails, the likelihood of reporting more than 2 tails (since 2 is the mean and the mode of the corresponding theoretical binomial distribution).

As secondary outcomes, we had pre-registered officers' aspirations and job satisfaction. The need to shorten the survey at endline, due to the phone implementation, led us to only keep one overall job satisfaction question in the survey. We report estimates of treatment effects on such measure in the Appendix. We had also pre-registered the collection of data on field behavior through surveys of tro-tro drivers (minibus drivers) and passengers. However, the COVID pandemic made this data collection impossible. Nevertheless, we sought to capture officers' field behavior through alternative means. Specifically, we obtained partial district-level administrative data on monthly road infractions filed by traffic police, as well as data on court orders, convictions, and fines issued. These administrative indicators serve as indirect measures of officers' propensity to engage in illegal behavior, under the assumption that corrupt officers prefer to solicit bribes from drivers rather than file formal charges. If this assumption holds, we would expect to see an increase in these administrative measures in treatment districts relative to controls following the training intervention, but not before. A limitation of this approach is that the available data are often aggregated across multiple districts—including both treatment and control areas. Our analysis is based on data for 17 districts - for which we have "clean data" - over 60 months (from January 2017 to December 2021). We describe these data, our empirical strategy and the estimated impacts of the training intervention on these measures of officer field behavior in Section 6.

[27]This is in addition to 30 GHS that each participant received for participating in the phone survey. Based on the wage data generated by our survey, the average hourly wage of a police officer in our sample was 12 GHS.

4.1.3 Social desirability bias and experimenter demand effects

While the implementation of the survey by phone was something that we did not plan in advance (when we started the project in 2018), it has a number of important advantages. First, by increasing the social distance between the surveyor and the respondent, and by allowing subjects to participate outside of their place of work and at their preferred time, it minimized social desirability bias and experimenter demand effects. In other words, we expect the method of implementation to have reduced the likelihood that respondents answered sensitive questions the way they thought it would be socially appropriate to answer, or the way they thought they were expected to answer after participating in a training program (for the trained group). Moreover, the phone implementation of the cheating game is likely to have reduced any feeling of being observed by colleagues or by the experimenter, therefore increasing the likelihood of cheating.

Another advantage of our methodology is that the endline survey took place 20 months after the training and 25 months after the baseline survey. The former time lag further decreases the risk that answers and decisions made at endline are the result of experimenter demand effects. The latter time lag makes it unlikely that subjects remembered the answers they provided at baseline and used such answers as reference points when participating in the endline survey. Moreover, both the time lag between baseline and endline, and the fact that we employed different cheating games at baseline and endline made sure that that officers could not use their experience in the baseline games, or potential discussions with colleagues about such games, to guide their behavior in the endline cheating game. Finally, in order to further reduce the risk of experimenter demand effects, we did not mention the training program, or the "Agent of Change" group identity, at any point during the phone survey.

We formally address social desirability bias and experimenter demand effects in our empirical analysis, as part of our robustness checks. Moreover, in Section 6, we evaluate the impact of the training using district-level administrative data, albeit from a limited subsample of districts.

4.2 Estimation strategy

To measure the impact of training on officers' attitudes, perceptions and behaviour, we exploit the random variation in training status induced through both random selection of districts to be treated, and pre-determined police duty rosters.

In a clustered design trial with partial compliance, one can estimate several models to capture treatment effects. As explained in Section 3.2, we assigned treatment status to 21 randomly selected police districts, out of 32. However, the actual training participation was

mandated by the Police headquarters, where it was decided that, in order not to disrupt the normal functions of traffic police in the Accra region, only about half of the officers operating in these districts could participate during the two weeks we had planned for the training. Actual participation was based on the officers' duty rosters that had been pre-determined for the two-week period that coincided with the training. According to police reports, duty assignments follow a random rota system. This is broadly confirmed by our baseline survey data; when asked how they are assigned their weekly duties, about 80 % of all officers, and 90% of low rank (rank <3) officers, answered that posts are assigned through a random system.[28] In Section 5.1, we test for significant differences between trained and untrained officers in treatment districts and find only only a few differences.

Nevertheless, given that we do not have access to historical data on officers' duty assignments, we are unable to verify the randomness of the duty assignment. Therefore, we cannot exclude the possibility of selection into the training among officers in treatment districts. To account for this, we follow our Pre-Analysis Plan and start our analysis by estimating a conservative intent-to-treat (ITT) model, comparing outcomes for officers in control and treatment districts, irrespective of their actual participation in the training. we estimatethe following equation:

$$Y_{id} = \alpha + \beta T_d + \gamma X_{id} + \delta Z_d + \varepsilon_{id} \tag{1}$$

where Y_{id} is one of our outcome variables for officer i in district d, i.e., behavior in the endline cheating game, or a survey-generated index measuring either values and beliefs, or the reporting of unethical behaviour, or the monitoring and disciplining of junior officers, or perceptions of corruption, or attitudes toward citizens. T_d is an indicator equal to 1 for all officers based in districts that were randomly assigned to the treatment group, regardless of training status. The vector of officer characteristics is captured by X_{id}, and is selected for each outcome through a Double Lasso procedure (Belloni et al., 2014), out of a large set of covariates measured at baseline.[29] X_{id} also includes the variables measured at baseline specific to each outcome. For instance, if the dependent variable is the Reporting Index, in addition to the Lasso-selected controls, we include in the specification the baseline survey measures of reporting behavior, aggregated using the weighted average method introduced by Anderson (2008). If the dependent variable is an outcome generated by the incentivized cheating game, we always include as a control the behavior in the baseline incentivized cheating game. Z_d

[28]It is possible that higher rank officers have some control over which weekly duties they are assigned.

[29]The Double Lasso method performs two separate Lasso regressions, one regression of the outcome on all covariates to identify relevant predictors of the outcome, and another regression of the treatment on all covariates to identify predictors of the treatment. It then takes the union of the selected covariates from both regressions as the control vector.

is a district-level control for district size, i.e., the number of traffic police officers operating in the district.

In all regressions, the standard errors are clustered at the police district level, our unit of randomization (Abadie et al., 2023). Given the small number of clusters (32), we also report p-values adjusted for small-sample clustering using wild cluster bootstrapping (Cameron and Miller, 2015). In addition, we conduct further robustness checks by estimating treatment effects through randomization inference (Heß, 2017). Since we have a large number of outcome variables, hence multiple hypotheses, we correct the p-values associated to individual hypotheses by employing the step-down multiple testing method developed by Romano and Wolf (2005).[30] One issue is how we should treat the non-compliers when estimating equation (3). We adopt the most conservative approach, which is to consider the 6 trained officers in control districts as *untreated* in Control districts, and the three (small) treatment districts where nobody was trained still as Treatment districts, while categorizing the officers from these districts as *untrained in T*. Any positive effect of the training on the 6 officers from Control districts would bias β_1 downward.

Next, we employ a two-stage least squares model to estimate the local average treatment effect (LATE) when selection into treatment is a concern, or there is non-compliance either in the treatment or in the control group. Regarding the latter, we had 6 officers from control districts participate in the training, as well three treatment districts (with a total of 9 employed officers) where no officer participated in the training.

$$\begin{aligned} Trained\_T_{id} &= \tau + \theta T_d + \gamma' X_{id} + \delta' Z_d + \eta_{id} \\ Y_{id} &= \alpha + \beta \hat{Trained\_T_{id}} + \gamma X_{id} + \delta Z_d + \varepsilon_{id} \end{aligned} \tag{2}$$

As per model (2), the estimation proceeds through a two-stage least squares regression. In the first stage, the actual training participation ($Trained\_T_{id}$) is predicted by the instrument, which is the initial assignment of districts to treatment and control groups (T_{id}). In the second stage, the outcomes of interest are regressed against the treatment take-up, as predicted in the first stage. The estimate of β is the local average treatment effect. Note that if the selection into treatment take-up and non-compliance are negligible, the IV LATE estimates from equation (2) will be very similar to the ATT estimates from equation (3).[31] As before, we cluster the standard errors at the police district level, and also report wild-bootstrapped p-values, and Romano-Wolf p-values.

[30] We correct for the number of hypotheses in each of the two sets of regressions, for the survey based-outcomes and the game outcomes.

[31] However, the LATE estimates are comparatively less efficient than OLS estimates, and result in smaller t-values.

Finally, we estimate a third model that assumes random assignment of officers to training in Treatment districts. This generates estimates of treatment impacts on trained versus untrained officers in Treatment districts, as opposed to officers in Control districts:

$$Y_{id} = \alpha + \beta_1 Trained\_T_{id} + \beta_2 Untrained\_T_{id} + \gamma X_{id} + \delta Z_d + \varepsilon_{id} \quad (3)$$

where $Trained\_T_{id}$ and $Untrained\_T_{id}$ are two indicators for the officers in treatment districts who participated in the training, and for those who did not participate, respectively. X_{id} and Z_d are officer and district-level controls, selected through the Double-Lasso method, as in models 1 and 2. Under the assumption of no systematic selection into the training, the β_1 coefficient is essentially the average treatment effect of our training program (ATE), while β_2 captures spillover effects from trained through untrained officers in Treatment districts. Ideally, to best analyze spillover effects from trained to untrained officers, we would need to know which officers where deployed together through duty rosters. However, we were unable to obtain these data from the MTDT. The only information we have is whether officers were employed in the same district and whether they participated in the training.

As a secondary analysis, we expand models (1) and (3) to examine possible mechanisms behind the treatment effects. We examine heterogeneous effects of the training across several district and officer characteristics: i) heterogeneity by intrinsic motivations of the officers to serve the community when they first joined the police; ii) heterogeneity by officer's rank; iii) heterogeneity by number of officers trained in a treated districts. When estimating the heterogeneous treatment effects by intrinsic motivations and rank, we use the full sample of officers and we simply add to models (1) and (3) the interaction terms between the variable of interest and the indicators for treatment district (in equation 1) or for trained and untrained in Treatment districts (in equation 3).

5 Results: Survey and Incentivized Cheating Game

5.1 Descriptive Statistics and Balance Tests

Table 1 reports an extensive set of officers' baseline demographic characteristics, survey answers and decisions in the cheating game. The table also displays p-values generated by tests of equalities across the relevant groups. The working sample consists of the 412 officers who were surveyed both at baseline and at endline; of them, 141 officers were based in control districts and 271 in treatment districts. About half of the surveyed officers in the treatment districts participated in the training program.

In Panel A of Table 1, we report average demographic and job-related characteristics; in

Panel B, we display baseline measures of (a large subset of) our survey-generated outcome variables;[32] in Panel C we report summary statistics of officers' behaviors in the baseline cheating game, i.e., the percentage of officers who reported that the number they rolled matched the number in their mind. Panel D reports descriptive statistics for a subset of questions pertaining to the monitoring and disciplining of junior staff. Only officers who had experience with leading a team of lower-rank officers at the time of the survey (178 officers in total) answered these questions.

The comparison of officers in Control and Treatment districts (columns 1 and 2) shows balance in all but two variables. In both groups, around one quarter of officers are women, the average age is in the 35-44 range, the average rank is 2-3 positions above the lowest rank (that of Constable), and the average monthly wage is around 1,950 GHS (324 USD). Panel B provides a full account of all the questions related to unethical behavior, which we asked at baseline. Note that we did not ask officers to answer questions about their own unethical behavior. Instead, we asked about the behaviors of other officers in their district and their propensity to report such behavior to higher authorities. We also asked questions about their perceived identity as service providers, and their attitudes toward citizens.

We see balance in most of our survey measures. We do however find that, at baseline, officers in Treatment districts witnessed bribery by colleagues more often than officers in Control district (significant at the 10 percent level), and a larger percentage of Treatment officers ever reported unethical behaviour (significant at the 1 percent level), as compared to officers in Control districts. These differences may be due either to a higher prevalence of corruption in the Treatment districts, or a higher propensity of Treatment officers to report such behavior. The only other significant difference is observed in the sub-sample of higher-rank officers who had experience managing a team of lower-rank officers (Panel D), where we see that officers in Treatment districts reported having disciplined juniors more often than officers in Control districts (significant at the 1 percent level). Again, this may be due to a higher need for disciplining, or a higher willingness of Treatment officers to discipline lower rank officials. We account for these imbalances in the empirical analysis, by always controlling for the baseline outcome measures when assessing the impact of the training on the corresponding dependent variables.

In columns (4) and (5) of Table 1, we compare trained and untrained officers within Treatment districts. According to the Police, participation in the training was determined by pre-established duty rosters, which follow a rota system. This process suggests a quasi-random assignment of officers to training. We assess this quasi-random assignment by conducting balance tests on observable characteristics of trained and untrained officers. The descriptive

[32] A few survey questions that we use to generate our endline outcome indexes were not included at baseline.

statistics displayed columns (4) and (5) of Table 1 show very few differences. First, trained officers are slightly younger and of lower rank. This is likely due to the fact that higher rank officers are less able to leave their duties for two full days to participate in the training.[33] We only see two other differences between trained and untrained officers in Treatment districts. The trained officers witnessed unethical behaviour slightly more often (significant at the 10 percent level) and were slightly less comfortable reporting corruption (significant at the 5 percent level). These differences run counter to the expected effects of the training. Notably, if a greater likelihood of witnessing unethical behavior and lower willingness to report it are correlated with selection into the program, they would likely introduce attenuation bias in our estimates.

Next, we test for balance in the officers' behaviors in the incentivized cheating game that we implemented at baseline (described in Section 3). The first row of Panel C, in Table 1, shows the percentages of officers who, after rolling the die, stated that the number they obtained matched the number in their mind. Both in Control and Treatment districts, the percentage of officers who reported a match in the baseline mind game is about 60%, as shown also in the top panel of Figure 1. The observed occurrence of matches is significantly higher than the theoretical prediction (of 16.7%). This indicates that cheating among officers was pervasive at baseline, with no significant differences across treatment groups.

5.2 Treatment effects

5.2.1 Impact on survey-generated outcomes

We study the impact of the "Proud to Belong" training program on five aggregate indexes measuring officers' values, attitudes and beliefs. The indexes were created by standardizing and averaging answers to individual survey questions, using the methodology introduced by Anderson (2008). In particular, we generated: 1) a Values, Beliefs and Identity index, 2) a Reporting (of unethical behavior) index, 3) a Monitoring index (defined only for officers who had managed a team of lower-rank officers), 4) a Corruption Perception index, and 5) a Relationship with Citizens index. Table A2 in Appendix displays the individual questions forming each index.[34] All indexes are standardized around the control mean, hence they are displayed in standard deviations from such mean.

When looking at the averages of each index for trained and untrained officers in Treatment districts (see Figure A2 in Appendix A), we see that the trained officers score significantly higher in the Values index and the Reporting index, when compared to the officers in Control

[33] It is also possible that higher-ranking officers have more control over their schedules and may have opted out of the training if selected. We consistently include officer rank in our set of controls.

[34] The full endline survey can be found in the Online Appendix.

districts, whose means are set to zero, and also when compared to the untrained officers in treatment districts. Trained officers also have higher means in the Citizen Relationship index and the Corruption Perception index, as compared to the control officers. None of the indexes display means that are significantly larger for the untrained officers in Treatment districts than for the officers in the Control districts.

We report the Intent-to-treat (ITT) estimates from equation (1) of Section 4.2 in Panel A Table 2. For each index, we display estimates without controls in odd columns, and estimates with Double-Lasso controls in even columns. We report robust standard errors, clustered at the police district level, in parentheses. In addition, in square brackets, we report p-values corrected for multiple hypotheses testing (for each family of hypotheses in Tables 2 and 3), using the Romano and Wolf (2005)'s procedure, and in curly brackets we report Wild-Clustered Bootstrap (Restricted) p-values, to correct for the small number of clusters. The ITT estimates displayed in Columns 9 and 10 show that the training had a marginally significant positive impact (by 0.16 standard deviations, significant at the 10 percent level) on the index measuring the relationship with citizens. Columns 1 and 3 indicate that the assignment to the training program had a positive and significant impact on the Values and teh Reporting indexes, which increased by around 0.17 and 0.15 standard deviations, respectively, in treatment districts, as compared to the control mean, although the estimated coefficients lose statistical significance when correcting for multiple hypotheses testing, and when including Double Lasso controls in the specification (column 2 and 4). We see no impact of the training on the indexes capturing the monitoring behavior of higher-rank officials toward subordinates (columns 4 - 6), and the index measuring perceptions of corruption in the district (columns 7 and 8).

Given that only about half of the officers in treatment districs were trained, we expect the LATE estimates to be larger than the ITT estimates. This is what we see in Panel B of Table 2. Specifically, the LATE estimates are more than double in magnitude compared to the ITT estimates, and significant at the 10 percent level for the Values index and at the 5 percent level for the Relationship index in the specifications including the Double Lasso controls. The LATE estimates indicate that the training increases the Values index by 0.315 standard deviations, and the Relationship with Citizens index by 0.34 standard deviations (with the level of signficance declining when adding Double Lasso controls, correcting for multiple hypotheses testing and for the small number of clusters). None of the other survey-generated indexes are significantly and robustly impacted by the training.

The difference in magnitudes between LATE and ITT estimates also suggests that there are minimum spillovers between trained and untrained officers in the treatment districts. This is confirmed when estimating model (3) of Section 4.2, i.e., when generating treatment

effect estimates separately for trained and untrained officers in treatment districts, as opposed to officers in control districts. Columns 1-5 of Table 4 show that the training significantly impacted the Values Index and the Citizen relationship Index for trained officers only. The average treatment effect estimates for this group of officers are roughly double the ITT estimates, and slightly smaller than the LATEs. We do not see any impact of the training on the untrained officers in Treatment districts, once again suggesting that the program did not spill over from trained to untrained officers.[35]

We had pre-registered job satisfaction as a secondary outcome of interest. In Appendix Table A3, we report the estimated treatment effects on officers' "overall satisfaction with their job." We find evidence of marginally significant positive impact of the training on this measure, suggesting that the program may have improved officers' perceptions of their work experience.

5.2.2 Impact on behavior in the incentivized cheating game

We start the analysis of officer behavior in the endline cheating game by assessing the distribution of the reported coin tosses. Officers had to toss a coin 4 times and report the number of obtained tails, as in Abeler et al. (2014). Figure 2 displays the percentages of officers who reported 0, 1, 2, 3, and 4 tails. The distributions are shown separately for the officers in the Control and Treatment districts in panel A, and for the untrained and trained officers in Treatment districts, as opposed to Control officers, in panel B. The figures also display the theoretical binomial distribution $B(N = 4, p = 0.5)$.

The distribution generated by the officers in the Control districts is heavily right-skewed. These officers under-report instances of one and two tails, while substantially over-reporting three and four tails. In fact, the frequencies of three and four tails are nearly twice as high as what would be expected under truthful reporting. In contrast, the distribution or reported tails in Treatment districts more closely aligns with the theoretical distribution. Although the empirical and theoretical distributions are still statistically different from each other, the probability of reporting two tails in the Treatment group matches the theoretical expectation, and reports of three tails are notably lower than those observed in the Control group. Kolmogorov-Smirnov tests of equality between the Control and Treatment empirical distributions confirm that officers in treatment districts report fewer tails in the coin toss game across the distribution (p-value=0.072 in a one-tailed test; p-value=0.144 in the combined test).

[35]The tests for the equality of the coefficients of "trained"and "untrained"indicators for the Values index and the Relationship index confirm that the training program had a differential impact on trained and untrained officers operating in the treatment districts.

When distinguishing between trained and untrained officers in Treatment districts, in panel B of Figure 2, we see that the shift of the Treatment districts' distribution to the left is entirely driven by the trained officers, whose distribution displays no over-reporting of four tails, and three tails reported less frequently than two tails. In contrast, the distribution observed for the untrained officers is similar to that of Control officers, displaying over-reporting of both three tails and four tails. Kolmogorov-Smirnov tests of equality of the distributions of tails reported by trained, untrained officers in treatment districts, and control officers indicate that the trained officers behave significantly differently than untrained and control officers. The p-values generated by two-sided tests comparing trained officers against untrained and control officers are 0.049 and 0.004, respectively. In contrast, the distributions observed for control and untrained officers are not different from each other (p-value=1.00). Overall, this analysis indicates that trained officers were less likely to cheat than officers in the other two groups.

We confirm these results in Table 3, which presents the estimates from equations (1) and (2) in panels A and B, where the dependent variables are three outcomes generated by the incentivized cheating game: 1) number of tails reported (columns 1-2); 2) the probability of reporting more than two tails (columns 3-4); 3) the probability of reporting exactly three tails (columns 5-6). As before, odd columns present ITT or LATE estimates without controls. Even columns report the corresponding estimates when including controls selected through the Double-LASSO procedure, and each outcome measured at baseline. We report robust standard errors, clustered at the police district level, in parentheses, p-values corrected for multiple hypotheses testing for each family of outcomes (survey measures and game outcomes), using the Romano and Wolf (2005)'s procedure, in brackets, and Wild Clustered Bootstrap (Restricted) p-values, in curly brackets.

The ITT estimates show that officers in Treatment districts are significantly less likely (by about 18 percentage points) to report three tails, and the results suggest they are less likely (by 10.9 percentage points) to report more than two tails than officers in Control districts (albeit the latter effect is not statistically significant in column 4).[36] Given that only around half of the officers in Treatment districts participated in the training, we find an impact of being offered training of about 50% of the magnitude of the LATE estimates displayed in Table 3, which are significant at 5% level for the probability of reporting three tails, and at 10% level for the probability of reporting more than two tails, as shown in Table 3, Panel B.

The estimates obtained separately for trained and untrained officers in Treatment districts, displayed in columns 4 to 6 of Table 4 tell the same story as the LATE estimates: trained officers are about 17 percentage point less likely to report more than two tails and 12

[36]The sizes of the estimates are similar across specifications with and without controls.

percentage points less likely to report three tails, as compared to Control officers. The null and insignificant coefficients obtained for the untrained officers in Treatment districts, shown in columns 6 to 8 of Table 4, indicate that the training program had no spillover effects from the trained to the untrained officers operating in the same district.

5.2.3 Robustness

As a first robustness check, we replicate our analysis using Randomization Inference. The resulting p-values (reported in square brackets in the Appendix Table A4) show that our ITT estimates for the survey and cheating game outcomes for which we found significant effects in our main Tables (Tables 2 and 3) remain marginally significant, and our estimated impacts on the trained officers remain significant at the 5 percent level. wild-bootstrapping]

We next test whether our treatment effects are robust to including Central district officers in the control group. We surveyed nearly all (253 out of 273) Central officers at baseline, and about half of them – randomly selected – at endline. We initially excluded the Central district from the RCT because Central officers often move between districts and could interact with both trained and untrained peers, making them a less clean control group. However, since we find no evidence of spillover effects between trained and untrained officers, this is less of a concern. Still, consistent with our pre-analysis plan, we excluded Central officers from the main analysis. Here, we first compare baseline characteristics of officers in the Central district versus those in our randomly selected Control districts. We do not see any notable significant differences, as shown in Table A5 in the Appendix. We then proceed to estimate our main regressions for a sample that now includes 133 Central officers (those surveyed both at endline and baseline) in the Control group. The estimated treatment effects, displayed in Table A6 in the Appendix, are similar in magnitude to those in Tables 2 and 3, but noisier for the survey outcomes and with higher levels of statistical significance for the cheating game outcomes.

Our last robustness check aims to address the possible concern that, given the self-reported nature of our survey-generated outcomes, our main findings may be driven by social desirability bias or experimenter demand effects. As we discussed in Section 3.2, the large time lags between the baseline and the endline survey (over 2 years), and between the treatment and the endline survey (20 months), together with the phone implementation of the endline survey, should attenuate these concerns. Nevertheless, we test whether social desirability and experimenter demand effects biased our measured effects upward. We do not have possible measures of social desirability bias in the survey. However, we conduct a test based on the assumption that subjects whose survey responses are more likely to be biased by social desirability and experimenter demand effects would show the greatest improvement, between

baseline and endline, in the survey-generated outcomes of interest. We identify these officers, and we test whether our results are robust to excluding them from the sample. We proceed as follows. First, we create an aggregate index of all the (standardized) survey questions of interest (i.e., those generating our outcomes) which were exactly the same in the baseline and endline surveys.[37] Then, we compute a measure of individual improvement as the difference between the endline and the baseline scores on this index. We identify as "most likely subject to social desirability bias" the 10% officers who recorded the highest improvement in the aggregate index over the 26 months between the two survey waves. The assumption here is that these improvements are the most likely to be inflated by social desirability.

We start by checking whether the "top improved officers" are more likely to be found among trained officers. This would be indicative of experimenter demand effects. We see no statistically significant difference in the percentage of top improved officers among untrained (9.5 percent) and trained (11.5 percent) officers (p-value = 0.529). We then proceed to replicate our analysis on the restricted sample, i.e., excluding the top improved officers. The results, displayed in Appendix Table A7, show that, despite the sample reduction from 412 to 370 officers, our treatment effects on the Citizen Relationship Index and the cheating game's measures are largely unaffected.

5.3 Exploratory Analysis: Heterogeneous Treatment Effects

In this section, we report findings from exploratory analyses of heterogeneous treatment effects.[38] The training focused heavily on the re-activation of intrinsic motivations and the priming of the identity of service provider, which we assumed (at least some) officers had when they first joined the police. In the baseline survey, we asked officers to state the primary reason they joined the police. About 60 percent of officers stated that they "wanted to help the people/their community."[39] Based on this answer, we create a dummy variable for *initial intrinsic motivations*. As shown in Table 1, the intrinsic motivation indicator is balanced among control and treatment officers, and among trained and untrained officers in treatment districts. We use this variable to conduct an analysis of heterogeneous treatment effects by initial intrinsic motivations.

[37] These are questions listed as questions 3, 4, 5, 6,12, 13 and 15 in Table A2 in the Appendix.

[38] We did not pre-register our analysis of mechanisms. In analyzing heterogeneous treatment effects, we focus exclusively on the channels through which we believe the training may have operated, based on how the program was conceived and designed. For details on how and where our analysis does or does not follow the pre-analysis plan, see Appendix C.

[39] Officers had to choose among the following options: 1) I wanted to make sure I had a job; 2) I wanted to make sure I get a job that pays well; 3) I wanted to continue the tradition of my family; 4) I wanted to serve the people (or my community); 5) I wanted to be respected by the people or the community; 6) I wanted to get a job that would allow for career advancement; 7) Other.

The estimates in Appendix Table A8 show that, for officers who did not join the police out of intrinsic motivation, the training had no effect on survey-based measures of values, attitudes, and perceptions (columns 1–4). In contrast, the training significantly affected the Value Index and Monitoring Index for intrinsically motivated officers (p-values of 0.039 and 0.077, respectively). However, for the index measuring relationships with citizens, the training had a positive effect only for non-intrinsically motivated officers. This likely reflects a ceiling effect, as intrinsically motivated officers already exhibited stronger relationships with citizens (the index is 0.39 standard deviations higher for intrinsically motivated officers, as opposed to not intrinsically motivated colleagues, in the control group). The null impact on the not-intrinsically motivated suggest that intrinsic motivations are difficult to instill if absent from the outset. On the other hand, the positive effect on the citizen relationship index indicates that training modules offering practical strategies for professional and effective citizen interaction can still benefit officers who joined the police for reasons other than public service.

Next, we focus on the officers' seniority in the police service. As a measure of seniority and experience, we take the officer's rank at baseline. Low-rank officers are Constable and Lance Corporal (ranks 1 and 2); high-rank officers are Sergeants, Inspectors and above (i.e., rank 3 and above). Around two thirds of the officers in our sample have high rank.[40] our analysis of heterogeneous impact by low vs. high rank, displayed in Table A9 in the Appendix, shows that the training had large and significant impacts on the Values, Monitoring and Relationship indexes of lower-rank officers. The magnitudes of the treatment effects for these officers are larger than the impacts estimated on the full sample (see Table 2), and are between 0.26 and 0.37 standard deviations above the control group means. depending on the index. The coefficients on the interactions between the training and the high-rank indicator display a negative sign for the three indexes that are significantly impacted among lower-rank officers (Value, Monitoring and Relationship Indexes), albeit not statistically significant. Linear combinations of the corresponding coefficients indicate that, in fact, the training did not impact any of the indexes for the higher-rank officials. We see similar patterns, but weaker, when examining behavior in the incentivized cheating game. Overall, this analysis provides suggestive evidence that the training was especially effective on lower-rank officers.

Lastly, we test whether an important component of the training was the awareness that one's colleagues also received the training, fostering the belief that collective change was possible in one's police district. We test whether the impact of the program increased with the number of trained colleagues from the same district (while controlling for the total number

[40]High rank officers have on average 18.3 years of experience compared to 10.4 years of experience for low rank officers. Replicating the analysis by employing officers' years of experience or officers' age as proxies for seniority leads to similar results.

of officers in a district). This analysis is restricted to Treatment districts only, and reported in Table A10 in the Appendix. The estimates show that the identified impacts of the training on trained officers are largely independent from the number of trained colleagues in the same district. This may be because the training was conceived as super-district, and the new identity of "Agent of Change" was envisioned and presented to officers as a Ghanaian police identity, rather than a district-specific identity. In fact, none of the activities implemented during the two days of training referred to districts. This analysis also provides some evidence of spillovers in attitudes towards citizens and behavior in the cheating game: as the number of trained officers increased, untrained officers reported more positive attitudes toward citizens and were less likely to report large numbers in the game.[41] A more accurate examination of spillovers effects would account for the fact that officers within a given district are deployed together to the field (in groups to 2 to 4), according to duty rosters, leading to variation in untrained officers' exposure to trained officers, and vice versa. nfortunately, despite our efforts, we were unable to obtain access to the weekly duty rosters data from the police.

6 Results: Administrative Data on Field Behavior

We have so far shown evidence of the positive impact of the training program on officers' values, beliefs and attitudes (survey-based) and their (un-)willingness to engage in unethical behavior, when acting in isolation from others (game-based). However, it is unclear whether these improvements can translate into better behavior in the field, where officers operate within a social environment and may be influenced by colleagues, superiors, or even by drivers offering bribes. Ideally, we would want officer-level administrative data on field behavior, i.e., complaints filed against officers, for instance for excessive use of force or for demanding bribes. In a context like Ghana, however, such data do not exist. We had originally planned to survey tro-tro (minivan) drivers and deploy enumerators to travel in these vehicles to record stops and potential interactions with police officers. Unfortunately, this became unfeasible due to the pandemic. As an alternative, we worked on obtaining available district-level administrative data on the monthly total cases (of road infractions) filed by traffic police officers, as well as issued court orders, convictions as decided by courts, and total fines.

These outcomes can reflect both officer integrity and effort. Specifically, following the framework in Conover et al. (2023)'s study of traffic enforcement in India, we assume that, when observing a traffic violations, officers have three choices: 1) enforce the law by issuing a

[41] Additional heterogeneity analyses by the gender of the officer (not reported in the paper but available upon request) show that while the impact of the training on the survey generated measures is not different from men and women, the impact of the training on behavior in the cheating game is significantly larger for female officers.

fine; 2) engage in corruption by collecting a bribe instead of filing the violation; 3) do nothing (shirking). Corruption may also involve officers stopping drivers who have not committed a violation and extorting money from them under threat of issuing a fine if they refuse to pay a bribe. Low levels of filed cases or court activity may therefore indicate either poor enforcement efforts (i.e., due to lack of on-the-job motivation) or the substitution of formal law enforcement with bribery, where officers let offenders (and possibly not offenders) go in exchange for payment. Accordingly, increases in filed cases, convictions and fines following the training program would be consistent with improvements in officer behavior, whether through increased effort, reduced corruption, or both.

6.1 Data

We obtained district-level administrative data for the 32 districts. However, data for 15 of these districts are aggregated in a manner that hinders our ability to cleanly compare our Treatment and Control districts. This is due to the reporting of combined data from some Control and Treatment districts, or the inclusion of data from other districts that do not have a traffic police unit, and are therefore not in our sample. Upon excluding these districts, we are left with 17 districts with "clean" administrative data, which we designate as our "Pure" Administrative Control and Treatment districts.[42] In Table 5, we compare the officer characteristics, as recorded in our baseline survey, in districts for which we have clean administrative data and in districts for which we have no such data. While districts with clean administrative data are slightly smaller on average (17 vs. 22 officers at baseline, $p = 0.246$), officer characteristics are otherwise highly comparable across the two groups, with no statistically significant differences.[43]

The district-level data at our disposal are monthly, ranging from September 2017 to December 2021 (60 months). Recall that our baseline data collection took place in October and November 2018, and our training was implemented during the last week of April and first week of May 2019. Following the training, the WhatsApp chat groups, involving officers who were trained on the same day, were formed and remained active between June and September 2019. The award ceremony took place in December 2019.

[42]For 10 of the "Pure" Treatment districts, the data are aggregated in groups of 2 or 4 districts.

[43]Balance tests for Control and Treatment districts for which we have clean administrative data, displayed in Table A11 in the Appendix, show that the Administrative Treatment districts are remarkably similar to the Administrative Control districts, based on officers' baseline demographics, attitudes and experiences, and their behavior in the baseline incentivized cheating game.

6.2 Analysis and Results

We start by plotting quarterly district-level data, averaged by the number of officers in a district, by district treatment status. Since the intervention took place the last week of April 2018 (day 1) and the first week of May (day 2) 2018, we set April as time 0, and display pre- and post-April quarterly trends for Control and Treatment Administrative Districts (i.e., the 17 districts for which we have clean administrative data) in Appendix Figure A3. We first note that the average monthly number of cases filed by an officer, the number sent to court and the number convicted are all remarkably low in both Control and Treatment districts. Specifically, in the pre-intervention months, 0.18 cases are filed monthly on average per officer, 0.14 are sent to court and 0.13 are convicted. The amount of Ghanaian Cedi collected in fines in a month by an officer on average is about 62 GHS – the equivalent of 13 USD in 2018. These low numbers suggest a potential lack of law enforcement activity by officers, raising the possibility that they may be substituting court orders with bribes.

Figure A3 shows evidence of parallel trends for all our administrative outcomes, up to April 2019, with no significant differences in the outcomes observed in Treatment and Control districts. However, in the quarter immediately following April 2019, Treatment and Control districts significantly diverge, with the treatment districts displaying more per officer filed cases, as well as more per officer court orders, court convictions and total fines (in Ghanaian Cedi).

Next, we estimate the dynamic treatment effects of the training program, by conducting a difference-in-differences (DiD) event study, which includes all (27) months before and all (32) months after the training, for which we have partial administrative data. Since the training was completed the first week of May 2019, we set May 2019 as time "0" and April 2019 as time "-1" for all districts. Hence, the reference month is April 2019, i.e., the month preceding the completion of the training. Formally, we estimate the following equation:

$$\begin{aligned} \mathrm{Y}_{td} = \alpha + \beta_{27}(\text{Lag } 27)_{td} + ... + \beta_2(\text{Lag } 2)_{td} + \gamma_0(\text{Lead } 0)_{td} + ... \\ + \gamma_{32}(\text{Lead } 32)_{td} + \mu_d + \lambda_t + \varepsilon_{td} \end{aligned} \tag{4}$$

where Y_{td} refers to one of our four outcome variables for district d in month t, i.e., 1) the number of cases filed in district d in month t, averaged by number of officers in d; 2) the number of cases sent to court in district d in month t, averaged by number of officers in d; 3) the number of convicted cases in district d in month t averaged by number of officers in d; 4) amount of Ghanaian Cedi collected in fines in district d in month t averaged by number of officers in d. We include time (month and year) fixed effects and district fixed effects, λ_t

and μ_d respectively.[44]

The resulting event study monthly estimates for the four outcome variables, averaged by the number of officers in a district, are shown in Figure 4. For ease of display and interpretation, we show the estimates for 13 months before and 13 months after April 2019 (i.e., 12 months post intervention), noting that April 2020 (month 12 in Figure 3) corresponds to the surge of the COVID-19 pandemic. We see significant differences between Treatment and Control districts emerge for all four outcomes in the immediate aftermaths of the training (June and July 2019).[45] The number of convictions and the amounts of court fines remain larger in the Treatment districts than the Control districts over the 2019 summer months (although only marginally significant). We also see a second increase in the Treatment districts, as compared to the Control districts, in February and March 2020, possibly in reaction to the award ceremony for trained officers that was held in late December 2019.

We display estimates generated by a standard difference-in-differences (DiD) model in Panel A of Table 6, and a DiD over multiple time periods, in Panel B of Table 6. In both models, the empirical specification includes time (month and year) fixed effects, as well as district fixed effects. Since data for 8 districts are aggregated into 3 administrative units, we conduct the analysis at the administrative unit level, i.e., we analyze data for 11 administrative units for 60 months, leading to 660 observations. We report standard errors clustered at the administrative unit level, as well as Wild-bootstrapped p-values correcting for the small number of clusters.[46]

The DiD estimates over only two time periods (pre- and post-training), in Panel A of Table 6, show no evidence of a significant impact of the training on our four outcome variables. However, when we estimate the DiD model for multiple time periods after the training, we find that the program had a positive and large impact on the four measures of field behavior in the three months following the interventions (Panel B). The training program led to an increase in the average monthly district-level number of cases, court orders and convictions per officer equal to 0.256, 0.242 and 0.237 cases, respectively, in the three months post training (over the pre-training Control district values of 0.19, 0.14 and 0.12 cases). The monthly average per officer fines increased by 126 Ghanaian Cedi, over a pre-treatment Control mean of 65 Cedi. The impact of the training on issued fines is sustained for up to 6 months post intervention.

[44]We use analytical weights for the number of officers in each district at the time of the baseline survey (October 2018). This is because the data are at the district-level and averaged by the number of officers in a district, yet this number varies by district.

[45]Recall that the training ended on the first week of May 2019. The fact that we do not see a significant impact on filed cases and cases sent to court in May is likely due to the timing of the reassignment of trained officers to road duties, which likely occurred in mid- to late-May.

[46]Given that the data are aggregated at the administrative unit level, and averaged by the number of officers, N, in a district, yet districts differ in N, we use analytical weights where, for each district, the weight is the number of officers in a district.

In sum, our analysis of administrative data, albeit on a smaller sample and based on district-level data, suggest that the training program induced officers to file more cases and issue more court orders up to three months post training. This resulted in more convictions and more money collected in fines. This is indicative of more effective law enforcement, and could also indicate a reduction in bribery, under the assumption that bribes are demanded or offered to avoid law enforcement. Given the rota system, where only a portion of officers are on duty each week, and considering that we trained only half of the officers in treatment districts, it is remarkable that we can discern a positive impact from the training, albeit only in the short term.

6.3 Discussion: Reconciling Impacts on Individual Outcomes and Field Behavior

The findings generated by our survey and incentivized cheating game offer encouraging evidence that an in-person ethics training centered around identify and intrinsic motivations can lead to long-term shifts in attitudes and willingness to engage in unethical behavior. At the same time, the effects on field behavior appear to be short-lived, concentrated in the first three months following training. This disconnect between persistent attitudes' changes and more temporary behavioral effects may reflect a tension between individual moral intentions and external pressures that shape officers' actions.

One possible explanation for the temporary nature of the impacts on field behavior is that officers who attempted to follow the training's principles may have faced resistance from colleagues or superiors and reverted to established norms in the longer term. This may be more likely for lower-ranked officers, who were most affected by the training and who may have risked reprimand or other forms of punishment from higher-rank officers. Our data on promotions and demotions, generated by endline and baseline survey data, provide some support for this interpretation. When comparing the ranks of the surveyed officers at baseline and endline (more than 2 years later), we find that about 40 percent of the officers got promoted to a higher rank, with no significant differences across Control officers (40%), untrained officers in Treatment districts (43%).[47]

Another complementary explanation is that the training's lessons need to be reinforced regularly to persist. The first three months post-training, which is when we see behavioral

[47]We also do not see any significant differences in rank demotions, which occurred for 19 percent of the officers across all groups. None of the lower-rank officers was demoted between baseline and endline. However, heterogeneity analysis by officer rank reveals that lower-rank officers in Treatment districts were about 12 percentage points less likely to be promoted than their counterparts in Control districts (who all got promoted to a higher rank), as shown in Table A12 in the Appendix. This suggests that efforts to act in line with the training's principles may have been discouraged, or at least not rewarded, within the police hierarchy.

changes in the field, correspond to the time when officers were most active in WhatsApp group chats with fellow trainees and were reminded about the trainings' content. This explanation is further supported by a second increase in officers' field activities after the award ceremony (i.e., months 9 and 10 in Figure 4) during which officers were given "Agent of Change" pins and certificates of training completion. This suggests that the ceremony may have acted as a refresher or reminder of the training's principles, leading to a second wave of (again, temporary) behavioral change.

A third, more optimistic interpretation is that field behavior changed in response to the training, but that its effects gradually extended to control districts through changes in drivers' behavior. In other words, it is possible that, after experiencing increased law enforcement in treatment districts, drivers adjusted their behavior by reducing opportunities for either law enforcement (i.e., fewer traffic infractions) or their likelihood to offer bribes, or both, across treatment and control districts. While this explanation is possible and aligns with a broader view of the training's potential to produce indirect effects through shifts in public expectations and behaviors, it seems less likely in our context given the relatively low level of overall enforcement activity observed even in treatment districts during the post-training period.

7 Conclusions

In this paper, we provided empirical evidence on the effectiveness of an ethics training program designed to leverage intrinsic motivation, individual identity, and shared group identity to improve the attitudes and behaviors of traffic police officers in Ghana. In collaboration with the Ghana Police Service, we randomly selected police districts within the Greater Accra Region to receive the training. About half of the traffic officers in these districts participated in a two-day program, implemented over two weeks in Spring 2019. The training was participatory and interactive, built around small group discussions, team-building activities, brainstorming exercises, and role-play scenarios. Its core objectives were to (re)activate officers' intrinsic motivation to serve the public and foster a shared identity as "Agents of Change."

Our findings from a survey and an incentivized cheating game conducted by phone 20 months after the implementation of the program, show that the training was successful in its primary objective of shifting officers' values, beliefs and perceptions related to unethical behavior, as well as their own willingness to engage in such behavior, as measured through an incentivized cheating game. The training also significantly affected officers' attitudes toward and relationships with citizens. We see no impact, however, on officers' reporting of

unethical behavior, their monitoring of subordinates and their perceptions of corruption in the police. We also find no evidence of spillovers from trained to untrained officers, which suggests that active participation in the program is necessary for changes in attitudes and behaviors to occur. In line with our expectations, the program operated primarily on officers who were intrinsically motivated to serve the public when they initially joined the police, and on younger, lower rank officers.

Although complete administrative data are not available, partial records from 17 of the 32 districts provide suggestive evidence of short-term improvements in officer field behavior. The fact that we observe lasting impacts on individual attitudes and propensities for unethical behavior 20 months post-training, but only short-lived effects on field behavior, suggest that enduring shifts in individual values alone may not be enough to sustain long-lasting behavioral changes. In particular, the fading of behavioral effects over time may reflect the absence of continued peer engagement and reinforcement, such as the reminders officers initially received through active WhatsApp group chats or later through public recognition at the award ceremony. Second, lower-ranked officers, who were most affected by the training, may have initially changed their behavior but then faced resistance or disincentives from within the hierarchy; our data show that they were less likely to be promoted than similar officers in control districts. Third, while less likely, it is possible that drivers changed their behavior after experiencing more law enforcement in treatment districts, and this may have reduced opportunities for officers to take action, either by issuing tickets or taking bribes, in both treatment and control districts.

Overall, we conclude that our ethics training program had a significant and lasting impact on police officers' attitudes, beliefs, and propensities to engage in unethical behavior, along with short-term effects on field behavior. While the localized setting and small sample size call for replication in other contexts, our study offers important insights into the potential of carefully designed ethics training programs centered around identity and intrinsic motivation to shift police officers' preferences and behaviors, even in highly corrupt environments. Our findings also suggest that for such changes to be sustained in the long term, reinforcement through periodic refresher trainings and meaningful support from leadership within the police hierarchy are important.

References

Abadie, A., S. Athey, G. W. Imbens, and J. M. Wooldridge (2023). When should you adjust standard errors for clustering? *The Quarterly Journal of Economics 138*(1), 1–35.

Abeler, J., A. Becker, and A. Falk (2014). Representative evidence on lying costs. *Journal of Public Economics 113*(5), 96–104.

Abeler, J., D. Nosenzo, and C. Raymond (2019). Preferences for truth-telling. *Econometrica 87*(4), 1115–1153.

Abril, V., E. Norza, S. M. Perez-Vincent, S. Tobón, and M. Weintraub (2023). Building trust in state actors: A multi-site experiment with the colombian national police.

Akerlof, G. A. and R. E. Kranton (2005). Identity and the economics of organizations. *Journal of Economic perspectives 19*(1), 9–32.

Akerlof, G. A. and R. E. Kranton (2008). Identity, supervision, and work groups. *American Economic Review 98*(2), 212–17.

Akerlof, R. (2016). "we thinking" and its consequences. *American Economic Review 106*(5), 415–19.

Ali, A. J., J. Fuenzalida, M. Gómez, and M. J. Williams (2021). Four lenses on people management in the public sector: An evidence review and synthesis. *Oxford Review of Economic Policy 37*(2), 335–366.

Anderson, M. L. (2008). Multiple inference and gender differences in the effects of early intervention: A reevaluation of the abecedarian, perry preschool, and early training projects. *Journal of the American statistical Association 103*(484), 1481–1495.

Ashraf, N., O. Bandiera, and B. K. Jack (2014). No margin, no mission? a field experiment on incentives for public service delivery. *Journal of public economics 120*, 1–17.

Ashraf, N., O. Bandiera, V. Minni, and L. Zingales (2024). Meaning at work.

Avis, E., C. Ferraz, and F. Finan (2018). Do government audits reduce corruption? estimating the impacts of exposing corrupt politicians. *Journal of Political Economy 126*(5), 1912–1964.

Banerjee, A., R. Chattopadhyay, E. Duflo, D. Keniston, and N. Singh (2021). Improving police performance in rajasthan, india: Experimental evidence on incentives, managerial autonomy, and training. *American Economic Journal: Economic Policy 13*(1), 36–66.

Banerjee, A., R. Hanna, and S. Mullainathan (2012). *27. Corruption*. Princeton University Press.

Banerjee, A. V., R. Banerji, E. Duflo, R. Glennerster, and S. Khemani (2010). Pitfalls of participatory programs: Evidence from a randomized evaluation in education in india. *American Economic Journal: Economic Policy 2*(1), 1–30.

Banerjee, R., T. Baul, and T. Rosenblat (2015). On self selection of the corrupt into the public sector. *Economics Letters 127*, 43–46.

Banuri, S. and P. Keefer (2016). Pro-social motivation, effort and the call to public service. *European Economic Review 83*, 139–164.

Barbosa, D. A., T. Fetzer, C. Soto, and P. C. Souza (2021). De-escalation technology: the impact of body-worn cameras on citizen-police interactions. Technical report, University of Warwick, Department of Economics.

Beeri, I., R. Dayan, E. Vigoda-Gadot, and S. B. Werner (2013). Advancing ethics in public organizations: The impact of an ethics program on employees' perceptions and behaviors in a regional council. *Journal of business ethics 112*(1), 59–78.

Belloni, A., V. Chernozhukov, and C. Hansen (2014). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies 81*(2), 608–650.

Bénabou, R. and J. Tirole (2011). Identity, morals, and taboos: Beliefs as assets. *The Quarterly Journal of Economics 126*(2), 805–855.

Besley, T. and M. Ghatak (2005). Competition and incentives with motivated agents. *American economic review 95*(3), 616–636.

Blair, G., J. M. Weinstein, F. Christia, E. Arias, E. Badran, R. A. Blair, A. Cheema, A. Farooqui, T. Fetzer, G. Grossman, et al. (2021). Community policing does not build citizen trust in police or reduce crime in the global south. *Science 374*(6571), eabd3446.

Blair, R. A., S. M. Karim, and B. S. Morse (2019). Establishing the rule of law in weak and war-torn states: Evidence from a field experiment with the liberian national police. *American Political Science Review 113*(3), 641–657.

Blattman, C., D. P. Green, D. Ortega, and S. Tobón (2021). Place-based interventions at scale: The direct and spillover effects of policing and city services on crime. *Journal of the European Economic Association 19*(4), 2022–2051.

Bloom, N., R. Lemos, R. Sadun, and J. Van Reenen (2020). Healthy business? managerial education and management in health care. *Review of Economics and Statistics 102*(3), 506–517.

Borcan, O., S. Heger, N. Grabher-Meyer, and A. Patel (2023). Right in the middle: A field experiment on the role of integrity training and norms in combating corruption. *University of East Anglia School of Economics Working Paper Series 2023-05*.

Borcan, O., M. Lindahl, and A. Mitrut (2014). The impact of an unexpected wage cut on corruption: Evidence from a "xeroxed" exam. *Journal of Public Economics 120*.

Borcan, O., M. Lindahl, and A. Mitrut (2017). Fighting corruption in education: What works and who benefits? *American Economic Journal: Economic Policy 9*(1), 180–209.

Bursztyn, L. and R. Jensen (2017). Social image and economic behavior in the field: Identifying, understanding, and shaping social pressure. *Annual Review of Economics 9*, 131–153.

Callen, M., S. Gulzar, A. Hasanain, M. Y. Khan, and A. Rezaee (2020). Data and policy decisions: Experimental evidence from pakistan. *Journal of Development Economics 146*, 102523.

Cameron, A. C. and D. L. Miller (2015). A practitioner's guide to cluster-robust inference. *Journal of human resources 50*(2), 317–372.

Canales, R., J. F. Santini, M. González Magaña, and A. Cherem (2025). Shaping police officer mindsets and behaviors: Experimental evidence of procedural justice training. *Management Science*.

Cassar, L. and J. Armouti-Hansen (2020). Optimal contracting with endogenous project mission. *Journal of the European Economic Association 18*(5), 2647–2676.

Chen, Y. and S. X. Li (2009). Group identity and social preferences. *American Economic Review 99*(1), 431–57.

Cohn, A. and M. A. Maréchal (2018). Laboratory measure of cheating predicts school misconduct. *The Economic Journal 128*(615), 2743–2754.

Cohn, A., M. A. Maréchal, and T. Noll (2015). Bad boys: How criminal identity salience affects rule violation. *The Review of Economic Studies 82*(4), 1289–1308.

Conover, E., D. Kraynak, and P. Singh (2023). The effect of traffic cameras on police effort: Evidence from india. *Journal of Development Economics 160*, 102953.

Dai, Z., F. Galeotti, and M. C. Villeval (2018). Cheating in the lab predicts fraud in the field: An experiment in public transportation. *Management Science 64*(3), 1081–1100.

Dal Bó, E., F. Finan, and M. A. Rossi (2013). Strengthening state capabilities: The role of financial incentives in the call to public service. *The Quarterly Journal of Economics 128*(3), 1169–1218.

Deserranno, E. (2019). Financial incentives as signals: experimental evidence from the recruitment of village promoters in uganda. *American Economic Journal: Applied Economics 11*(1), 277–317.

Dhaliwal, I. and R. Hanna (2017). The devil is in the details: The successes and limitations of bureaucratic reform in india. *Journal of Development Economics 124*, 1–21.

Dube, O., S. J. MacArthur, and A. K. Shah (2025). A cognitive view of policing. *The Quarterly Journal of Economics 140*(1), 745–791.

Duflo, E., R. Hanna, and S. P. Ryan (2012). Incentives work: Getting teachers to come to school. *American Economic Review 102*(4), 1241–78.

Dunsch, F. A., D. K. Evans, E. Eze-Ajoku, and M. Macis (2017). Management, supervision, and health care: A field experiment. Technical report, National Bureau of Economic Research.

Ferraz, C. and F. Finan (2008). Exposing corrupt politicians: the effects of brazil's publicly released audits on electoral outcomes. *The Quarterly journal of economics 123*(2), 703–745.

Finan, F., B. A. Olken, and R. Pande (2017). The personnel economics of the developing state. *Handbook of economic field experiments 2*, 467–514.

Fischbacher, U. and F. Föllmi-Heusi (2013). Lies in disguise—an experimental study on cheating. *Journal of the European Economic Association 11*(3), 525–547.

Fisman, R. and E. Miguel (2007). Corruption, norms, and legal enforcement: Evidence from diplomatic parking tickets. *Journal of Political economy 115*(6), 1020–1048.

Foltz, J. D. and K. A. Opoku-Agyemang (2015). Do higher salaries lower petty corruption? a policy experiment on west africa's highways. *Unpublished Working Paper, University of Wisconsin-Madison and University of California, Berkeley*.

Fracchia, M., T. Molina-Millán, and P. C. Vicente (2023). Motivating volunteer health workers in an african capital city. *Journal of Development Economics 163*, 103096.

Ghana Integrity Initiative, l. (2017). Press release: The corruption perception index 2016.

Hanna, R. and S.-Y. Wang (2017). Dishonesty and selection into public service: Evidence from india. *American Economic Journal: Economic Policy 9*(3), 262–90.

Heldring, L. (2021). The origins of violence in rwanda. *The Review of economic studies 88*(2), 730–763.

Heß, S. (2017). Randomization inference with stata: A guide and software. *The Stata Journal 17*(3), 630–651.

Hogg, M. A., D. Abrams, and M. B. Brewer (2017). Social identity: The role of self in group processes and intergroup relations. *Group Processes & Intergroup Relations 20*(5), 570–581.

Kajackaite, A. and U. Gneezy (2017). Incentives and cheating. *Games and Economic Behavior 102*, 433–444.

Kaptein, M. (2015). The effectiveness of ethics programs: The role of scope, composition, and sequence. *Journal of Business Ethics 132*(2), 415–431.

Kaptein, M. and M. S. Schwartz (2008). The effectiveness of business codes: A critical examination of existing studies and the development of an integrated research model. *Journal of Business Ethics 77*(2), 111–127.

Karim, S. (2020). Relational state building in areas of limited statehood: Experimental evidence on the attitudes of the police. *American Political Science Review 114*(2), 536–551.

Khan, A. Q., A. I. Khwaja, and B. A. Olken (2019). Making moves matter: Experimental evidence on incentivizing bureaucrats through performance-based postings. *American Economic Review 109*(1), 237–70.

Khan, M. Y. (2020). Mission motivation and public sector performance: Experimental evidence from pakistan.

Kumasey, A. S., J. N. Bawole, and F. Hossain (2017). Organizational commitment of public service employees in ghana: do codes of ethics matter? *International Review of Administrative Sciences 83*(1_suppl), 59–77.

Lameke, A. A., A. M. Nkuku, R. S. de la Sierra, V. Tanutama, and K. Titeca (2023). On the governance of corrupt exchange: how citizens and officials build social ties to reduce corruption's transaction costs. Technical report, National Bureau of Economic Research.

Linos, E. (2018). More than public service: A field experiment on job advertisements and diversity in the police. *Journal of Public Administration Research and Theory 28*(1), 67–85.

Lowes, S., N. Nunn, J. A. Robinson, and J. L. Weigel (2017). The evolution of culture and institutions: Evidence from the kuba kingdom. *Econometrica 85*(4), 1065–1091.

Norman, I. D., D. Dzidzonu, M. A. Aviisah, F. Norvivor, W. Takramah, M. Kweku, et al. (2017). The incidence of money collected by the ghana police from drivers during routine traffic stops and ad hoc road blocks. *Advances in applied sociology 7*(05), 197.

Olken, B. A. (2007). Monitoring corruption: evidence from a field experiment in indonesia. *Journal of political Economy 115*(2), 200–249.

Olken, B. A. and R. Pande (2012). Corruption in developing countries. *Annu. Rev. Econ. 4*(1), 479–509.

Owens, E. and B. Ba (2021). The economics of policing and public safety. *Journal of Economic Perspectives 35*(4), 3–28.

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

Park, H. and J. Blenkinsopp (2013). The impact of ethics programmes and ethical culture on misconduct in public service organizations. *International Journal of Public Sector Management*.

Peyton, K., M. Sierra-Arévalo, and D. G. Rand (2019). A field experiment on community policing and police legitimacy. *Proceedings of the National Academy of Sciences 116*(40), 19894–19898.

Potters, J. and J. Stoop (2016). Do cheaters in the lab also cheat in the field? *European Economic Review 87*, 26–33.

Pring, C. and J. Vrushi (2019). Global corruption barometer: Africa 2019. *Transparency International*.

Rasul, I. and D. Rogger (2018). Management of bureaucrats and public service delivery: Evidence from the nigerian civil service. *The Economic Journal 128*(608), 413–446.

Romano, J. P. and M. Wolf (2005). Stepwise multiple testing as formalized data snooping. *Econometrica 73*(4), 1237–1282.

Roodman, D., M. Ørregaard Nielsen, J. G. MacKinnon, and M. D. Webb (2019). Fast and wild: Bootstrap inference in stata using boottest. *The Stata Journal 19*(1), 4–60.

Sánchez De La Sierra, R., K. Titeca, H. Xie, A. A. Lameke, and A. Malukisa Nkuku (2024). The real state: Inside the congo's traffic police agency. *American Economic Review 114*(12), 3976–4014.

Serra, D. and L. Wantchekon (2012). Experimental research on corruption: Introduction and overview. In *New advances in experimental research on corruption*. Emerald Group Publishing Limited.

Sowatey, E. A. and J. Tankebe (2019). Doing research with police elites in ghana. *Criminology & Criminal Justice 19*(5), 537–553.

Tajfel, H. (1974). Social identity and intergroup behaviour. *Social science information 13*(2), 65–93.

Tankebe, J. (2010). Public confidence in the police: Testing the effects of public experiences of police corruption in ghana. *The British Journal of Criminology 50*(2), 296–319.

Urminsky, O., C. Hansen, and V. Chernozhukov (2016). Using double-lasso regression for principled variable selection. *Available at SSRN 2733374*.

Warren, D. E., J. P. Gaspar, and W. S. Laufer (2014). Is formal ethics training merely cosmetic? a study of ethics training and ethical organizational culture. *Business Ethics Quarterly 24*(1), 85–117.

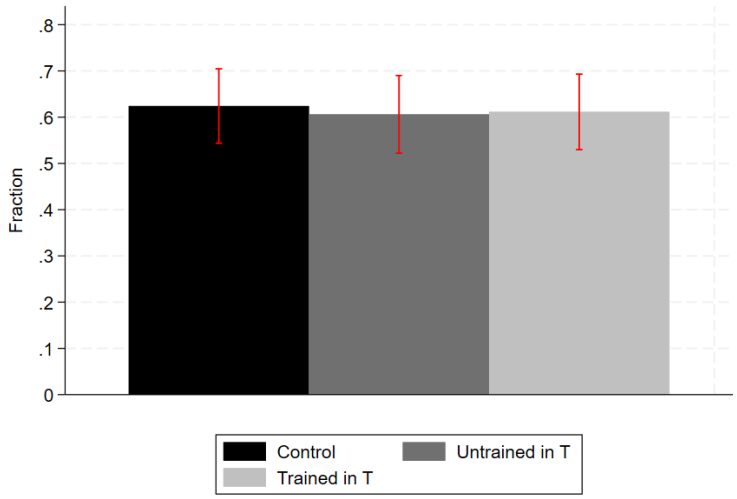
Weisel, O. and S. Shalvi (2015). The collaborative roots of corruption. *Proceedings of the National Academy of Sciences 112*(34), 10651–10656.

Williams Jr, M. C., N. Weil, E. A. Rasich, J. Ludwig, H. Chang, and S. Egrari (2021). Body-worn cameras in policing: Benefits and costs.

Wood, G., P. A. V. R. J. Tyler, Tom R, and P. H. Sant'Anna (2021). Revised findings for "procedural justice training reduces police use of force and complaints against officers". *Proceedings of the National Academy of Sciences 118*(27).

FIGURES AND TABLES

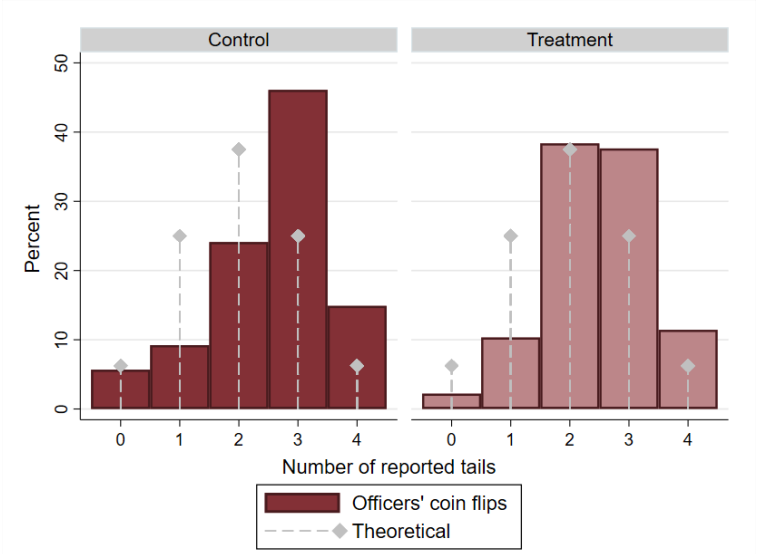
Figure 1: Behavior in the Baseline Incentivized cheating Game



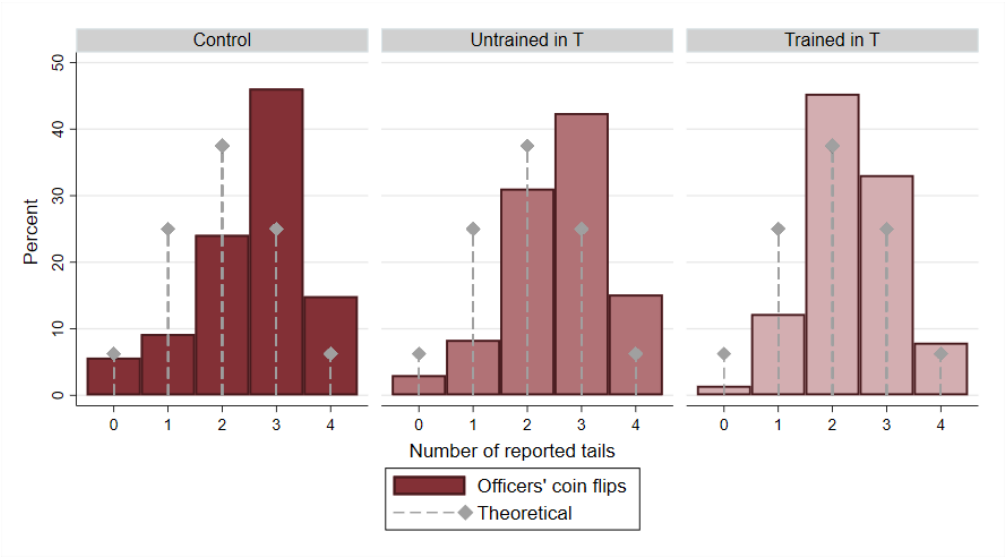
Notes: At baseline (October 2018) all survey officers engaged in an incentivized Mind Dice Game. Officers had to think of a number, roll a dice and report whether the number they obtained was equal to the number they had thought of. Reporting a match would generate them monetary earnings. The figure shows the percentages of Control officers (black) and Treatment officers – differentiated by whether they subsequently received the training – who reported a match between the number they thought of and the number they rolled. The Mind Game was conducted about 6 months prior to the implementation of the training intervention.

Figure 2: Officers' Behavior in the Endline Incentivized Cheating Game

(a) Control vs. Treatment Officers



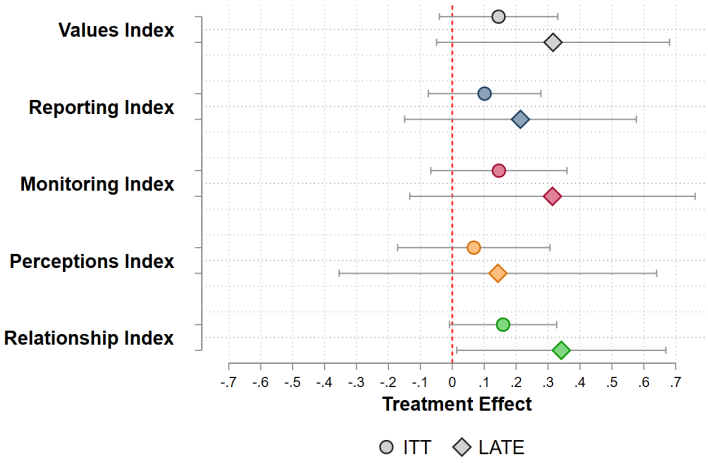
(b) Control vs Treatment Officers, by Training Status



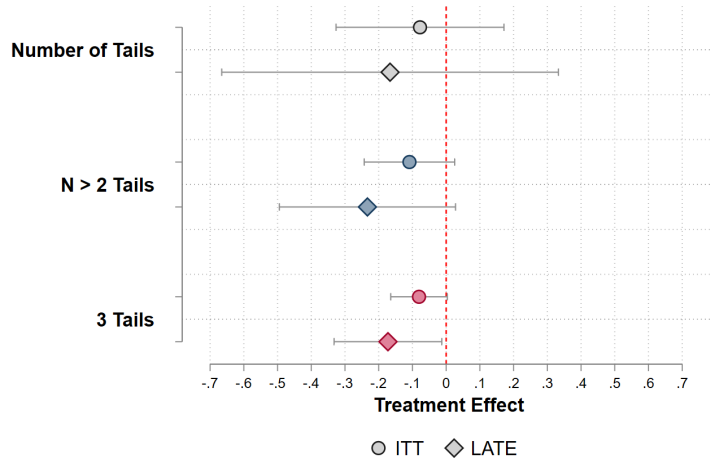
Notes: The figures show the distributions of the number of tails reported by the officers in the incentivized cheating game at endline. The officers had to flip a coin 4 times, and report the number of tails they obtained, with higher numbers leading to higher earnings. In the top figure, we report the empirical distributions of coin tosses generated by officers in Control versus Treatment districts. In the bottom figure, we differentiate between trained and untrained officers in Treatment districts. In gray, we display the theoretical binomial distribution of coin flips.

Figure 3: Estimated Impacts of the Training Program

(a) Survey Outcomes

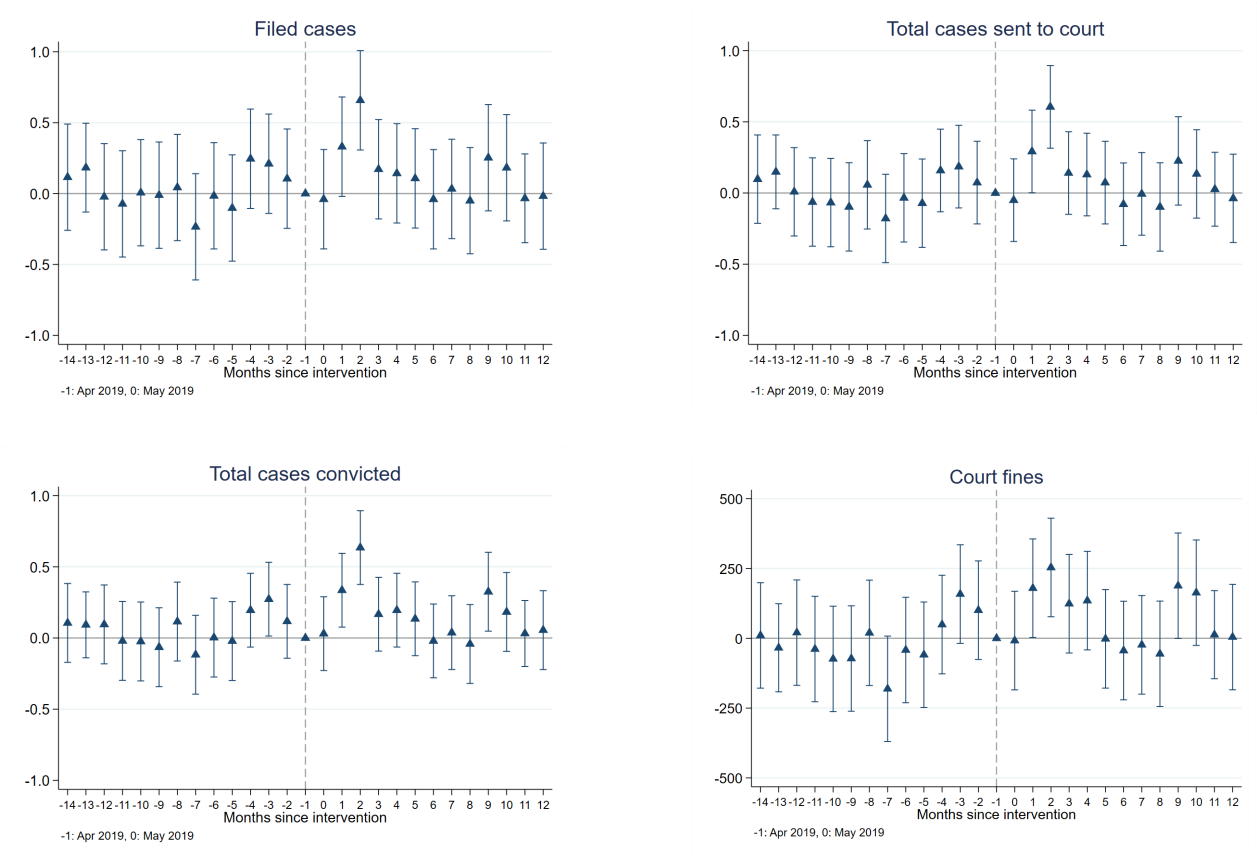


(b) Cheating Game Outcomes



Notes: The figures display the estimated coefficients obtained for each of the listed outcomes, along with their 90 percent confidence intervals, as derived from the regression analysis reported in Tables 2 and 3. The survey outcomes are aggregated indexes generated by survey answers of police officers during the endline phone interviews; they are calculated as weighted averages of the underlying variables, where the weights are inversely related to the degree of correlation between the variables (see Anderson, 2008). The survey questions in each index can be found in Table A2 in Appendix A. All indexes are standardized around their Control mean; estimates are expressed in standard deviations from the control mean. Panel (b) displays the estimated treatment impacts on the number reported in the cheating game, the likelihood of reporting a number greater than 2, and the likelihood of reporting the number 3. For each outcome, we display ITT and LATE estimates, corresponding to Panels A and B of the corresponding regression table. The LATE estimates are generated by IV regressions, where the instrument is the random assignment of districts to treatment or control groups. Standard errors are clustered at the police district level, and the specifications generating the reported estimates include Double Lasso controls.

Figure 4: Event Study Estimates of Treatment Effects on Field Behavior



Notes: The figures report the estimated impacts of the training from an event study (which include time and district fixed effects) over 60 months, of which 27 pre-intervention and 32 post-intervention. For ease of interpretation, the figures report estimates for 13 months pre-intervention and 12 months post-intervention. Note that the 12th month after the intervention corresponds to the surge of the COVID-19 pandemic. The intervention took place the last week of April 2019 and the first week of May. We set May 2019 as "0" and April 2019 as "-1," i.e., the omitted comparison month. The four dependent variables are monthly district-level outcomes, averaged over the number of officers in the district. The figure on the top left displays the estimated impact of the training on the number cases filed on average per officer in a district, for each of the months before and after the intervention. The figure on the top right displays the estimated impact of the training on the monthly cases sent to court on average per officer in a district. The figure on the bottom left displays the estimated impact of the training on the number of monthly convicted cases on average per officer in a district. The figure on the bottom right displays the estimated impact of the training on the total monthly fines issued to drivers on average per officer in a district (in Ghanaian Cedi).

Table 1: Descriptive Statistics

	Mean C	Mean T	p-value (1-2)	Mean Untrainted (T)	Mean Trained (T)	p-value (4-5)
	(1)	(2)	(3)	(4)	(5)	(6)
District size at baseline	18.636	20.056	0.777	-	-	-
Panel A - Demographics						
Female	0.277	0.255	0.668	0.295	0.216	0.134
Age (1/5: 3=35-44 years)	3.014	3.059	0.714	3.182	2.942	0.022**
Year graduated (baseline)	2003.454	2002.550	0.462	2001.765	2003.295	0.119
Officer rank (baseline)	3.589	3.694	0.640	3.871	3.525	0.083*
Wage (baseline)	1939.582	1962.328	0.811	1936.280	1987.065	0.492
Intrinsic motivation to serve the community y/n (baseline)	0.631	0.594	0.471	0.598	0.590	0.887
Panel B - Outcome measures at baseline						
Education most effective against corruption y/n (baseline)	0.121	0.140	0.583	0.159	0.122	0.385
Comfortable reporting corruption 1/5 (baseline)	3.170	3.269	0.609	3.470	3.079	0.029**
Know where to report corruption 1/5 (baseline)	4.312	4.380	0.557	4.379	4.381	0.978
Ever reported unethical behaviour y/n (baseline)	0.213	0.365	0.003***	0.326	0.403	0.189
Police corrupt 1/5 (baseline)	2.546	2.491	0.760	2.417	2.561	0.359
Witness unethical behaviour 1/4 (baseline)	2.723	2.897	0.177	2.795	2.993	0.090*
Witness bribe 1/4 (baseline)	2.390	2.668	0.073*	2.576	2.755	0.171
Police as Service Provider y/n (baseline)	0.333	0.280	0.275	0.280	0.281	0.996
Aggressive more useful than curteous 1/5 (baseline)	3.220	3.196	0.908	3.174	3.216	0.811
Overall workplace satisfaction 1/5 (baseline)	3.638	3.705	0.525	3.788	3.626	0.170
Panel C - Behavior in the baseline cheating game						
Mind Game: Number matched	0.624	0.609	0.765	0.606	0.612	0.927
Observations (max)	141	271		139	132	
Panel D - Outcome measures at baseline: reduced sample						
How often monitor juniors 1/4 (baseline)	3.566	3.656	0.442	3.719	3.590	0.270
Caught juniors' unethical behaviour 1/4 (baseline)	2.453	2.704	0.124	2.594	2.820	0.177
Disciplined juniors 1/4 (baseline)	1.774	2.272	0.006***	2.375	2.164	0.267
Observations (max)	53	125		61	64	

Notes: The table displays officer characteristics and answers to survey questions of interest during the baseline data collection (October 2018) p-values in column (3) are based on clustered standard errors (at police district level). The measures in Panel D are for a subsample of officers who reported leading a team of lower-rank officers.

Table 2: Treatment effects on survey-generated outcomes

	Values Index		Reporting Index		Monitoring Index		Perceptions Index		Relationship Index	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: ITT estimates										
Treatment District	0.171*	0.145	0.152*	0.101	0.123	0.146	0.127	0.068	0.178*	0.159*
	(0.096)	(0.091)	(0.085)	(0.086)	(0.099)	(0.105)	(0.132)	(0.117)	(0.098)	(0.082)
	[0.194]	[0.221]	[0.194]	[0.275]	[0.272]	[0.249]	[0.272]	[0.484]	[0.194]	[0.152]
	{0.118}	{0.153}	{0.097}	{0.270}	{0.243}	{0.185}	{0.402}	{0.585}	{0.126}	{0.092}
Control Mean	0.000	0.000	0.000	0.000	-0.000	-0.000	-0.000	-0.000	0.000	0.000
Observations	412	412	412	412	248	248	412	412	412	412
LASSO Controls	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Panel B: LATE estimates										
Treatment District	0.371*	0.316*	0.315*	0.214	0.261	0.314	0.270	0.143	0.381*	0.342**
	(0.194)	(0.186)	(0.180)	(0.185)	(0.206)	(0.228)	(0.285)	(0.254)	(0.208)	(0.167)
	[0.153]	[0.201]	[0.153]	[0.310]	[0.260]	[0.282]	[0.270]	[0.505]	[0.153]	[0.119]
	{0.095}	{0.133}	{0.104}	{0.284}	{0.236}	{0.194}	{0.416}	{0.601}	{0.122}	{0.077}
Control Mean	0.043	0.043	-0.015	-0.015	0.001	0.001	-0.003	-0.003	-0.007	-0.007
Observations	412	412	412	412	248	248	412	412	412	412
LASSO Controls	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes

Notes: Panel A displays OLS estimates where the dependent variables are aggregated indexes generated by survey answers of police officers during the endline phone interviews. The indexes are calculated as weighted averages of the underlying variables, where the weights are inversely related to the degree of correlation between the underlying variables (see Anderson 2008). The survey questions entering each index can be found in Table A2 in Appendix A. All indexes are standardized around their Control mean. Therefore, estimates are expressed in standard deviations from the control mean. Panel B displays IV regression estimates. The instrument is the random assignment of districts to treatment or control groups. Standard errors clustered at the police district level are reported in parentheses. Multiple hypothesis corrected p-values using the Romano-Wolf (2005) procedure are reported in square brackets. Due to small number of clusters (32), we report Wild-bootstrapped p-values in curly brackets. The controls are selected using the Double Lasso procedure (Belloni et al., 2014; Urminsky et al., 2016) from the full set of covariates displayed in Table 1. The available baseline variables forming the index that is used as the dependent variable are always included in the set of controls.

Table 3: Treatment effects on game-generated outcomes

	Number of tails		Coin flip: N > 2 tails (y/n)		3 tails (y/n)	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: ITT estimates						
Treatment District	-0.096	-0.078	-0.119*	-0.109	-0.085*	-0.080*
	(0.123)	(0.122)	(0.068)	(0.066)	(0.045)	(0.041)
	[0.341]	[0.440]	[0.088]	[0.100]	[0.088]	[0.080]
	{0.472}	{0.562}	{0.126}	{0.138}	{0.101}	{0.082}
Control Mean	2.553	2.553	0.610	0.610	0.461	0.461
Observations	412	412	412	412	412	412
LASSO Controls	No	Yes	No	Yes	No	Yes
Panel B: LATE estimates						
Treatment District	-0.203	-0.166	-0.253*	-0.233*	-0.180**	-0.173**
	(0.256)	(0.255)	(0.139)	(0.133)	(0.090)	(0.082)
	[0.332]	[0.437]	[0.072]	[0.076]	[0.072]	[0.057]
	{0.470}	{0.558}	{0.111}	{0.120}	{0.081}	{0.061}
Control Mean	2.553	2.553	0.610	0.610	0.461	0.461
Observations	412	412	412	412	412	412
LASSO Controls	No	Yes	No	Yes	No	Yes

Notes: Panel A displays OLS estimates where the dependent variables are the outcomes reported by police officers in the incentivized 4-time coin toss game. In columns 1 and 2, the dependent variable is the number or reported tails, ranging from 0 to 4. In columns 3 and 4, the dependent variable is a dummy equal to 1 if the officers reported more than 2 tails, and 0 otherwise. In columns 5 and 6, the dependent variable is a dummy equal to 1 if the officer reported 3 tails, and 0 otherwise. Panel B displays IV regression estimates. The instrument is the random assignment of districts to treatment or control groups. Standard errors clustered at the police district level are reported in parentheses. Multiple hypothesis corrected p-values using the Romano-Wolf (2005) procedure are reported in square brackets. Due to small number of clusters (32), we report Wild-bootstrapped p-values in curly brackets. The controls are selected using the Double Lasso procedure (Belloni et al., 2014; Urminsky et al., 2016) from the full set of covariates displayed in Table 1. The set of controls always include behavior in the baseline mind game, i.e., whether the officers stated that they rolled a number that matched the one in their head.

Table 4: Treatment Effects: Trained versus Untrained Officers in Treatment Districts

	Values Index	Reporting Index	Monitoring Index	Perceptions Index	Relationship Index	Number of tails	N> 2 tails (y/n)	3 tails (y/n)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Trained in T	0.273**	0.135	0.154	0.127	0.238**	-0.162	-0.169**	-0.116**
	(0.100)	(0.091)	(0.129)	(0.146)	(0.103)	(0.145)	(0.073)	(0.046)
	[0.052]	[0.500]	[0.708]	[0.815]	[0.106]	[0.431]	[0.058]	[0.051]
	{0.021}	{0.171}	{0.241}	{0.418}	{0.063}	{0.323}	{0.071}	{0.036}
Untrained in T	-0.032	-0.074	0.140	0.026	0.050	0.092	0.005	-0.006
	(0.105)	(0.119)	(0.133)	(0.155)	(0.094)	(0.136)	(0.069)	(0.051)
	[0.913]	[0.898]	[0.782]	[0.913]	[0.898]	[0.657]	[0.976]	[0.976]
	{0.767}	{0.548}	{0.322}	{0.868}	{0.627}	{0.524}	{0.948}	{0.900}
Control Mean	0.000	0.000	-0.000	-0.000	0.000	2.553	0.610	0.461
Trained=Untrained (p-val)	0.005	0.060	0.925	0.585	0.035	0.030	0.002	0.030
Observations	412	412	248	412	412	412	412	412
LASSO Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The tables displays OLS estimates for trained and untrained officers in treatment districts as compared to officers in Control districts. The dependent variables in columns 1 to 5 are aggregated indexes generated by survey answers of police officers during the endline phone interviews. The indexes are calculated as weighted averages of the underlying variables, where the weights are inversely related to the degree of correlation between the underlying variables (see Anderson 2008). The survey questions entering each index can be found in Table A2 in Appendix A. All indexes are standardized around their Control mean. Therefore, estimates are expressed in standard deviations from the control mean. In columns 6, 7 and 8, the dependent variables are the outcomes reported by police officers in the incentivized 4-time coin toss game. In columns 6, the dependent variable is the number or reported tails, ranging from 0 to 4. In columns 7, the dependent variable is a dummy equal to 1 if the officers reported more than 2 tails. In columns 8, the dependent variable is a dummy equal to 1 if the officer reported 3 tails. Standard errors clustered at the police district level are reported in parentheses. Multiple hypothesis corrected p-values using the Romano-Wolf (2005) procedure are reported in square brackets. The correction applies separately to the hypotheses tested in columns 1 to 5, and those tested in columns 6 to 8, i.e., each family of outcomes. Due to small number of clusters (32), we report Wild-bootstrapped p-values in curly brackets. The controls are selected using the Double Lasso procedure (Belloni et al., 2014; Urminsky et al., 2016) from the full set of covariates displayed in Table 1. The available baseline variables forming the index that is used as the dependent variable are always included in the set of controls.

Table 5: Comparing Baseline Officer Characteristics in Districts with and without *Clean* Administrative Data

	No Admin Data	Admin Data	p-value
District size at baseline	22.600	16.471	0.163
Panel A - Demographics			
Female	0.253	0.274	0.669
Age (1/5: 3=35-44 years)	3.000	3.103	0.381
Year graduated (baseline)	2003.160	2002.451	0.555
Officer rank (baseline)	3.578	3.766	0.386
Wage (baseline)	1923.599	1996.451	0.426
Intrinsic motivation to serve the community y/n (baseline)	0.629	0.577	0.298
Panel B - Outcome measures at baseline			
Education most effective against corruption y/n (baseline)	0.148	0.114	0.333
Comfortable reporting corruption 1/5 (baseline)	3.203	3.280	0.681
Know where to report corruption 1/5 (baseline)	4.376	4.331	0.697
Ever reported unethical behaviour y/n (baseline)	0.295	0.337	0.454
Police corrupt 1/5 (baseline)	2.422	2.629	0.229
Witness unethical behaviour 1/4 (baseline)	2.819	2.863	0.729
Witness bribe 1/4 (baseline)	2.506	2.663	0.315
Police as Service Provider y/n (baseline)	0.300	0.297	0.958
Aggressive more useful than curteous 1/5 (baseline)	3.165	3.257	0.650
Overall workplace satisfaction 1/5 (baseline)	3.671	3.697	0.795
Panel C - Behavior in the baseline cheating game			
Mind Game: Number matched	0.616	0.611	0.925
Panel D - Outcome measures at baseline: reduced sample			
How often monitor juniors 1/4 (baseline)	3.615	3.644	0.796
Caught juniors' unethical behaviour 1/4 (baseline)	2.582	2.678	0.516
Disciplined juniors 1/4 (baseline)	2.242	2.000	0.142

Notes: We obtained *clean* monthly district-level administrative data for 17 districts, 4 in the Control and 13 in the Treatment group, for 60 months, from January 2017 to December 2021. We also obtained data for the remaining 15 districts, but such data are aggregated at a higher administrative level, which pulls together Control and Treatment districts in our sample, or adds districts that do not have traffic police units, and for which we do not have officer-level or other data at baseline or endline. This table compares district size (i.e., number of officers) and officer characteristics, as measured through our baseline survey, for the districts for which we have clean administrative data, and for those for which we do not have such data. The variables in the table are the same as those in Table 1. The sample size is all the 475 officers surveyed at baseline, including those for whom we do not have endline data.

Table 6: Treatment effects on District-Level Administrative Data (per officer per month)

	Total filed cases	Total cases sent to court	Total cases convicted	Court fines
	(1)	(2)	(3)	(4)
Panel A				
Post	0.092	0.087	0.074	42.334
	(0.071)	(0.059)	(0.054)	(36.465)
Post x Training	0.018	0.020	0.029	10.456
	(0.051)	(0.043)	(0.039)	(26.308)
	[0.716]	[0.638]	[0.468]	[0.750]
Pre-Control Mean	0.19	0.14	0.12	65.24
Observations	660	660	660	660
Analytical Weights	Yes	Yes	Yes	Yes
Month and Yr FE	Yes	Yes	Yes	Yes
Admin unit FE	Yes	Yes	Yes	Yes
Panel B				
1-3 months post	0.018	0.027	0.026	1.660
	(0.114)	(0.095)	(0.086)	(58.403)
1-3 months post x Training	0.256**	0.242**	0.237***	126.382**
	(0.119)	(0.099)	(0.089)	(60.904)
	[0.070]	[0.084]	[0.087]	[0.070]
4-6 months post	0.070	0.053	0.029	-11.392
	(0.114)	(0.095)	(0.086)	(58.403)
4-6 months post x Training	0.075	0.087	0.107	122.917**
	(0.119)	(0.099)	(0.089)	(60.904)
	[0.515]	[0.374]	[0.252]	[0.071]
After 6 months post	-0.061	-0.064	-0.074	-45.333
	(0.093)	(0.077)	(0.070)	(47.651)
After 6 months post x Training	-0.016	-0.013	-0.004	-15.897
	(0.053)	(0.044)	(0.040)	(27.305)
	[0.746]	[0.756]	[0.924]	[0.676]
Pre-Control Mean	0.19	0.14	0.12	65.24
Observations	660	660	660	660
Analytical Weights	Yes	Yes	Yes	Yes
Month and Yr FE	Yes	Yes	Yes	Yes
Admin unit FE	Yes	Yes	Yes	Yes

Notes: Panel A reports estimates from a standard two-period difference-in-differences (DiD) model. Panel B reports DiD estimates over multiple time periods. In both models, we include time (month and year) fixed effects, as well as district fixed effects. Since data for 10 districts are aggregated into 4 administrative units, we conduct the analysis at the administrative unit level, i.e., we analyze data for 11 administrative units for 60 months, leading to 660 observations. The four dependent variables are monthly district-level outcomes, averaged over the number of officers in the district (or administrative unit). Given that the data are aggregated at the district level, and averaged by the number of officers, N, in a district, yet districts differ in N, we use analytical weights where, for each district, the weight is the number of officers in a district. "Total filed cases" refers to the monthly number cases filed on average per officer in a given district or administrative unit. "Total cases sent to court" refers to the monthly cases sent to court on average per officer in a given district or administrative unit. "Total cases convicted" refers to the number of convicted cases on average per officer in a given district or administrative unit. "Court fines" refers to the monthly amount of issued fines (in Ghanaian Cedi) on average per officer in a given district/administrative unit. Standard errors reported in parentheses. Wild Clustered Bootstrap (Restricted) p-values are reported in square brackets (as per Roodman et al. (2019)) to account for the small number of districts or administrative unit.

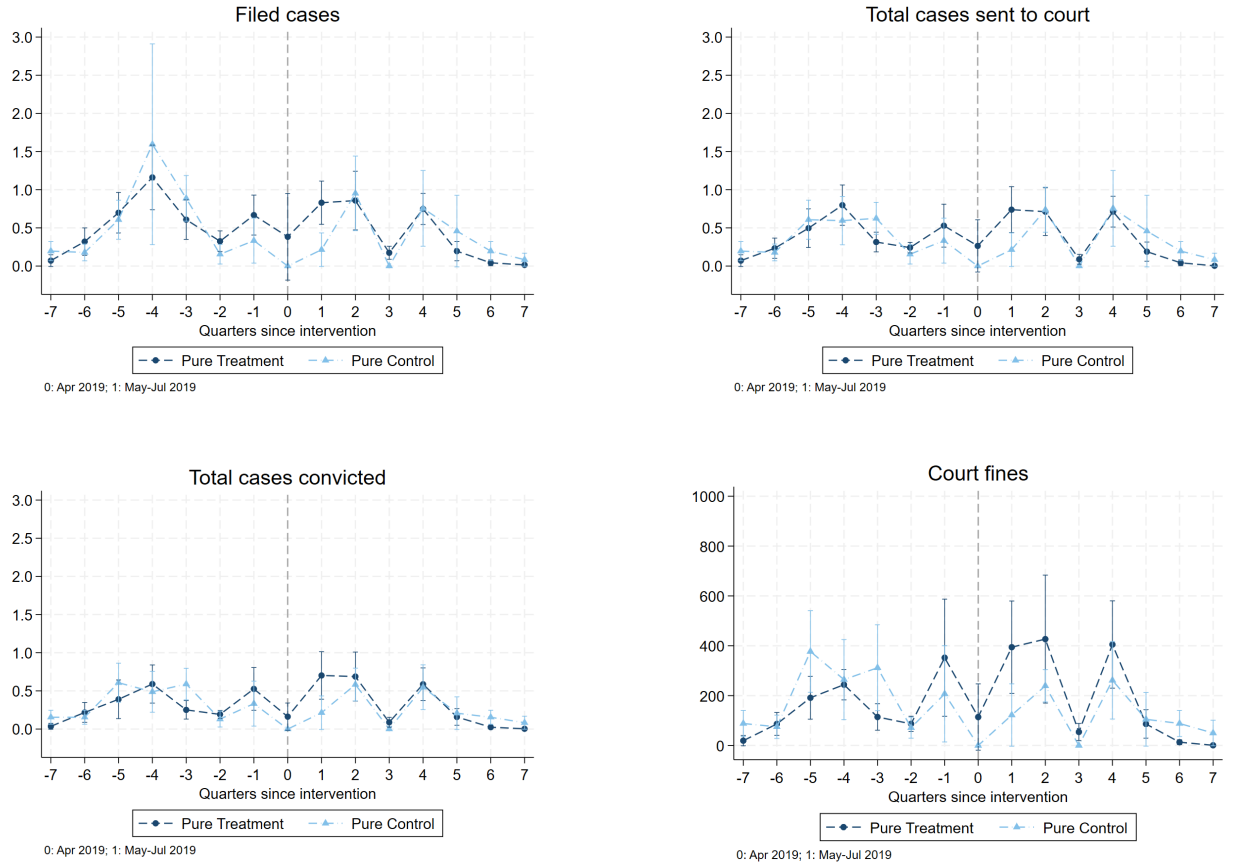
APPENDIX A: FIGURES AND TABLES

Figure A1: Agent of Change pin delivered to trained officers at the award ceremony

Figure A2: Survey-based indexes

Notes: The figure shows the means and 95% confidence intervals of four survey-generated indexes, for trained and untrained officers in treatment police districts. The aggregated indexes are generated by survey answers of police officers during the endline phone interviews. The indexes are calculated as weighted averages of the underlying variables, where the weights are inversely related to the degree of correlation between the underlying variables (see Anderson 2008). Each index is expressed in standard deviations from the standardized Control mean (equal to 0). Table A2 in Appendix A reports the individual questions constituting each index.

Figure A3: District-level Administrative Data on Field Behavior



Notes: Each panel of the figure reports quarterly district-level administrative data for cases filed by traffic police officers, cases sent to court, total cases convicted and total court fines, averaged by the number of officers in a district (as recorded at baseline, in October 2018). The intervention took place the last week of April and the first week of May 2019. We set April 2019 as the "0" in the figure, and we include May 2019 in the first quarter post-intervention. "-7" refers to the quarter of July to September 2017, and "7" refers to the quarter of November 2020 to January 2021. The "5" quarter corresponds to the beginning of the COVID-19 pandemic. The figures are based on data for 17 districts, 4 in the Control and 13 in the Treatment group. Data for other districts are aggregated at a higher level, which pulls together Control and Treatment districts in our sample, or adds districts without traffic police units, for which we have no officer or other data. We refer to the 17 district for which we have clean administrative data as "Pure Control" and "Pure Treatment" in the figure. "Filed cases" are the cases filed on average by a traffic police officer in a district in a given quarter. "Total cases sent to court" are the cases subsequently sent to court by an officer on average in a district per quarter. "Total cases convicted" refer to the per-officer average number of cases that are convicted and that originate from a given district. "Court fines" are the quarterly fines, in Ghanaian Cedi, issued on average by an officers to drivers in a district in a given quarter.

Table A1: Likelihood of Attrition in Treatment versus Control Districts

	Attrited at endline	
	(1)	(2)
Treatment District	-0.009	-0.013
	(0.034)	(0.037)
	{0.826}	{0.771}
Observations	479	479
Control Mean	0.145	0.145
Controls	No	Yes

Notes: The table displays OLS regressions where the dependent variable is a 0-1 dummy, equal to 1 if the officer was interviewed at baseline but not at endline. Controls include age, gender, rank at baseline and whether or not the officer reported a match in the baseline incentivized cheating game. Standard errors, clustered at district level, are reported in parentheses. Wild bootstrap standard errors, to account for the small number of clusters (N=32) are reported in curly brackets.

Table A2: Aggregate survey-based indexes

Index	Underlying survey questions
Values and Beliefs Index	- What are the most important qualities of a police officers? (Answered "being honest and professional") - Can organizational norms be changed? (Answered "yes") - Do you see yourself more as a crime preventer, a law enforcer or a service provider for the citizen? (Answered "service provider") - What do you think would be the most effective way of tackling corruption? (Answered "Sensitize/educate the people on evils of corruption")
Reporting Index	- I know where to report unethical behavior (5 point Likert agree/disagree scale) - Have you ever reported unethical behavior? (Answered "Yes") - How often do you talk about unethical behavior with colleagues? (4-point Likert scale)
Monitoring Index . (Officers who manage lower-rank officers)	- How often do you monitor juniors? (4-point Likert scale); - Over the past 24 months, have you caught officers behaving unethically?(4-point Likert scale); - Over the past 24 months, have you disciplined/punished junior officers for behaving unethically? (4-point Likert scale).
Perceptions Index	- Do you think that in your police district corruption is not a problem, a small problem, a moderate problem, a quite serious problem, or a very serious problem?; - There have been several reports that brand the police as corrupt. To what extent do you agree with that assessment? (5-point Likert scale) - How often have you witnessed (or heard about) a colleague accepting a bribe from a citizen over the past 24 months? (4-point Likert scale)
Citizens Relationship Index	- In your opinion, how is the relationship between police officers and citizens this year compared to last year? (5-point Likert scale) - In certain situations, it is more useful for an officer to be aggressive than to be courteous (5 point Likert scale)

Notes: The table displays the survey questions that form each index. The full endline questionnaire could be found in the Online Appendix.

Table A3: Treatment Effects on Job Satisfaction

	Job Satisfaction	
	(1)	(2)
Panel A: ITT estimates		
Treatment District	0.161	0.178*
	(0.115)	(0.099)
	{0.175}	{0.074}
Control Mean	-0.000	-0.000
Observations	412	412
LASSO Controls	No	Yes
Panel B: LATE estimates		
Trained in T	0.314	0.354*
	(0.226)	(0.199)
	{0.181}	{0.080}
Control Mean	-0.000	-0.000
Observations	412	412
LASSO Controls	No	Yes

Notes: The table displays OLS (Panel A and C) and IV estimates (Panel B) where the dependent variable is the rating of job satisfaction (measured on a Likert 1-5 scale, from least to most satisfied). The job satisfaction is standardized around the control group mean. Therefore, estimates are expressed in standard deviations from the control mean. "Treatment District" is an indicator equal to 1 for officers in Treatment districts, and 0 for officers in Control districts. Standard errors clustered at the police district level are reported in parentheses. Due to the small number of clusters (32), we report Wild-bootstrapped p-values in curly brackets. The controls are selected using the Double Lasso procedure (Belloni et al., 2014; Urminsky et al., 2016) from the full set of covariates displayed in Table 1.

Table A4: Randomization Inference

	Values Index	Reporting Index	Monitoring Index	Perceptions Index	Relationship Index	Number of tails	N> 2 tails (y/n)	3 tails (y/n)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: ITT estimates								
Treatment District	0.145	0.101	0.146	0.068	0.159*	-0.078	-0.109	-0.080*
	(0.091)	(0.086)	(0.105)	(0.117)	(0.082)	(0.122)	(0.066)	(0.041)
	[0.132]	[0.318]	[0.189]	[0.633]	[0.093]	[0.534]	[0.160]	[0.104]
Control Mean	0.000	0.000	-0.000	-0.000	0.000	2.553	0.610	0.461
Observations	412	412	248	412	412	412	412	412
LASSO Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Panel B: Trained vs Untrained in Treatment								
Trained in T	0.273**	0.135	0.154	0.127	0.238**	-0.162	-0.169**	-0.116**
	(0.100)	(0.091)	(0.129)	(0.146)	(0.103)	(0.145)	(0.073)	(0.046)
	[0.005]	[0.166]	[0.210]	[0.221]	[0.026]	[0.107]	[0.000]	[0.025]
Untrained in T	-0.032	-0.074	0.140	0.026	0.050	0.092	0.005	-0.006
	(0.105)	(0.119)	(0.133)	(0.155)	(0.094)	(0.136)	(0.069)	(0.051)
	[0.762]	[0.508]	[0.289]	[0.809]	[0.644]	[0.344]	[0.929]	[0.882]
Control Mean	0.000	0.000	-0.000	-0.000	0.000	2.553	0.610	0.461
Trained=Untrained (p-val)	0.005	0.060	0.925	0.585	0.035	0.030	0.002	0.030
Observations	412	412	248	412	412	412	412	412
LASSO Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The table reports p-values obtained by conducting Randomization Inference (Heß, 2017) in square brackets. Clustered standard errors are reported in parentheses. Panel A reports estimates and p-values corresponding to our ITT analysis. Panel B reports estimates and p-values corresponding to the analysis of trained and untrained officers in treatment districts and opposed to officers in control districts. The dependent variables in columns 1 to 5 are aggregated indexes generated by survey answers of police officers during the endline phone interviews. The indexes are calculated as weighted averages of the underlying variables, where the weights are inversely related to the degree of correlation between the underlying variables (see Anderson 2008). The survey questions entering each index can be found in Table A2 in Appendix A. All indexes are standardized around their Control mean. Therefore, estimates are expressed in standard deviations from the control mean. In columns 6, 7 and 8, the dependent variables are the outcomes reported by police officers in the incentivized 4-time coin toss game. In columns 6, the dependent variable is the number or reported tails, ranging from 0 to 4. In columns 7, the dependent variable is a dummy equal to 1 if the officers reported more than 2 tails. In columns 8, the dependent variable is a dummy equal to 1 if the officer reported 3 tails.

Table A5: Balance Tests - Control vs Central district officers

	Mean Control	Mean Central	p-value
Panel A: Demographics			
District size at baseline	23	273	0.000***
Female	0.277	0.128	0.356
Age (1/5: 3=35-44 years)	3.014	2.880	0.462
Year graduated (baseline)	2003.454	2004.489	0.673
Officer rank (baseline)	3.589	3.346	0.542
Wage (baseline)	1939.582	1863.767	0.781
Intrinsic motivation to serve the community y/n (baseline)	0.631	0.617	0.807
Panel B: Outcome measures at baseline			
Education most effective against corruption y/n (baseline)	0.121	0.158	0.389
Comfortable reporting corruption 1/5 (baseline)	3.170	3.421	0.629
Know where to report corruption 1/5 (baseline)	4.312	4.165	0.626
Ever reported unethical behaviour y/n (baseline)	0.213	0.316	0.079*
Police corrupt 1/5 (baseline)	2.546	2.541	0.990
Witness unethical behaviour 1/4 (baseline)	2.723	2.850	0.647
Witness bribe 1/4 (baseline)	2.390	2.368	0.958
Police as Service Provider y/n (baseline)	0.333	0.406	0.239
Aggressive more useful than curteous 1/5 (baseline)	3.220	3.421	0.728
Overall workplace satisfaction 1/5 (baseline)	3.638	3.594	0.829
Panel C: Behavior in the baseline cheating game			
Mind Game: Number matched	0.624	0.549	0.233
Panel D: Outcome measures at baseline (restricted sample)			
How often monitor juniors 1/4 (baseline)	3.566	3.566	0.998
Caught juniors' unethical behaviour 1/4 (baseline)	2.453	2.618	0.379
Disciplined juniors 1/4 (baseline)	1.774	2.026	0.201
Observations (max)	141	133	

Notes: The table displays descriptive statistics for officers in the randomly selected Control districts, and in the Central district. P-values in column (3) are based on clustered standard errors (at police district level). Samples include officers surveyed at baseline and at endline.

Table A6: Robustness - Treatment effects, including Central district officers

	Values Index	Reporting Index	Monitoring Index	Perceptions Index	Relationship Index	Number of tails	N > 2 tails (y/n)	3 tails (y/n)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: ITT estimates								
Treatment District	0.126	0.073	0.150*	0.034	0.073	-0.184*	-0.122**	-0.075**
	(0.082)	(0.082)	(0.076)	(0.091)	(0.081)	(0.101)	(0.048)	(0.032)
	[0.303]	[0.640]	[0.134]	[0.640]	[0.640]	[0.069]	[0.030]	[0.037]
	{0.152}	{0.387}	{0.080}	{0.715}	{0.406}	{0.195}	{0.043}	{0.041}
Control Mean	-0.000	0.000	0.000	0.000	0.000	2.657	0.620	0.449
Observations	545	545	315	545	545	545	545	545
R-squared	0.050	0.066	0.063	0.126	0.059	0.020	0.026	0.010
LASSO Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Panel B: LATE estimates								
Treatment District	0.248	0.145	0.297*	0.066	0.142	-0.362*	-0.240***	-0.148**
	(0.151)	(0.158)	(0.156)	(0.178)	(0.154)	(0.195)	(0.088)	(0.059)
	[0.277]	[0.645]	[0.181]	[0.645]	[0.645]	[0.070]	[0.023]	[0.027]
	{0.133}	{0.383}	{0.091}	{0.720}	{0.401}	{0.188}	{0.031}	{0.029}
Control Mean	-0.000	0.000	0.000	0.000	0.000	2.657	0.620	0.449
Observations	545	545	315	545	545	545	545	545
R-squared	0.062	0.071	0.054	0.127	0.062	0.027	0.038	0.014
LASSO Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Panel C: Treatment effects								
Trained in T	0.278***	0.177**	0.171	0.078	0.164*	-0.306**	-0.204***	-0.123***
	(0.091)	(0.082)	(0.111)	(0.122)	(0.089)	(0.121)	(0.058)	(0.038)
	[0.043]	[0.175]	[0.447]	[0.945]	[0.284]	[0.036]	[0.008]	[0.009]
	{0.007}	{0.038}	{0.145}	{0.566}	{0.083}	{0.061}	{0.008}	{0.009}
Untrained in T	-0.031	-0.033	0.129	-0.013	-0.023	-0.058	-0.037	-0.025
	(0.089)	(0.121)	(0.104)	(0.132)	(0.088)	(0.112)	(0.047)	(0.040)
	[0.991]	[0.991]	[0.635]	[0.991]	[0.991]	[0.704]	[0.621]	[0.704]
	{0.739}	{0.798}	{0.229}	{0.926}	{0.822}	{0.650}	{0.441}	{0.548}
Control Mean	-0.000	0.000	0.000	0.000	0.000	2.657	0.620	0.449
Trained=Untrained (p-val)	0.002	0.106	0.781	0.600	0.014	0.042	0.003	0.050
Observations	545	545	315	545	545	545	545	545
R-squared	0.063	0.071	0.063	0.127	0.062	0.029	0.040	0.015
LASSO Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The table reports estimates obtained when including officers working in the Central district office. Panel A displays ITT estimates. Panel B displays LATE estimates where participation in the training is instrumented by treatment assignment. Panel C displays estimates obtained for trained and untrained officers in Treatment districts as opposed to officers in Control districts. Standard errors clustered at the police district level are reported in parentheses. Multiple hypothesis corrected p-values using the Romano-Wolf (2005) procedure are reported in square brackets. Due to small number of clusters (32), we report Wild-bootstrapped p-values in curly brackets. The controls are selected using the Double Lasso procedure (Belloni et al., 2014; Urminsky et al., 2016) from the full set of covariates displayed in Table 1.

Table A7: Robustness - Treatment effects, excluding top 10% improvement in survey scores

	Values Index	Reporting Index	Monitoring Index	Perceptions Index	Relationship Index	Number of tails	N > 2 tails (y/n)	3 tails (y/n)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: ITT estimates								
Treatment District	0.112	0.067	0.148	0.143	0.202**	-0.066	-0.113*	-0.074
	(0.097)	(0.094)	(0.116)	(0.113)	(0.082)	(0.122)	(0.063)	(0.048)
	[0.410]	[0.410]	[0.410]	[0.410]	[0.055]	[0.508]	[0.120]	[0.161]
	{0.278}	{0.497}	{0.211}	{0.250}	{0.031}	{0.622}	{0.110}	{0.160}
Control Mean	0.000	0.000	0.000	-0.000	-0.000	2.548	0.619	0.460
Observations	370	370	219	370	370	370	370	370
R-squared	0.075	0.096	0.088	0.158	0.060	0.012	0.029	0.028
LASSO Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Panel B: LATE estimates								
Treatment District	0.244	0.150	0.333	0.310	0.434**	-0.142	-0.245*	-0.161
	(0.198)	(0.206)	(0.259)	(0.257)	(0.171)	(0.257)	(0.131)	(0.100)
	[0.423]	[0.423]	[0.423]	[0.423]	[0.032]	[0.437]	[0.076]	[0.057]
	{0.255}	{0.499}	{0.213}	{0.274}	{0.026}	{0.620}	{0.099}	{0.150}
Control Mean	0.000	0.000	0.000	-0.000	-0.000	2.548	0.619	0.460
Observations	370	370	219	370	370	370	370	370
R-squared	0.096	0.098	0.069	0.148	0.042	0.022	0.051	0.038
LASSO Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Panel C: Treatment effects								
Trained in T	0.317***	0.116	0.141	0.123	0.218**	-0.223	-0.209***	-0.132***
	(0.102)	(0.096)	(0.151)	(0.141)	(0.102)	(0.140)	(0.069)	(0.047)
	[0.025]	[0.675]	[0.815]	[0.815]	[0.147]	[0.225]	[0.022]	[0.031]
	{0.013}	{0.266}	{0.357}	{0.458}	{0.052}	{0.143}	{0.012}	{0.016}
Untrained in T	-0.007	-0.032	0.139	0.082	0.128	0.048	-0.017	-0.001
	(0.097)	(0.130)	(0.149)	(0.125)	(0.084)	(0.127)	(0.058)	(0.055)
	[0.920]	[0.920]	[0.815]	[0.815]	[0.424]	[0.912]	[0.918]	[0.980]
	{0.944}	{0.813}	{0.375}	{0.529}	{0.142}	{0.728}	{0.783}	{0.986}
Control Mean	0.000	0.000	0.000	-0.000	-0.000	2.548	0.619	0.460
Trained=Untrained (p-val)	0.001	0.251	0.992	0.794	0.305	0.065	0.007	0.038
Observations	370	370	219	370	370	370	370	370
R-squared	0.092	0.096	0.082	0.155	0.060	0.025	0.053	0.039
LASSO Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The table reports estimates obtained when excluding officers who are in the top 10% of the distribution based on the extent to which their survey outcome measures improved over their baseline measures. Panel A displays ITT estimates. Panel B displays LATE estimates where participation in the training is instrumented by treatment assignment. Panel C displays estimates obtained for trained and untrained officers in Treatment districts as opposed to officers in Control districts. Standard errors clustered at the police district level are reported in parentheses. Multiple hypothesis corrected p-values using the Romano-Wolf (2005) procedure are reported in square brackets. Due to small number of clusters (32), we report Wild-bootstrapped p-values in curly brackets. The controls are selected using the Double Lasso procedure (Belloni et al., 2014; Urminsky et al., 2016) from the full set of covariates displayed in Table 1.

Table A8: Heterogeneous effects by initial intrinsic motivations

	Values Index	Reporting Index	Monitoring Index	Perceptions Index	Relationship Index	Number of tails	N > 2 tails (y/n)	3 tails (y/n)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ITT estimates								
Treatment District	0.005	0.183	-0.055	-0.113	0.496***	0.097	-0.078	-0.072
	(0.170)	(0.156)	(0.176)	(0.195)	(0.148)	(0.228)	(0.117)	(0.079)
	[0.968]	[0.678]	[0.891]	[0.867]	[0.054]	[0.886]	[0.752]	[0.669]
	{0.977}	{0.261}	{0.766}	{0.582}	{0.008}	{0.696}	{0.556}	{0.401}
Treatment District x	0.219	-0.123	0.336	0.296	-0.546**	-0.282	-0.047	-0.011
Intrinsic Motivations	(0.195)	(0.183)	(0.255)	(0.238)	(0.214)	(0.240)	(0.125)	(0.108)
	[0.696]	[0.867]	[0.598]	[0.632]	[0.116]	[0.508]	[0.886]	[0.886]
	{0.283}	{0.505}	{0.232}	{0.257}	{0.021}	{0.273}	{0.721}	{0.921}
Intrinsic Motivations when	-0.314**	0.211	-0.165	-0.147	0.379***	0.226	0.077	0.060
joined the police	(0.146)	(0.155)	(0.199)	(0.176)	(0.127)	(0.215)	(0.093)	(0.082)
	[0.201]	[0.573]	[0.867]	[0.867]	[0.073]	[0.586]	[0.715]	[0.747]
	{0.088}	{0.192}	{0.448}	{0.447}	{0.023}	{0.330}	{0.428}	{0.475}
Treat + Treat*IM=0	0.039**	0.541	0.077*	0.214	0.698	0.150	0.076*	0.164
Observations	412	412	248	412	412	412	412	412
LASSO Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The table displays OLS estimates where the dependent variables in columns 1 to 5 are the aggregated indexes generated by survey answers of police officers during the endline phone interviews. The indexes are calculated as weighted averages of the underlying variables, where the weights are inversely related to the degree of correlation between the underlying variables (see Anderson 2008). All indexes are standardized around their Control mean. Therefore, estimates are expressed in standard deviations from the control mean. The outcomes in columns 6 to 8 are the outcomes generated by the endline incentivized cheating game. "Intrinsic motivation" is an indicator for officers who responded "I wanted to serve the community" to the baseline survey question "What is the most important reason why you chose to become a police officer?". "Treatment District" is an indicator equal to 1 for officers in Treatment districts, and 0 for officers in Control districts. Standard errors clustered at the police district level are reported in parentheses. Multiple hypothesis corrected p-values using the Romano-Wolf (2005) procedure are reported in square brackets. Due to the small number of clusters (32), we report Wild-bootstrapped p-values in curly brackets. The controls are selected using the Double Lasso procedure (Belloni et al., 2014; Urminsky et al., 2016) from the full set of covariates displayed in Table 1.

Table A9: Heterogeneous effects by officer rank

	Values Index	Reporting Index	Monitoring Index	Perceptions Index	Relationship Index	Number of tails	N > 2 tails (y/n)	3 tails (y/n)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ITT estimates								
Treatment District	0.303**	-0.031	0.372**	-0.032	0.266**	-0.080	-0.143*	-0.058
	(0.127)	(0.158)	(0.181)	(0.217)	(0.103)	(0.221)	(0.082)	(0.069)
	[0.193]	[0.957]	[0.285]	[0.957]	[0.151]	[0.977]	[0.282]	[0.790]
	{0.059}	{0.847}	{0.036}	{0.892}	{0.024}	{0.736}	{0.101}	{0.421}
Treatment District x	-0.226	0.197	-0.312	0.144	-0.134	0.001	0.049	-0.034
High Rank	(0.182)	(0.189)	(0.207)	(0.270)	(0.203)	(0.221)	(0.072)	(0.102)
	[0.666]	[0.697]	[0.561]	[0.925]	[0.925]	[0.997]	[0.839]	[0.977]
	{0.249}	{0.319}	{0.171}	{0.623}	{0.537}	{0.997}	{0.504}	{0.753}
High Rank	0.225	-0.085	0.298	-0.146	0.363	-0.048	-0.065	0.024
	(0.171)	(0.137)	(0.253)	(0.217)	(0.220)	(0.202)	(0.055)	(0.091)
	[0.629]	[0.925]	[0.666]	[0.925]	[0.489]	[0.977]	[0.578]	[0.977]
	{0.237}	{0.565}	{0.274}	{0.557}	{0.130}	{0.823}	{0.301}	{0.806}
Treat + Treat*HR=0	0.534	0.115	0.655	0.451	0.349	0.530	0.194	0.134
Observations	412	412	248	412	412	412	412	412
LASSO Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The table displays OLS estimates where the dependent variables in columns 1 to 5 are the aggregated indexes generated by survey answers of police officers during the endline phone interviews. The indexes are calculated as weighted averages of the underlying variables, where the weights are inversely related to the degree of correlation between the underlying variables (see Anderson 2008). All indexes are standardized around their Control mean. Therefore, estimates are expressed in standard deviations from the control mean. The outcomes in columns 6 to 8 are the outcomes generated by the endline incentivized cheating game. "High rank" is an indicator for officers who, at baseline, had police ranks above the two bottom ranks of Constable and Lance Corporal. "Treatment District" is an indicator equal to 1 for officers in Treatment districts, and 0 for officers in Control districts. Standard errors clustered at the police district level are reported in parentheses. Multiple hypothesis corrected p-values using the Romano-Wolf (2005) procedure are reported in square brackets. Due to the small number of clusters (32), we report Wild-bootstrapped p-values in curly brackets. The controls are selected using the Double Lasso procedure (Belloni et al., 2014; Urminsky et al., 2016) from the full set of covariates displayed in Table 1.

Table A10: Heterogeneous effects by number of trained officers

	Values Index	Reporting Index	Monitoring Index	Perceptions Index	Relationship Index	Number of tails	N > 2 tails (y/n)	3 tails (y/n)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Trained in T	0.518***	0.134	-0.149	0.093	0.321**	-0.514**	-0.175	0.030
	(0.180)	(0.193)	(0.320)	(0.252)	(0.120)	(0.211)	(0.106)	(0.100)
	[0.106]	[0.919]	[0.971]	[0.984]	[0.142]	[0.099]	[0.293]	[0.968]
	{0.022}	{0.520}	{0.649}	{0.718}	{0.038}	{0.026}	{0.104}	{0.785}
N trained (-i)	0.011	0.009	-0.012	0.000	0.015**	-0.023**	-0.010*	0.001
	(0.013)	(0.013)	(0.011)	(0.017)	(0.006)	(0.010)	(0.005)	(0.005)
	[0.902]	[0.919]	[0.813]	[0.997]	[0.189]	[0.119]	[0.232]	[0.968]
	{0.461}	{0.538}	{0.363}	{0.980}	{0.013}	{0.086}	{0.159}	{0.882}
Trained in T x	-0.018*	0.002	0.015	-0.001	-0.015	0.023	0.001	-0.011*
N trained (-i)	(0.010)	(0.014)	(0.020)	(0.017)	(0.010)	(0.016)	(0.006)	(0.006)
	[0.434]	[0.997]	[0.916]	[0.997]	[0.560]	[0.391]	[0.968]	[0.232]
	{0.139}	{0.883}	{0.479}	{0.972}	{0.366}	{0.147}	{0.834}	{0.064}
Observations	271	271	163	271	271	271	271	271
LASSO Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The table displays OLS estimates where the dependent variables in columns 1 to 5 are the aggregated indexes generated by survey answers of police officers during the endline phone interviews. The indexes are calculated as weighted averages of the underlying variables, where the weights are inversely related to the degree of correlation between the underlying variables (see Anderson 2008). All indexes are standardized around their Control mean. Therefore, estimates are expressed in standard deviations from the control mean. The outcomes in columns 6 to 8 are the outcomes generated by the endline incentivized cheating game. "N trained (-i)" is the number of trained officers in a district, excluding officer i. The analysis is restricted to officers in Treatment districts. "Trained in T" is an indicator equal to 1 for trained officers in Treatment districts, and 0 for untrained officers in Treatment districts. Standard errors clustered at the police district level are reported in parentheses. Multiple hypothesis corrected p-values using the Romano-Wolf (2005) procedure are reported in square brackets. Due to the small number of clusters (32), we report Wild-bootstrapped p-values in curly brackets. The controls are selected using the Double Lasso procedure (Belloni et al., 2014; Urminsky et al., 2016) from the full set of covariates displayed in Table 1.

Table A11: Comparing Baseline Officer Characteristics in C and T Admin Data Districts

	Admin C	Admin T	p-value
District size at baseline	15.00	16.92	0.816
Female	0.261	0.279	0.816
Age (1/5: 3=35-44 years)	3.000	3.140	0.483
Year graduated (baseline)	2003.152	2002.202	0.639
Officer rank (baseline)	3.783	3.760	0.937
Wage (baseline)	2011.391	1991.124	0.869
Intrinsic motivation to serve the community y/n (baseline)	0.609	0.566	0.623
Panel B: Outcome measures at baseline			
Education most effective against corruption y/n (baseline)	0.087	0.124	0.555
Comfortable reporting corruption 1/5 (baseline)	3.370	3.248	0.750
Know where to report corruption 1/5 (baseline)	4.283	4.349	0.717
Ever reported unethical behaviour y/n (baseline)	0.283	0.357	0.378
Police corrupt 1/5 (baseline)	2.652	2.620	0.909
Witness unethical behaviour 1/4 (baseline)	2.804	2.884	0.680
Witness bribe 1/4 (baseline)	2.587	2.690	0.600
Police as Service Provider y/n (baseline)	0.391	0.264	0.123
Aggressive more useful than curteous 1/5 (baseline)	3.022	3.341	0.380
Overall workplace satisfaction 1/5 (baseline)	3.630	3.721	0.575
Panel C: Behavior in the baseline cheating game			
Mind Game: Number matched	0.609	0.612	0.965
Panel D: Outcome measures at baseline (restricted sample)			
How often monitor juniors 1/4 (baseline)	3.583	3.667	0.601
Caught juniors' unethical behaviour 1/4 (baseline)	2.458	2.762	0.225
Disciplined juniors 1/4 (baseline)	1.708	2.111	0.149

Notes: We obtained *clean* monthly district-level administrative data for 17 districts, 4 in the Control and 13 in the Treatment group, from September 2017 to December 2021. We also obtained data for the remaining districts, but such data are aggregated at a higher administrative level, which pulls together Control and Treatment districts in our sample, or adds districts that do not have traffic police units, and for which we do not have officer-level or other data at baseline or endline. This table compares district size (i.e., number of officers) and officer characteristics, as measured through our baseline survey, for the Control and Treatment districts for which we have clean administrative data. The variables in the table are the same as those in Table 1.

Table A12: Treatment Effects on the Probability of Promotion and Demotion

	Promoted		Demoted
	(1)	(2)	(3)
Panel A: ITT estimates			
Treatment District	0.014	-0.124**	0.023
	(0.040)	(0.047)	(0.035)
	{0.766}	{0.022}	{0.535}
Treatment District x High Rank		0.153**	
		(0.061)	
		{0.024}	
High Rank		-0.785***	
		(0.084)	
		{0.000}	
Treat + Treat*HR=0		0.476	
Control Mean	0.397	1.000	0.164
Observations	412	412	412
LASSO Controls	Yes	Yes	Yes
Panel B: LATE estimates			
Treated in T	0.033	-0.218**	0.016
	(0.081)	(0.087)	(0.060)
	{0.729}	{0.027}	{0.535}
Treated in T x High Rank		0.290**	
		(0.119)	
		{0.032}	
High Rank		-0.789***	
		(0.087)	
		{0.000}	
Treat + Treat*HR=0		0.385	
Observations	412	412	412
Control Mean	0.397	1.000	0.164
LASSO Controls	Yes	Yes	Yes

Notes: The table displays OLS ITT estimates where the dependent variable is a dummy equal to 1 if the officer was promoted (columns 1 and 2) or demoted (column 3) in the 20 months between baseline and endline surveys, and equal to 0 otherwise. Promotions and demotions are computed by comparing the self-reported ranks recorded at baseline and at endline. "Treatment District" is an indicator equal to 1 for officers in Treatment districts, and 0 for officers in Control districts. High rank is an indicator equal to 1 for officers whose rank is 3 or above. Standard errors clustered at the police district level are reported in parentheses. Due to the small number of clusters (32), we report Wild-bootstrapped p-values in curly brackets. The controls are selected using the Double Lasso procedure (Belloni et al., 2014; Urminsky et al., 2016) from the full set of covariates displayed in Table 1.

APPENDIX B: THE TRAINING PROGRAM

Day 1 of training: Intrinsic motivations, individual identity and values

The day started with an ice-breaking activity. Officers were asked to talk about their favourite toy that they played with during their childhood and the positive feelings that such experience created. Officers formed a circle and one-by-one they shared their story about their favourite toy, regardless of their ranks. The idea behind this activity was to bring everyone back to their initial sense of self and thus, the same starting point.

Each officer was then asked to explain what motivated them to become a police officer in the first place and what kind of police officer they wanted to be at the beginning of their career. The officers were divided into groups of five and each group discussed the qualities of the police officers they aspired to be when they first joined the Ghana Police Service. This was the first part of the re-building of their intrinsic motivations. Officers were then encouraged to discuss what the average police officer looks like today and where the gaps between the police officer they once wanted to be and the average police officer are. The second component of Part 1 of the training focused on the values that would help fill these gaps. Officers were reminded of the Ghana Police's Core Values, which are: 1) Commitment to personal and professional development of staff at all levels; 2) Treatment of all people – offenders, staff and the general public with dignity, respect and understanding; 3) Demonstration of professionalism through vigilance, fortitude, integrity, accountability and pride in work; 4) Encouraging positive interaction with the public to promote public safety and understanding; 5) Zero tolerance for crime and full commitment for human rights. Officers were encouraged to come up with additional values that they thought were important for the "ideal" police officer.

Next, officers were then asked to consider four work scenarios, and discuss, within their group, the best way to handle each situation using a particular value or set of values. For example, one of the four settings was the following: "You are on patrol and considering whether you have reasonable grounds to stop and search a particular vehicle which you suspect may be carrying illegal drugs. You step towards the car and the driver is an influential figure. He calls your supervisor and the supervisor asks you to let him go. What value is the most important?" Each group had to pick values that they thought were most important in that specific context, justify their reasons, and explain how they would use the value(s) to deal with the situation.

The final component of Part 1 focused on effective communication with citizens and public perceptions of the police. The primary message was that each interaction with a citizen should be treated as the most important interaction, because it only takes one *non-value-driven action* for public perceptions of the police *as a whole* to turn negative. The officers also learned about the principles of effective communication i.e., the 3 Vs – visual (i.e., body language), vocal (i.e., speech), and verbal (i.e., content) – and participated in a role-play exercise where some officers played in the role of citizens and others in the role of police officers. Each group took turns to enact the role play. A class discussion on communication strategies followed.

Day 2 of training: The development of a new group identity

By design, the second day of training took place one week after participation in the first day of training. This means that officers who participated on a given day in Week 1, returned on the same day in Week 2. The seven day break aimed to allow officers to internalize what was discussed during Day 1, and observe their every-day work realities, including interactions with colleagues and citizens, with new eyes.

Day 2 of training focused on building the new group identity of "Agent of Change" based on the shared values that the officers discussed and agreed to be important on Day 1. The session started with the facilitator, Inspector Schettini, showing a picture of a group of eagles – the symbol of the Ghana Police – and discussing how eagles are much stronger when they hunt together. The purpose was to introduce the idea of shared goals and enhanced strength through joint efforts, and to start creating a new group identity of "Agent of Change." In order to build the sense of belonging and solidify this new group identity, throughout the day, officers participated in several team-building activities. For instance, one of the activities required officers within each group of 5, to work together to create a drawing that represented a specific value that they had discussed during their first day of training. Another activity was a "sense exercise," where within each group of five officers, each person had to face some constraints in terms of their senses and needed others' help, e.g., some were blind-folded, or had one arm tied, or both arms tied, or were not allowed to speak, or were not able to hear, etc. The objective of this task was to get the group to find creative ways to communicate and work together, despite the physical challenges they faced. This was a metaphor for the challenges they would have to face if they wanted to change the culture of their organization.

An important component of Day 2 was to show evidence that change is possible. To this end, officers watched a 12 minute documentary video about Inspector Schettini's own journey at the Federal Highway Police in Brazil. The main purpose of the documentary was to show a story and narratives that the officers could identify with, i.e., the experience

of another police transport department, in which corruption was the norm, but where, by working together, officers managed to change the prevailing norms and permanently shift officers' mindsets. In the documentary, Inspector Schettini and his colleagues talked about how things were within their Department, e.g., how corruption was the norm, how change started when a new Head of the Department took office and gained the support of a group of like-minded officers who wanted to change, and how the new norms are maintained through rigorous training programs.[48] Participants watched the documentary as a group, and then expressed their thoughts and discussed how the measures that were discussed in the video could be applied in Ghana. This triggered a lively debate about what would be feasible and what other ideas could be implemented.

In order to turn these ideas into reality, the officers also learned about S.M.A.R.T (Specific, Measurable, Attainable, Relevant, and Time-bound) goal setting. The task was to create their own S.M.A.R.T plan for a change they wanted to make in their Unit or Department. Officers were asked to write down their goal, put it in a small envelope and keep it in their wallet as a reminder of what they were working towards with their plan.

The final task of Day 2 was for all the officers participating in a session (40 on average) to work together to create and perform a group dance inspired by the "Haka," which is a posture dance in Māori culture.[49] The idea behind having officers come up with a group dance and chant was to reinforce a sense of unity between the officers participating in the training. The dance was performed by the full group at the end of the day, and served to cement their sense of belonging to the new group

[48]The video is available here `https://www.bsg.ox.ac.uk/multimedia/video/proud-belong-brazilian-federal-highway-polices-fight-against-corruption`.

[49]The original Haka dance involves vigorous movements and stamping of the feet with rhythmically shouted or chanted accompaniment. It is often performed by the New Zealand Rugby team before a competition.

APPENDIX C: Deviations from the Pre-Analysis Plan

We pre-registered the study in the AEA Registry, with reference number AEARCTR-0003631, in September 2019. This Appendix outlines how the data analysis in the paper deviates from the Pre-Analysis Plan (PAP).

Timing of data collection

The endline data collection was initially scheduled for November 2019, seven months after the implementation of the ethics training. Due to administrative delays, primarily driven by the need for the police cooperation, the data collection was re-scheduled for April 2020, one year after the training. We were in the field training enumerators in March 2020, when the COVID pandemic forced us to postpone the data collection once again. We eventually decided to implement the data collection via phone, rather than in person, in December 2020, 20 months after the training. The change in the mode of survey implementation affected the lenght of the survey (which had to be shorter over the phone) as well as the behavioral game that we could use to measure willingness to engage in unethical behavior at endline. We elaborate further below.

Outcome Measures

Our pre-registration specified that our primary outcome variables would be: "1) unethical/corrupt behavior by members of the police force, including their own propensity to engage in such behavior; 2) their relationship with the citizens." We pre-specified that we would use survey measures and behavioral games to measure our outcomes of interest. With respect to the survey, we also stated that we would measure "sentiments towards the citizens, perceptions of corruption, tolerance for bending/breaking rules, propensity to notice and report unethical behavior of colleagues." While we did not pre-specify the aggregation of specific survey questions into indexes in the PAP, we stated that we would apply multiple hypothesis testing (MHT) corrections. For exposition clarity and to reduce the number of hypotheses tested (especially given the small sample size), we decided to group together the questions pertaining to 5 different indexes. In line with the PAP, they measure: 1) officers' values; 2) officers perceptions of corruption; 3) officers reporting of unethical behavior; 4) officers' monitoring of subordinates; 5) relationship with citizens. We still implement the MHT correction and report Romano-Wolf corrected p-values in the regression tables.

With respect to the game-generated outcomes, we deviated from the PAP by implementing a different cheating game at endline. We had originally planned to use the same "mind game" employed at baseline. However, this game required the rolling of a dice. Since the pandemic forced us to conduct a phone survey at endline, we opted for a cheating game that required the tossing of a coin. The assumption was that all respondents would be able to find a coin in their pockets or in their home, to be able to participate in the game over the phone.

At baseline, officers also played a second cheating die-rolling game (Weisel and Shalvi (2015)), which required sequential play between anonymous pairs of individuals. However, since members of a pair played different roles in the game (first mover versus second mover) the game was not well suited for between-subject comparisons of unethical behavior. Additionally, the inability to implement a die-rolling game at endline led us to exclude this game from the endline data collection.

In the PAP we pre-specified our intention to collect two sets of secondary outcomes: 1) measures of officers' aspirations and job satisfaction, and 2) measures of officer field behavior. Due to the need to shorten the survey instrument for phone implementation, we decided to refocus the endline survey to primarily measure our primary outcomes. As a result, we excluded questions on aspirations and retained only a single question on overall job satisfaction (while we had originally planned to measure satisfactions with different aspect of the job). We report the analysis of treatment effects in the Appendix.

In order to generate our second set of secondary outcomes (officers' field behavior), we had originally planned to collect survey data from tro-tro drivers, i.e., operators of minibuses commonly used for urban transportation. We initially planned to map common routes and classify them based on their location in control or treatment districts. Additionally, we intended to survey passengers and deploy decoy passengers on tro-tros to record police stops along these routes. This way, we aimed to record police stops along these routes to collect data on the frequency and duration of interactions between drivers and traffic officers, drivers' reports of payment exchanges, and perceptions of officers' attitudes. Due to the COVID-19 pandemic, we were unable to implement this data collection. Nevertheless, throughout the pandemic, we maintained our relationship with the police and worked toward securing administrative data on officers' behavior. Given the local context, the available data are limited. Ultimately, we obtained administrative records on the number of case violations reported by officers, cases referred to court, convictions, and the fines imposed on drivers. We use these data to provide some evidence of the intervention's impact on officers' field behavior. This analysis is not pre-registered, as we were unaware of the possibility of accessing these data when we first designed the project.

Estimation strategy

Our estimation strategy closely follows our pre-analysis plan (PAP). We pre-specified the estimation of Intent-to-Treat (ITT) effects (Equation 1 in the PAP) and Local Average Treatment Effects (LATE) (Equation 2 in the PAP) as our primary specifications. The resulting estimates are reported in Panels A (ITT) and B (LATE) of our primary regression tables.

In the PAP, we had also proposed estimating difference-in-difference regressions (equation 3), as we expected to have the same outcome variables at baseline and endline for both treatment and control groups. However, we deviated from this plan because modifications to our survey design at endline, necessitated by the need to shorten the survey, led us to include some of the questions from baseline and add new ones that we believed would more directly measure our primary outcomes of interest. As a result, the outcomes at baseline and endline are not identical. To address this, we always control for the baseline variables related to each outcome of interest in our empirical analysis, while selecting other controls using the Double Lasso procedure (more on this below).

In addition, we augmented our pre-specified analysis in several ways. First, we also report separate estimates for trained and untrained officers in treatment districts (Table 4). Assuming random assignment of officers to training within treatment districts, these estimates would represent the treatment effects on the treated. However, since we cannot empirically verify the randomization of officers into the training, we interpret these estimates with caution. Second, we implement the Double Lasso procedure (Belloni et al., 2014; Urminsky et al., 2016) to select control variables for inclusion in the empirical specifications. Third, we report Wild-bootstrapped p-values to correct for the small number of clusters. Finally, the PAP pre-specified heterogeneity analysis by officer age, gender, and rank. We report heterogeneity by rank but not by age, as age and rank are highly correlated, and the results are equivalent. We also conducted heterogeneity analysis by gender but do not report these results in the paper due to space constraints. These results are mentioned in footnote 41. Additionally, we conduct heterogeneity analysis by officers' initial intrinsic motivations and by the number of officers trained in a district (to capture potential spillovers). These additional analyses are not pre-registered but are important for studying the main mechanisms expected to drive the results, given the nature of the training.