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

March 21, 2024

Abstract

We examine the impact of ethics and integrity training on police officers in Ghana through a randomized field experiment. The program, informed by theoretical work on the role of identity and motivation in organizations, aimed to re-activate intrinsic motivations to serve the public, and to create a new shared identity of "Agent of Change." Data generated by a survey conducted 20 months later, show that the program positively affected officers' values and beliefs regarding on-the-job unethical behavior and improved their attitudes toward citizens. The training also lowered officers' propensity to behave unethically, as measured by an incentivized cheating game at endline. District-level administrative data for a subsample of districts is consistent with a significant impact of the program on field behavior in the short-run.

Keywords: Ethics training, police, experiment.

JEL classification codes: H76, K42, M53, D73

* 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 invaluable help with the design of the training modules. We are grateful to Chris Blattman, Alex Coutts, Jennifer Doleac, Andrew Foster, Emily Owens, Imran Rasul and Abu Siddique, and participants in the NBER Crime Summer Institute, the BREAD/IZA conference, and various seminars, for useful comments. We thank Daniel Gomez-Vasquez for excellent research assistance. We are forever 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.

†Authors' affiliations. Harris: Department of Economics, University of Oxford; donhatai.harris@economics.ox.ac.uk. Borcan: Department of Economics, University of East Anglia; o.borcan@uea.ac.uk. Serra: 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 chronic problem affecting many countries is poor performance of government organizations in delivering public services. A growing number of studies, recently reviewed by Finan et al. (2017),[1] have examined whether and how the recruitment, monitoring and compensation of public sector personnel could affect their performance.[2] Other studies have assessed the importance of management practices, including supervision, goal-setting and autonomy in decision-making (Bloom et al., 2020; Dunsch et al., 2017; Rasul and Rogger, 2018).

In this paper, we contribute to this literature by evaluating, through a randomized controlled trial, the impact of an ethics and integrity training program on values, attitudes and behaviors of police officers. The training centered around individual and group identity, and aimed to (re-)activate the officers' intrinsic motivations to *serve* the public, and to create a sense of belonging to a new social group that shared the mission of *bringing change* to the police.

The reputation of the police is often very poor, especially in developing countries. For example, the 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.[3] Among public organizations, the police is one of the sectors most difficult to regulate, monitor and incentivize.[4] This is for a number of reasons. First, the police is characterized by a more hierarchical structure than other public sectors. In settings where corruption and unlawful decision-making permeate all levels, including higher offices, this structure renders motivated lower ranked officers powerless to change prevailing organizational norms when those changes are not welcome by their superiors. Therefore, traditional mechanisms to motivate officers by aligning their behaviors with the original public service mission (e.g. monitoring, supervision, and financial rewards) often fail. Second, any abuse of power, e.g. any illegal or sub-par behaviors towards citizens, would need to be reported to the police itself, making it especially unlikely for citizens to come forward. Finally, officers have extreme discretionary power over the recipients of their services - being able, for instance, to imprison or physically harm them. For these reasons, alternative approaches that have proved successful in other public services are less likely to work in this setting. For instance, giving the police more autonomy is unlikely to work, unless the personal mission of police officers is also re-aligned with the original mission of the organization.

Given these challenges, we designed an ethics and integrity training program that is

[1] See also Ali et al. (2021), Avis et al. (2018) and Ashraf and Bandiera (2018).

[2] These include: Ashraf et al. (2020); Banerjee et al. (2015); Banuri and Keefer (2016); Borcan et al. (2017, 2014); Callen et al. (2020); Cassar and Armouti-Hansen (2020); Deserranno (2019); Dal Bó et al. (2013); Dhaliwal and Hanna (2017); Duflo et al. (2012); Ferraz and Finan (2008); Fisman et al. (2019); Hanna and Wang (2017); Khan et al. (2019); Linos (2018); Olken (2007).

[3] See `https://images.transparencycdn.org/images/2019_GCB_Africa3.pdf`

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

theoretically and empirically informed, specifically by the work on identity in economics (Akerlof and Kranton, 2005, 2008; Akerlof, 2016; Bénabou and Tirole, 2011; Chen and Li, 2009), studies of social identity and intra-intergroup dynamics in social psychology (Hogg et al., 2017; Costa Pinto et al., 2016; Tajfel, 1974; Travaglino et al., 2017), and analyses of intrinsic motivations, mission-matching and reputational concerns in economics (Andreoni and Bernheim, 2009; Bénabou and Tirole, 2006; Besley and Ghatak, 2005; Bursztyn and Jensen, 2017; Cassar and Armouti-Hansen, 2020; Prendergast, 2008). While some of this literature has highlighted the importance of matching the missions of workers and those of organizations at the job selection stage to enhance worker productivity, there is little evidence (Khan, 2020; Fracchia et al., 2021)[5] on how emphasizing the mission of an organization to current workers (rather than new potential workers) may impact their behaviors. Moreover, we add to this literature by designing and evaluating a program that aims to align the mission of existing workers to the *de jure* mission of the organization, while the *de facto* norms prevailing within the organization, including among superiors, are in contrast with it. This scenario is typical in highly corrupt government sectors.

We conducted 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).[6]

About half of the traffic police officers operating in 21 randomly selected police districts, out of 32, in the greater Accra region, received training between April and May 2019. The program, which we called "Proud to Belong," required participation in two full days of training delivered by an expert in police training, one day per week over two consecutive weeks and was implemented with the full cooperation of the Ghana's Motor Traffic and Transport Department (MTTD). The first day consisted of a workshop combining presentations and interactive activities including role-playing exercises. The aim of the first day was to get the officers to think about why they joined the police, and what they thought would be the characteristics of an ideal officer, which (at least some) participants may have aspired to be

[5]In particular, Khan (2020) assesses the effectiveness of a training program for community health workers in Pakistan, which emphasized the mission of the job and facilitated reflections and discussions centered around such mission.

[6]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.

when they first joined the police. Officers were then invited to jointly identify and discuss important values that are needed to be this ideal police officer. The training emphasized professionalism, honesty, respect and cooperative relationships with both colleagues and citizens. It also gave officers tools to better interact with citizens, and to challenge current behaviors and attitudes. The second day of training, held a week later, consisted of several team-building activities aimed at generating a new sense of unity between the trained officers, altering beliefs about other officers' willingness to join forces to change organizational norms, and creating a new shared identity of "Agent of Change." Post-training treatment reinforcements consisted of WhatsApp group chats and in an award ceremony that was held by the MTTD Police in December 2019, about eight months after the intervention. During the ceremony, the Inspector General of the Police awarded all trained officers with a certificate of completion, a formal letter praising and encouraging their efforts as Agents of Change, and a symbolic "Agent of Change" lapel pin.[7]

We conducted a baseline face to face survey of police officers operating in the districts under study in October 2018, about 7 months prior to the implementation of the training program. In December 2020, about 20 months post training and 12 months post award ceremony, we implemented the endline survey of police officers through phone interviews, due to the COVID-19 pandemic. Our outcome variables are measures of: 1) values and beliefs related to honesty, integrity, and the identity of a police officer; 2) likelihood of reporting the unethical behaviors of other officers; 3) likelihood of monitoring subordinates and punishing unethical behavior; 4) perceptions of corruption in one's police district; 5) relationship with citizens. Importantly, as part of our data collection efforts, we involved officers in an incentivized cheating game modeled after Abeler et al. (2014). 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 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 been used widely in recent years to measure preferences for (dis)honesty. Several studies have shown that behavior in this and similar incentivized games where subjects are asked to report the privately observed outcome of one or multiple dice rolls or coin flips, correlates well with dishonest or unethical behavior in the field, e.g., absenteeism of health workers in India (Hanna and Wang, 2017), illegal drug possession of inmates in prison (Cohn et al., 2015), high school students' misconduct (Cohn and

[7]This component of the program relates to interventions leveraging social recognition of public service providers to improve their performance, as in Ashraf et al. (2014) and in Fracchia et al. (2021). However, in our case, the symbolic award was not based on officers' absolute or relative performance or specific achievements. Instead, it was purely linked to the completion of the training, and served to reinforce the new collective identity created (Agents of Change), hence strengthening officers' commitment to honesty and integrity.

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

Maréchal, 2018), passengers' use of public transportation without a ticket (Dai et al., 2018), and university students' misappropriation of unearned income (Potters and Stoop, 2016).

Our data show that, nearly two years after participating in the "Proud to Belong" program, trained officers score significantly higher than untrained officers in our aggregate survey indexes capturing values, identity and beliefs, and attitudes toward citizens. We do not see significant differences in our survey-generated measures of officers' reporting of unethical behavior, their monitoring of subordinates, and their perceptions of corruption. Crucially, however, we find that the training program significantly lowered officers' likelihood to engage in unethical behavior, as measured by cheating in the incentivized game. As this measures actual behaviour (and not self-reported attitudes or perceptions), we take our results as evidence that the program was successful in activating and/or reinforcing values of honesty and integrity, in line with the *de jure* mission to serve the public. Heterogeneity analysis suggests that this may have operated through re-activating intrinsic motivation at the time of joining the police: the impact is driven by officers reporting high intrinsic motivation at the time of joining.

In order to examine whether changes in individual values and beliefs led to changes in field behavior, we obtained district-level data spanning 60 months (from January 2017 to December 2021) on number of filed cases, number sent to court, number of convictions, and fine amounts. These are important records, as traffic infractions in Ghana (should) result in court orders, requiring drivers to appear in court either to pay a fine or to await trial. However, due to data limitations, wherein values from multiple districts (including treatment and control districts, and adjacent areas) are often reported together, we are able to use only data for a subset of 17 of the original 32 districts. Despite the limited data, 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 impact of the training on officers' behavior, detectable within the first 3 months post-training. In particular, we see large increases in filed cases and court orders, leading to more convictions and fines. This aligns with the hypothesis that, in the absence of training, officers may have been substituting court orders with bribes.

Our study contributes to the growing literature on ways to improve the performance of public sector providers by focusing on different types of incentives, other than financial penalties or rewards, or increased monitoring. We are not aware of any previous evaluations of training programs for police officers or other civil servants centered specifically around ethics and integrity, with an emphasis on organizational mission, intrinsic motivations, and group identity. There are a few studies that have implemented training programs focusing on procedural justice and targeting police officers in the US (McLean et al., 2020; Owens et al., 2018; Wood et al., 2020), the UK (Miller et al., 2020; Wheller et al., 2013) and Mexico (Canales et al., 2019). These investigations have provided evidence of improved interactions between officers and citizens, for instance in the form of reduced use of force (Owens et al.,

2018; Wood et al., 2020) and increased likelihood of procedurally just behaviors (Canales et al., 2019). Similarly, a soft skill training program conducted with police officers in India, and centered around communication skills, improved the quality of interactions between officers and citizens, as perceived by citizens (Banerjee et al., 2021). Other studies have investigated the impact of community policing interventions aimed at increasing the frequency and quality of social interactions between police officers and community members, primarily in developing countries (Blair et al., 2021, 2019; Karim, 2020; Peyton et al., 2019), producing mixed results. A recent study conducted with the Chicago police (Dube et al., 2023) shows that a "Situational Decision-making" training, designed to mitigate impulse decision-making in cognitively complex situations, reduced officers' use of force and discretionary arrests.

The empirical evidence on the effectiveness of ethics training, more generally, is virtually non-existent.[9] 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 conclusion on the causal effect of ethics training on attitudes and behaviors.[10]

Overall, our findings suggests that carefully designed (interactive) face to face training programs, centered around intrinsic motivations and shared group identities, and aimed at shifting beliefs about others' willingness to jointly push for collective change, can significantly affect attitudes and behaviors, even in environments where corruption and abuses of power are pervasive. Our study has important implications for the design of anti-corruption interventions in developing countries,[11] as well as the design of programs aimed at incentivizing the ethical behaviors of police officers, and of employees both in governmental and private

[9]There is also very little empirical evidence on the causal effects of diversity training programs. Chang et al. (2019) conducted one of the few evaluations of an online diversity training program by implementing a field experiment involving employees of a large private firm. The study provides evidence of a positive impact on attitudes toward women, but no impact on behaviors. For a meta-analysis of research on diversity training, see Bezrukova et al. (2016).

[10]For instance, Beeri et al. (2013) assess the impact of ethical 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. Banerjee and Mitra (2018) study the effectiveness of ethics training on the ethical behavior of university students, as measured in an incentivized bribery game, and find no long-term impact. Similarly, Borcan et al. (2023) study the impact of an integrity training for Ukrainian law students on their propensity to facilitate bribes in an incentivised game, and find that training is only effective when students 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).

[11]Countering corruption in settings where corrupt behavior on the part of service providers is widespread is challenging. While increased monitoring and auditing have proven effective in these settings (e.g., Ferraz and Finan, 2008; Olken, 2007), the implementation of top-down accountability systems is often hindered by poor institutions, lack of transparency and collusion between different layers of the bureaucratic apparatus. In these environments, participatory accountability interventions, relying on the monitoring on the part of service recipients, also tend to fail (Banerjee et al., 2010; Olken, 2007). For a review of issues related to corruption in developing countries, see Banerjee et al. (2012), Olken and Pande (2012) and Serra and Wantchekon (2012).

organizations. More broadly, our findings provide empirical support to theoretical studies showing that individual and social identities play an important role in shaping values and attitudes.

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.[12] 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.[13]

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.[14]

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.[15] 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

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

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

[14] They document that tracks 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.

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

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.

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 typically 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. This framework introduces an incentive for bribery transactions, as officers may accept (solicited or unsolicited) bribes instead of issuing court orders.

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.[16]

3 The "Proud to Belong" training program

3.1 Design

When designing our training program we faced the crucial question: How do we shift individual attitudes and behaviors in the presence of sub-optimal organizational norms, when not following such norms comes with the risks of being punished by superiors, and of being socially ostracised by colleagues?

We primarily drew from the literature on identity and organizations (Akerlof and Kranton, 2005, 2008) and on group identity and collective behavior (Akerlof, 2016). Central to the former line of research is the alignment between employees' identity and the goals of their organisation. That is, the idea that when employees identify with the organisation's

[16]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).

identity, mission, and goals, they tend to work harder. Besley and Ghatak (2005), Cassar and Armouti-Hansen (2020) and Prendergast (2007) show similarly that production is enhanced when organizations hire workers who share their mission. However, aligning an employee's identity with that of the organization is problematic in environments where the *de facto* norms prevailing within an organization are in contrast with what may have been the original (or the *de jure*) mission of the organization. In these settings, it may be beneficial to re-activate the aspirational identity that employees may have had when first joining the organization, before being fully exposed to its sub-par norms, and to create a new shared identity among members of the organization. The latter strategy is in line with the idea that group identity could facilitate collective action through the activation of group pride, which in turns may lead individuals to switch from "I-thinking" to "We-thinking," as shown and discussed in Akerlof (2016).

We adopted this framework and designed our program with the aim of first (re-)activating officers' intrinsic motivations to serve the public with integrity and professionalism, and then creating a new group identity – that of "Agent of Change" – centered around the belief that change is possible through joint efforts.

When designing the training program, we consulted with Professor Alan Stein, a leading psychiatrist at the University of Oxford, and with Professor Michelle G. Craske, Professor of Psychology, Psychiatry and Biobehavioral Sciences, at the University of California, Los Angeles. We also drew from the vast experience of one of our co-authors – Bruno Schettini – who used to be a Police Inspector in Brazil and led a large number of training programs for the Bazilian Federal Highway Police. The training manual and modules were reviewed by Superintendent Samuel Sasu-Mensah (PhD) who is responsible for traffic law enforcement, training, and research at the Motor Traffic and Transport Department within the Ghana Police Service.

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. Details on the implementation of the program are provided in Section 3.2.

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

Part 1 of the training, delivered on Day 1, focused on: (i) rebuilding intrinsic motivations and the individual identity of a police officer; (ii) identifying and reflecting on the important values that define a "good police officer," and (iii) discussing how to effectively implement such values and communicate them to the public both verbally and through actions.

The day started with an ice-breaking activity.[17] Each officer was then asked to explain

[17]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

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[18] and were encouraged to come up with additional values that they thought were important for the "ideal" police officer. They 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.[19]

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.

3.1.2 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

back to their initial sense of self and thus, the same starting point.

[18]The Core Values 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.

[19]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.

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.[20]

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.[21] 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.[22] 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

[20]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.

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

[22]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.

cement their sense of belonging to the new group identity of "Agent of Change."

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.[23] 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.[24] Of the 32 MTTD districts in the sample, we randomly selected 21 to participate in the training program.[25]

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.[26]

[23]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.

[24]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.

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

[26]We computed these numbers based on the duty roasters for the month preceding the baseline data

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 in Treatment districts are not significantly different from each other (Table 1, columns (4)-(6)).[27]

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 329 officers in the Treatment districts and 165 in Control districts, for a total of 494 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 training - 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. Despite careful scrutiny of the lists of attendees, six officers from the Control districts, four officers from the Central district and three officers from non-MTTD districts were also present on at least one day of the training. Moreover, three small Treatment districts, which employed, respectively, a total of 2, 3 and 4 officers, were unable to send officers to the training. This was most likely due to their small sizes, which made it difficult to keep normal operations had one or more officers been absent to participate in the training.[28]

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,

collection.

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

[28]We discuss how non-compliers may affect the treatment evaluation, and how we deal with them in the empirical analysis, in Section 4.3.

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 83% (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.

We provide details on the baseline and endline surveys, as well as the incentivised cheating games, in Section 4 below. In Section 6, we also describe administrative data obtained from a subset of the districts, and explore the impact of the training on district-level field behavior.

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. 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.

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 Administration, and partly to generate baseline measures of officers' perceptions and experiences of corruption in the police, as well as their relationship with citizens, their willingness to report unethical behavior, and their beliefs about values associated with being a police officer and effective ways to fight corruption.

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.[29] Interviews took place in private rooms or outdoor space set up with marquee.

At the conclusion of the survey, officers participated in an incentivized cheating game – the mind game first introduced by Kajackaite and Gneezy (2017). 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. In our setting, after the four officers assigned to a given time slot had been privately surveyed, we had them sit in the same room, facing different sides of the room. We also used cardboard privacy screens to protect officers' decisions from each other's view. 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.[30]

During the session, officers were identified through a randomly assigned player number, and were never asked to state or write down their names. At the end of the experiment, they were paid in private through mobile credit sent directly to their phone on the same day.

We chose to include an incentivized cheating game in the baseline survey for a number of reasons. Previous research has shown 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). 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, despite the fact that nobody could observe the true outcome of a dice roll. The literature shows that individuals' preferences for being *seen* as honest are important, with subjects cheating less when the true outcome is observable to

[29]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.

[30]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, and thus, we did not include this game in the analysis.

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). In the mind game, subjects' concerns about being identified as a liar (when reporting a high number) are lower, and, as shown in Kajackaite and Gneezy (2017), subjects are more likely to cheat, especially when monetary incentives are large.

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 particular, we edited our baseline questionnaire to contain a restricted number of questions.[31]

Our primary objective was to measure the impact of the program on a set of outcomes that we hypothesized would be affected by the training. First and foremost, the training aimed to activate the identity of a police officer as a service provider, and the values of honesty and professionalism. These were central topics of the Day 1 of training. Other central issues were the ability to change organizational norms through joint efforts and the need to educate others about the harm generated by unethical behavior (Day 2). Another central module covered the relationship with citizens, including communication strategies. Specific activities and role-play exercises induced officers to talk about the reporting of unethical behavior, and the monitoring of other officers.

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. Moreover, we correct the p-values associated to individual hypotheses concerning to each treatment by employing the step-down multiple testing method developed by Romano and Wolf (2005), as further discussed in Section 4.2.

To give an example of the index creation process, let us consider our primary outcome variable, i.e., the values, identity and belief index. We have four survey-based proxies, generated by answers to the following questions: i) What are the most important qualities of a good police officer? (answers "being honest and professional"); ii) Do you see yourself more

[31]The baseline questionnaire, in its paper form, was 20 page long. The endline questionnaire was reduced to 7 pages only.

as a protector/crime preventer, a law enforcer or a service provider for citizens (answers "service provider"); iii) Do you think that organizational norms can change? (answers "yes"); iv) What do you think would be the most effective way of tackling corruption? (answers "Sensitize/educate the people on the evils of corruption"). We aggregate the answers to these four questions into an index by first demeaning each measure and dividing it by the standard deviation of the corresponding control group variable (so that all measures are on a comparable scale) and then taking their weighted average, as in Anderson (2008). We do the same for all sets of variables measuring the same outcome. All indexes are fully described in Table A1 in Appendix.

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.[32]

Our game-generated outcomes of interest are i) the reported number of tails, ii) the likelihood of reporting more than 2 tails, and iii) the likelihood of reporting exactly 3 tails. However, note that, 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. Our analysis will be primarily based on the comparisons of the empirical distributions of reported tails obtained for trained and untrained officers, and the corresponding theoretical distribution.

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

[32]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.

increasing the likelihood of cheating.

Another advantage of our methdology is that the endline survey took place 20 months after the training and 26 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: average treatment effect on the treated, intent-to-treat and local average treatment effect. As explained in Section 3.2, we assigned treatment status to 21 randomly selected police districts. 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. Given that, according to police reports, duty assignments follow a random rota system, we can consider participation in the training *within* treatment districts as essentially random (we discuss this further below). Thus, we proceed to estimate a model where we allow the treatment assignment to have different effects on trained and untrained officers in Treatment districts compared to officers in Control districts:

$$Y_{id} = \alpha + \beta_1 Trained\_T_{id} + \beta_2 Untrained\_T_{id} + \gamma X_{id} + \delta Z_d + \varepsilon_{id} \quad (1)$$

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. $Trained\_T_{id}$ 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} is a vector of officer characteristics measured at baseline: gender, age, wage and officer rank at baseline, officer graduation year, and intrinsic motivations to join the police officer to serve the community, as measured at baseline. X_{id} also includes outcome variables measured at baseline specific to each outcome: e.g, if the dependent variable is the values index, we include as control any baseline survey measures of values and beliefs; if the dependent variable is any of the outcomes in the incentivized coin flip game, we include as a control the behavior in the baseline mind game.[33] Z_d is a district-level control for district size, i.e., number of traffic police officers operating in the district. The standard errors are clustered at the police district level.

Our main estimates for the impact of the training program are based on equation (1). In particular, β_1 captures the average treatment effect on the treated (ATT). 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). Our partners within the Police repeatedly stated that participation in the training would be determined centrally on the basis of the duty rosters, which rely on a rota system. This was discussed early on, prior to the district-level randomization, when we were told that our training could not interfere with the pre-determined duties of the officers in the treatment districts. In fact, this led us to oversample treatment districts. Our baseline survey data are consistent with officers being assigned to posts through a random process.[34]

While we obtained numerous reassurances from the Police headquarters that the duty rota system was followed in assigning officers to the training, there remains a possibility that some officers were able to select in or out of the training. In Section 5.1, we test for significant differences between trained and untrained officers in treatment districts. We see little difference. In order to further account for possible selection into the training within treatment districts, we also estimate a more restrictive model, in which we compare our main outcomes of interest in districts randomly assigned to the treatment group and those in the control group:

$$Y_{id} = \alpha + \beta T_d + \gamma X_{id} + \delta Z_d + \varepsilon_{id} \tag{2}$$

where 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. Given that only about half of officers in treatment districts participated in the training, the coefficient β measures the

[33] We cannot compute exactly the same outcome indexes at baseline, as we added some survey questions that we use for such indexes at endline.

[34] At baseline, we asked officers how they were assigned their weekly duties. About 80 % answered that posts are assigned through a random system. The percentage increases to over 90% for the officers with the lower ranks. It is possible that higher rank officers have some control over which weekly duties they are assigned.

Intent-to-Treat (ITT) effect of the program.

We estimate also a third model, which captures 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.

$$\begin{aligned} Trained\_T_{id} &= \tau + \theta T_d + \gamma' X_{id} + \delta' Z_d + \eta_{id} \\ Y_{id} &= \alpha + \beta \hat{Trained\_T_{id}} + \gamma X_{id} + \delta Z_d + \varepsilon_{id} \end{aligned} \tag{3}$$

As per model (3), 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 will be very similar to the ATT estimates from equation (1). However, the LATE estimates are comparatively less efficient than OLS estimates, and result in smaller t-values.

We report estimates from equation (1) in the main text, and ITT and LATE estimates from models (2) and (3) in the appendix. In all specifications, we cluster the standard errors at the police district level. Given that 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).

One issue is how we should treat the non-compliers when estimating equation (1). 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. In section 5.2.3 we conduct robustness checks where we exclude the 6 trained Control officers from the sample.

Finally, as a secondary analysis, in order to identify the 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 seniority; iii) heterogeneity by number of officers trained in a treated districts. When estimating the heterogeneous treatment effects by intrinsic motivations and seniority, we use the full sample and we simply add to model (1) the interaction terms between the variable of interest and the indicators for trained and untrained in Treatment districts, as shown below for the intrinsic motivation

(IM) variable:

$$\begin{aligned} Y_{id} = \tau + \delta_0 Trained\_T_{id} + \delta_1 Untrained\_T_{id} + \delta_2 Trained\_T_{id} * IM_{id} \\ + \delta_3 Untrained\_T_{id} * IM_{id} + \delta_4 IM_{id} + \gamma_1 X_{id} + \gamma_2 \mathrm{Z}_d + \varepsilon_{id} \end{aligned} \tag{4}$$

As a result, δ_0 and δ_1 capture the effects of the program on trained and untrained officers in T who were not intrinsically motivated to serve the public when they first join the police; δ_2 and δ_3 represent the differential effects of the program on intrinsically motivated officers in T districts who were trained or not trained, respectively. The sums of δ_0 and δ_2, and of δ_1 and δ_3 capture the total effects of the program on intrinsically motivated trained and untrained officers in Treatment districts.

When estimating heterogeneous treatment effects by the number of trained officers, we restrict the sample to treatment districts, and estimate the following model:

$$\begin{aligned} Y_{id} = \tau + \delta_0 Trained\_T_{id} + \delta_1 N\_T_d + \delta_2 Trained\_T_{id} * N\_T_d \\ + \gamma_1 X_{id} + \gamma_2 \mathrm{Z}_d + \varepsilon_{id} \end{aligned} \tag{5}$$

where N_T_d is the number of trained officers in Treatment district d, excluding officer i. The coefficient δ_0 captures the marginal effect of training one officer in districts with no other officers trained. δ_2 captures the difference in training effects on officers in districts where more peers are trained, compared to those where no peers are trained. δ_1 represents the effect on untrained officers of having an additional officer in their district trained; therefore, from this coefficient we can infer whether there are any spillovers of the program from trained to untrained officers in treatment districts.

5 Results: Survey and Incentivized Cheating Game

5.1 Balance tests

To verify that the randomization resulted in balanced covariates across treatment and control districts, as well as across trained and untrained officers in treatment districts, in Table 1 we report an extensive set of baseline demographic characteristics, survey questions and decisions in the cheating game, and test for differences across the relevant groups. In Panel A, 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;[35] 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. The working sample consists of the 412 officers who were surveyed both at baseline

[35] A few survey questions that we use to generate our endline outcome indexes were not included 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. 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 (278 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 the unrestricted sample. In both groups, around 26-27% 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 controlling for the baseline measures of reporting and monitoring behaviors when assessing the impact of the training on the corresponding dependent variables.

To ensure that the β_1 coefficient in equation (1) is an unbiased estimate of the average treatment effect, it is also critical to have balanced pre-treatment covariates across trained and untrained officers in the Treatment districts. The descriptive statistics displayed in 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.[36] We only see two

[36]However, we cannot exclude the possibility that higher rank officials may have greater control over their schedule, and may have refused to participate in the training, if selected. We always include officer rank in our set of controls.

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 runs counter to the expected impact of the training. Specifically, if the higher likelihood of witnessing unethical behavior and the lower willingness to report such behavior are correlated with selection into the program, we would expect them to generate attenuation bias.

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. In all treatment groups - Control, *Untrained* in Treatment and *Trained* in Treatment - the percentage of officers who reported a match in the baseline mind game is about 60%, as shown also in Figure A2 in the appendix. This is significantly higher than the theoretical prediction (of 16.7%), and indicates that cheating among officers was pervasive at baseline, with no significant differences across treatment groups.

Overall, we conclude that the district-level randomization was successful. Moreover, the comparison of trained and untrained officers in Treatment districts gives us reasons to believe that participation of officers in the training program was in fact based on pre-determined, randomly generated, duty roasters. We account for the slight differences seen in our balance tests by including the full set of demographics and outcome-relevant baseline measures in our set of controls, as discussed in Section 4.2.

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.[37] All indexes are standardized around the control mean, hence they are displayed in standard deviations from such mean.

Figure 1 displays the average of each index for trained and untrained officers in Treatment districts. The figure shows that the trained officers score significantly higher in the Values index and the Reporting index, when compared to the officers in Control districts, whose means are set to zero, and also when compared to the untrained officers in treatment districts.

[37]Table A1 in Appendix displays the individual questions forming each index.

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.

The coefficients reported in Table 2, generated from estimating equation (1) of Section 4.2, confirm most of these findings. In the table, for each index, we report estimates without controls in odd columns, and estimates with controls in even columns. We report robust standard errors, clustered at the police district level, in parentheses, and p-values corrected for multiple hypotheses testing, using the Romano and Wolf (2005)'s procedure, in brackets (accounting for 8 hypotheses across Tables II and III).

Columns 1 and 2 of Table 2 show that the training program had a positive and significant impact on the Values index, which increased by around 0.33 standard deviations, as compared to the control mean. Columns 9 and 10 show that the training also positively impacted (by 0.24 standard deviations) the index measuring the relationship with citizens. We do not see any impact of the training on the untrained officers in Treatment districts, suggesting that the program did not spill over from trained to untrained officers.[38] We see no impact of the training on the indexes capturing reporting unethical behaviour and monitoring behavior of higher-rank officials toward subordinates (columns 3 - 6), and the index measuring perceptions of corruption in the district (columns 7 and 8).

Note that these estimates are conservative, since we treat the six officers based in Control districts who erroneously participated in the training as *untreated* in Control districts. We present robustness checks where we drop these six officers in Section 5.2.3.

We report the Intent-to-treat (ITT) estimates from equation (2) of Section 4.2 in Table A2 in the appendix. There, we test whether officers in treatment districts display higher values for the indexes of interest as compared to the officers in control districts. The estimates suggest that the program did have an impact on the Values index and the Relationship with citizens index, albeit the estimated coefficients are smaller in magnitude (i.e., about half the size) as compared to those in Table 2 and either marginally statistically significant or insignificant. This is to be expected, given that only about half of the officers in treatment districts were trained, and there were no spillover effects from trained to untrained officers, as shown in Table 2. Finally, the LATE estimates generated by equation (3) of Section 4.2, and reported in Table A4 in the appendix, are slightly larger in magnitude compared to the ATT estimates in Table 2 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 controls. However, the LATE estimates are less precise due to the loss of efficiency that comes from the two-stage estimation. Overall, we take the LATE estimates as a further indication that participation

[38]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.

in the training was not selective.

5.2.2 Impact on behavior in the incentivized cheating game

We start by assessing the distribution of reported coin tosses in the endline incentivized cheating game, modeled after Abeler et al. (2014), where officers had to toss a coin 4 times and report the number of obtained tails. Figure II 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 districts, and for the untrained and trained officers in Treatment districts, along with the theoretical binomial distribution $B(N = 4, p = 0.5)$.

The distributions generated by the officers in the Control districts and by the untrained officers in the Treatment districts are heavily left-skewed. Control officers under-report one and two tails, and far over-report three and four tails (e.g. around 47% report three tails, compared to the theoretical probability of 25%). The distribution for the untrained officers in Treatment districts is similar, displaying slightly less over-reporting for three tails, but still more frequent reports of three than two tails, and similar over-reporting of four tails. By contrast, the distribution of the number of tails reported by trained officers is closer to the theoretical distribution, with three tails reported less frequently than two tails, and with no over-reporting of four tails. Kolmogorov-Smirnov tests of equality between the three distributions indicate that the trained officers behave significantly differently than untrained and control officers (p-value=0.004).[39] Overall, the analysis of the distributions of reported coin tosses provides evidence of cheating among officers in Control districts and among untrained officers in Treatment districts, and suggests that trained officers were less likely to cheat than officers in the other two groups.

We confirm this result in Table 3, which presents the estimates from equation (1), 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 treatment effects for trained and untrained officers in Treatment districts, compared to the officers in the Control districts, without controls. Even columns report the estimates of treatment effects when including demographics and other individual characteristics measured at baseline, as well as behavior in the mind game at baseline.[40] We report robust standard errors, clustered at the police district level, in parentheses, and p-values corrected for multiple hypotheses testing, using the Romano and Wolf (2005)'s procedure, in brackets (accounting for 8 hypotheses across Tables 2 and 3).

[39]Binomial probability tests lead us to reject the null hypothesis that the observed probability of reporting three tails is equal to the theoretical probability of 25% for all officer groups; only for the treated officers, the probability of reporting four tails is insignificantly different from the theoretical probability of 6.5%.

[40]The sample size difference between even and odd columns is due to one missing value in the baseline mind game.

The estimates show that trained officers are significantly less likely (by about 18 percentage points) to report more than two tails, and significantly less likely (by 12.5 percentage points) to report exactly three tails.[41] The null and insignificant coefficients obtained for the untrained officers in Treatment districts indicate that the training program had no spillover effects from the trained to the untrained officers operating in the same district. P-values generated by testing for the equality of the coefficients of the "trained" and the "untrained" dummy variables confirm that the average treatment effect on the trained is significantly different from the average treatment effect on the untrained. Given that only around half of the officers in Treatment districts participated in the training, when estimating the Intent-To-Treat (ITT) effect of the program, i.e., equation (2) of Section 4.2, we find an impact that is about 60% of the magnitude of the Average Treatment on the Treated (which in our case is the same as the ATE) displayed in Table 3, and that is only marginally significant, as shown in Table A3 in the appendix.

Finally, our LATE estimates, i.e., equation (3) of Section 4.2, which are reported in Table A5 in the appendix, are strikingly similar to the ATT estimates displayed in Table 3 (and some are larger), i.e., they show that trained officers are about 22 percentage point less likely to report more than two tails. However, as before, the LATE estimates are less precise due to the loss of efficiency that comes from the two-stage estimation.

5.2.3 Robustness

While the LATE estimates of the training impact on officers' outcomes in the survey and cheating game are similar to the ATT estimates (and for some estimates slightly larger in magnitude), we conduct further robustness checks to understand if and how non-compliance may have affected our results. In this section, for some of the tests we also expand the sample size to include officers in the Central district. Finally, we conduct some checks to rule out the possibility that our results – and primarily those generated by survey questions - are driven by social desirability bias.

As a first robustness check, we exclude from the sample the six officers from Control districts who participated in the training. The estimates in appendix Tables A6 and A7 are very similar in magnitude to the main estimates in Tables 2 and 3, respectively. Moreover, the coefficients remain statistically significant in all specifications for which they were significant in the corresponding results tables.

Our next tests assess whether and how our estimated treatment effects change when including the Central district in the control group. Recall that we surveyed almost all (i.e., 253 of 273) officers based in the Central district at baseline, and about half of them, randomly selected, at endline. Our rationale for excluding the Central district from the main analysis was that Central districts officers are likely to move from district to district, filling in for ab-

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

sent officers, for instance, and therefore interacting with both trained and untrained officers. For this reason, we thought that, even if we did not allow them to participate in the training (and we did not), their potential interactions with trained officers would make them a less clean control group than the officers based in Control districts. However, our estimates show no evidence of spillover effects from trained to untrained officers, making our initial concerns less worrisome. Still, we have followed our pre-analysis plan and excluded the Central officers from our primary data 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 A10 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 results, displayed in Tables A11 and A12 in the appendix, are remarkably similar to those in Tables 2 and 3, but have higher levels of statistical significance. Overall, the inclusion of Central officers leaves the results unaffected, and, if anything, make them stronger (due to the larger sample size caused by the near-doubling of the control group). There is even some evidence that the training also significantly impacted the Reporting (unethical behaviour) index. This confirms, on the one hand, that there was no contamination from Treatment districts to Central and Control districts, and, on the other, that our program was highly effective (only) on the officers who participated in the training.

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.[42] 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

[42]These are questions Q3, Q4, Q5, Q6, Q12, Q13 and Q15 - displayed in Table A1 in the Appendix.

desirability.

We first check whether, due to experimenter demand effects, the "top improved officers" are more likely to be found among trained officers. This is not the case. 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 estimate our main model on the restricted sample, i.e., excluding the top improved officers. The results, displayed in Tables 4 and 5, show that, despite the sample reduction from 412 to 370 officers, our treatment effects are largely unaffected, which lends credence to the assumption that the identified impacts of the training program are unlikely to be the result of social desirability in survey answers and/or experimenter demand effects.

5.3 Secondary analysis: Mechanisms

In this section, we examine the channels that may have enabled the positive treatment effects that we have identified in our primary data analysis.[43] 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 fact, the central message of the integrity training (especially during the first day) was the idea of returning to the officers' motivations and aspirational identity. 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."[44] Based on this answer, we create a dummy variable for *initial intrinsic motivations*. As shown in Table 1, this variable is balanced among control and treatment officers, and among trained and untrained officers in treatment districts. In Tables A13 and A14 we interact our *Trained in T* and our *Untrained in T* dummies with our measure for initial intrinsic motivations, as per model (4) in section 4.2.

The estimates in Table A13 in the Appendix, columns 1 to 8, show that the values, reporting and monitoring behaviors, and the corruption perceptions of officers without initial intrinsic motivations were unaffected by the training. The interaction terms between training and intrinsic motivations are sizeable, albeit not precisely estimated. However, the t-tests for the overall treatment effects for intrinsically motivated officers produce p-values below the 10 percent significance level, suggesting that the training worked only for officers with a strong initial drive to serve the community. In particular, intrinsically motivated officers saw their values, beliefs and attitudes positively impacted by the training ($p = 0.011$), as

[43] We note that 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.

[44] 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.

well as their propensity to monitor subordinates ($p = 0.084$). The estimates in columns 9 and 10 show that in the model where the dependent variable is the relationship with citizen index, training did have an effect on not-intrinsically motivated officers, whereas the impact is null for the intrinsically motivated officers. This is likely due to the fact that intrinsically motivated officers already have better Relationships with Citizens, as indicated by the estimated coefficient of the uninteracted intrinsic motivations dummy. The estimates in Table A14, where the dependent variables are the outcomes in the incetivized game, show that the training effects are insignificant for officers lacking the original intrinsic motivation to serve the community, while the overall training effect on intrinsically motivated officers is significantly different from zero for all outcomes.

Overall, we take this as evidence that the training operated as intended, i.e., by re-activating intrinsic motivations and the identity of service provider. The null impact on the not-intrinsically motivated suggest that intrinsic motivations are difficult to generate if they were never originally there. On the other hand, the positive impact on the citizen relationship index suggests that training modules providing practical strategies on how to effectively and professionally communicate and interact with citizens, could be impactful also on officers who joined the police for reasons other than to serve their community.

Our second analysis of heterogeneous effects focuses on the officers' seniority in the police service. If intrinsic motivations and initial identity as a service provider are stronger when joining the police and are likely to weaken over time, we would expect the impact of the training to be stronger on the younger and less experienced officers. As a measure of seniority and experience, we take the officer's rank at baseline. High rank officers are Sergeants, Inspectors and above, while low-rank officers are Constable and Lance Corporal. Around two thirds of the officers in our sample are high rank.[45] Table A15 in the Appendix reveals that the training had large and significant impacts on the Values, Monitoring and Relationship indexes of low-rank officers. The magnitudes of the treatment effects for low rank officers on these three indices are larger than the baseline results in Table 2, and are between 0.4 and 0.5 SD above the control group means. The coefficients on the interactions between the training and the high-rank dummy mostly display a negative sign, albeit not statistically significant, suggesting that the training had a lower impact on higher-rank officials. In fact, only the Values Index of high-rank officials is significantly impacted by the training, albeit only at the 10 percent level ($p = 0.096$). We see similar patterns when examining behavior in the incentivized cheating game, in Table A16. Overall, the results suggest that the training proved effective mostly for low rank officers. One concern could be that trained low-rank officers, by being more likely to change their attitudes and behaviors in response to the program, may encounter resistance from (not as positively affected) higher rank officers,

[45] 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.

possibly leading to undesirable outcomes, such as demotions or less frequent promotions. This is not supported by our data. In fact, 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 (41%) and trained officers in Treatment districts (45%).[46]

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. To this end, we examine the number of officers trained within each treatment district and test whether the impact of the program on an officer increased with the number of trained colleagues (while controlling for the total number of officers in a district, as well as peers trained in that district; see model (5) in section 4.2). This analysis is restricted to Treatment districts only, and reported in Tables A17 and A18 in the Appendix. The estimates show that the identified impacts of the training on an officer's values and beliefs, his or her relationship with citizens, and his or her propensity to report high numbers in the cheating game, are largely independent from the number of trained officers in a district. In other words, the effects exist also for officers who were the only ones being trained in their district. Moreover, the sizes of the effects are not larger when the number of trained officers in a district increases. 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. The program also induced officers to separate their own identity, values and attitudes toward (un-)ethical behavior from their beliefs about the attitudes and behaviors of their colleagues. This may have contributed to the observed independence of treatment effects from number of trained colleagues in the same district.

The estimates in Tables A17 and A18 also show that, no matter the number of trained officers in a district, there were no spillovers to untrained officers in the same district. Put differently, in order for an officer to be affected, he or she had to have actively participated in the training.

6 Results: Administrative Data

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, whether the training has led to improved behavior in real-world scenarios, where officers operate within a social

[46]We also do not see any significant differences in rank demotions, which occurred for 19 percent of the officers across all groups.

context and may face influence from peers or superiors, is an open question, which we cannot answer with our survey and behavioral data. Ideally, we would use officer-level administrative data on field behavior, i.e., complaints filed against an officer, for instance for excessive use of force or for demanding bribes. In a context like Ghana, however, such data do not exist.

We were able to obtain partial 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 data are informative of unethical behavior because, when witnessing or discovering an infraction or a traffic offense, officers could substitute a formal case filing, and subsequent court order, with the demand (or acceptance) of a bribe. In other words, if on-the-road bribery is widespread, we anticipate a suboptimal number of cases being filed and forwarded to court by traffic police officers. Through an analysis of the available administrative data, we can evaluate whether our training program has led to an increase in the overall number of filed cases, court orders, convictions, and issued fines. We can interpret such changes as suggestive evidence of a decrease in bribery incidents in the field.

We obtained district-level administrative data for the 32 districts. However, the data for 15 of these districts is aggregated in a manner that hinders our ability to uphold our Treatment versus Control district comparison. 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.[47] The district-level data at our disposal are monthly, ranging from January 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.

In Table 6, 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 the districts with clean administrative data tend to be slightly smaller on average (22.60 versus 16.47 officers employed at baseline, $p = 0.163$), officer characteristics are remarkably similar across the two sets of districts, with no statistical significant difference identified in the data.[48]

We start by plotting quarterly district-level data, averaged by the number of officers in a

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

[48]Balance tests for Control and Treatment districts for which we have clean administrative data, displayed in Table A19 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 incentivized Mind Game.

district, by district treatment status, in Figure A3 in the Appendix. 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). 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, only 0.18 cases are filed monthly on average per officer, only 0.14 are sent to court and only 0.13 are convicted. The amount of Ghanaian Shilling collected in fines in a month by an officer on average is also low, at 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 significantly more per officer filed cases, as well as more per officer court orders, court convictions and total fines (in Ghanaian Shilling).

Next, we estimate the dynamic treatment effects of the training program, by conducting a difference-in-differences (DiD) event study,[49] 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 20219, 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:

$$\begin{aligned} \mathrm{Y}_{td} = \alpha + \beta_{27}(\text{Lag } 27)_{td} + ... + \beta_{2}(\text{Lag } 2)_{td} + \gamma_{0}(\text{Lead } 0)_{td} + ... \\ + \gamma_{32}(\text{Lead } 32)_{td} + \mu_{d} + \lambda_{t} + \varepsilon_{td} \end{aligned} \tag{6}$$

where Y_{td} refers 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

[49] We also estimate a standard difference-in-differences (DiD) model, in Panel A of Table A20, and a DiD over multiple time periods, in Panel B of Table A20. 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. 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. The DiD estimates over only two time periods (pre- and post-training), in Panel A, 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).

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 Shilling 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.[50]

The resulting event study monthly estimates for the four outcome variables, averaged by the number of officers in a district, are shown in Figure 3. 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). 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. The short-term impacts of the training on field behavior are large. Difference-in-differences estimates, shown in Table A20 in the Appendix, indicate that 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 Shelling, over a pre-treatment Control mean of 65 Shelling.[51]

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.[52] This resulted in more convictions and more money collected in fines. This is indicative of more effective law enforcement, hence less bribery, under the assumption that bribes are demanded or offered to avoid law enforcement. It is noteworthy that we are able to identify a positive impact of the training on field behavior despite the small sample, which only included half of all police districts

[50] As in our standard DiD specification, 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.

[51] The impact of the training on issued fines is sustained for up to 6 months post intervention, as shown in Panel B of table A20.

[52] This is in the three months post treatment were we saw the most activity in the WhatsApp group chats among the trained officers. It is possible that officers got transferred across districts later in the year, or the following year, which would compromise our identification strategy, and coul explain the lack of treatment effects in the long run. Our analysis of endline survey data (which involved about 80 percent of officers surveyed at baseline), however, indicate that only 5 percent of officers got transferred in the 20 months post training. Therefore, transfers are unlikely to be the reason for the lack of sustained treatment effects in the long run.

involved in the study. 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 with no evidence of spillover effects, it is remarkable that we can discern a positive impact from the training, albeit only in the short term.

7 Conclusions

In this paper, we provided empirical evidence on the effectiveness of an ethics training program that aimed to leverage intrinsic motivations, individual identity and a shared group identity with the aim of improving 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 training program, implemented over two weeks, in Spring 2019. The training modules were participatory and interactive, relying on a set of small group discussions, team-building and brainstorming activities and role-play scenarios. The overall core objectives were to first (re-)activate intrinsic motivations related to the identity of a police officer as "service provider," and to then generate a new group identity of "Agent 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 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.

While a full set of matched administrative data are not available, partial data from half of the districts in our data are suggestive of short-term large impacts on actual officer behaviour consistent with less predatory actions across the districts in which (some) officers were trained through our intervention. The fact that we observe a lasting impact on individual values 20 months post training, but only a short-term effect on field behavior may indicate that enduring shifts in individual values alone may not suffice to maintain sustained changes in field outcomes. Nevertheless, it is important to highlight that our examination of individual values focuses on treatment effects for those who participated in the training, while our assessment of field behavior constitutes an intent-to-treat analysis based on a limited sample

of districts.

Overall, we conclude that our ethics training program had a significant impact on police officers' values, beliefs and propensity to engage in unethical behavior, as well as their attitudes toward citizens, and seemingly also on actual behaviour. While the localized context of our study and the relatively small sample size make it desirable to replicate the program in other settings, we see our study as an important first step toward a better understanding of the role that carefully designed ethics training programs, centered around individual and shared identities, may play in shifting preferences and behaviors of police officers and, possibly, other service providers.

References

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.

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.

Andreoni, J. and B. D. Bernheim (2009). Social image and the 50–50 norm: A theoretical and experimental analysis of audience effects. *Econometrica 77*(5), 1607–1636.

Ashraf, N. and O. Bandiera (2018). Social incentives in organizations. *Annual Review of Economics 10*, 439–463.

Ashraf, N., O. Bandiera, E. Davenport, and S. S. Lee (2020). Losing prosociality in the quest for talent? sorting, selection, and productivity in the delivery of public services. *American Economic Review 110*(5), 1355–94.

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.

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.

Banerjee, R. and A. Mitra (2018). On monetary and non-monetary interventions to combat corruption. *Journal of economic behavior & organization 149*, 332–355.

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

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.

Bénabou, R. and J. Tirole (2006). Incentives and prosocial behavior. *American economic review 96*(5), 1652–1678.

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.

Bezrukova, K., C. S. Spell, J. L. Perry, and K. A. Jehn (2016). A meta-analytical integration of over 40 years of research on diversity training evaluation. *Psychological Bulletin 142*(11), 1227.

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.

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.

Canales, R., A. Charem Maus, M. Gonzales Magana, and J. F. Santini (2019). Shaping police officer mindsets and behaviors: Experimental evidence of procedural justice training.

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

Chang, E. H., K. L. Milkman, D. M. Gromet, R. W. Rebele, C. Massey, A. L. Duckworth, and A. M. Grant (2019). The mixed effects of online diversity training. *Proceedings of the National Academy of Sciences 116*(16), 7778–7783.

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.

Costa Pinto, D., W. M. Nique, M. Maurer Herter, and A. Borges (2016). Green consumers and their identities: how identities change the motivation for green consumption. *International Journal of Consumer Studies 40*(6), 742–753.

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 (2023). A cognitive view of policing. Technical report, National Bureau of Economic Research.

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., F. Schulz, and V. Vig (2019). Financial disclosure and political selection: Evidence from india (working paper).

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 (2021). Motivating volunteer health workers in an african capital city.

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.

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.

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.

McLean, K., S. E. Wolfe, J. Rojek, G. P. Alpert, and M. R. Smith (2020). Randomized controlled trial of social interaction police training. *Criminology & Public Policy 19*(3), 805–832.

Miller, J., P. Quinton, B. Alexandrou, and D. Packham (2020). Can police training reduce ethnic/racial disparities in stop and search? evidence from a multisite uk trial. *Criminology & Public Policy 19*(4), 1259–1287.

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.

Prendergast, C. (2007). The motivation and bias of bureaucrats. *American Economic Review 97*(1), 180–196.

Prendergast, C. (2008). Intrinsic motivation and incentives. *American Economic Review 98*(2), 201–05.

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.

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.

Travaglino, G. A., D. Abrams, and G. Russo (2017). Dual routes from social identity to collective opposition to criminal organisations: Intracultural appropriation theory and the roles of honour codes and social change beliefs. *Group Processes & Intergroup Relations 20*(3), 317–332.

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.

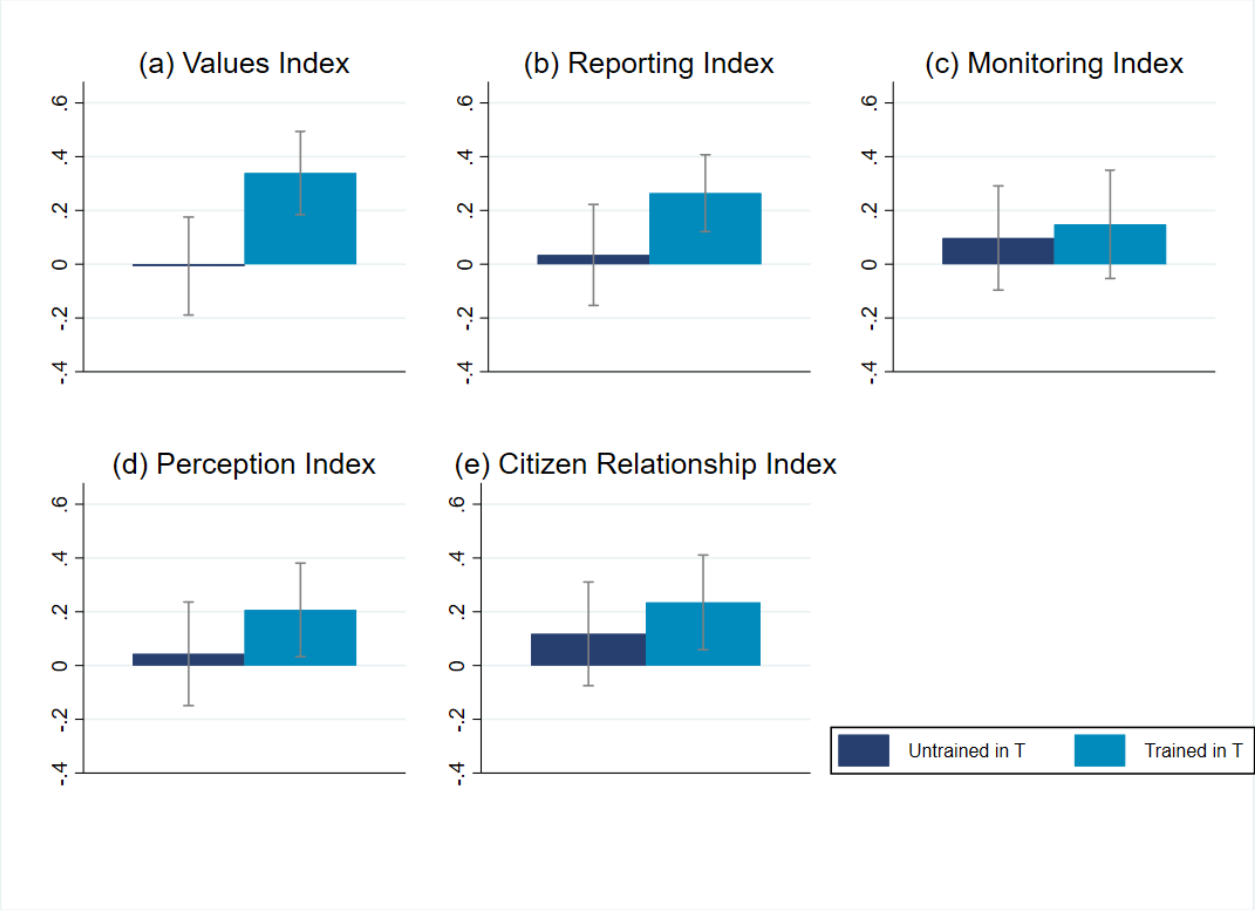
Weisel, O. and S. Shalvi (2015). The collaborative roots of corruption. *Proceedings of the National Academy of Sciences 112*(34), 10651–10656.

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

Wood, G., T. R. Tyler, and A. V. Papachristos (2020). Procedural justice training reduces police use of force and complaints against officers. *Proceedings of the National Academy of Sciences 117*(18), 9815–9821.

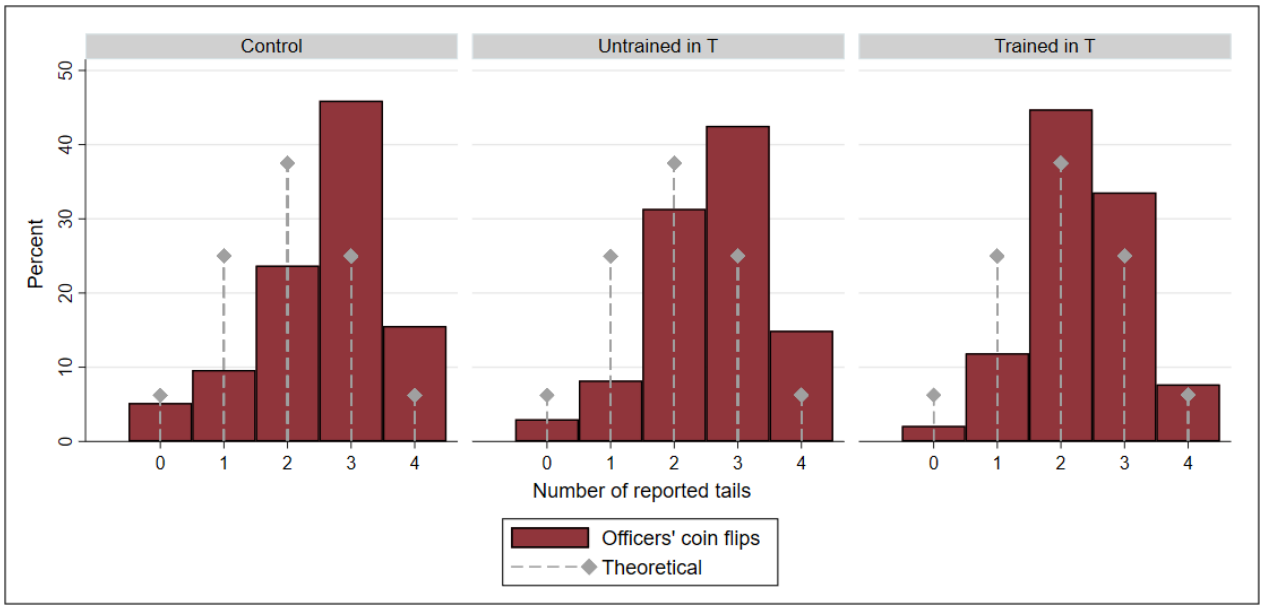
FIGURES AND TABLES

Figure 1: Survey-based indexes



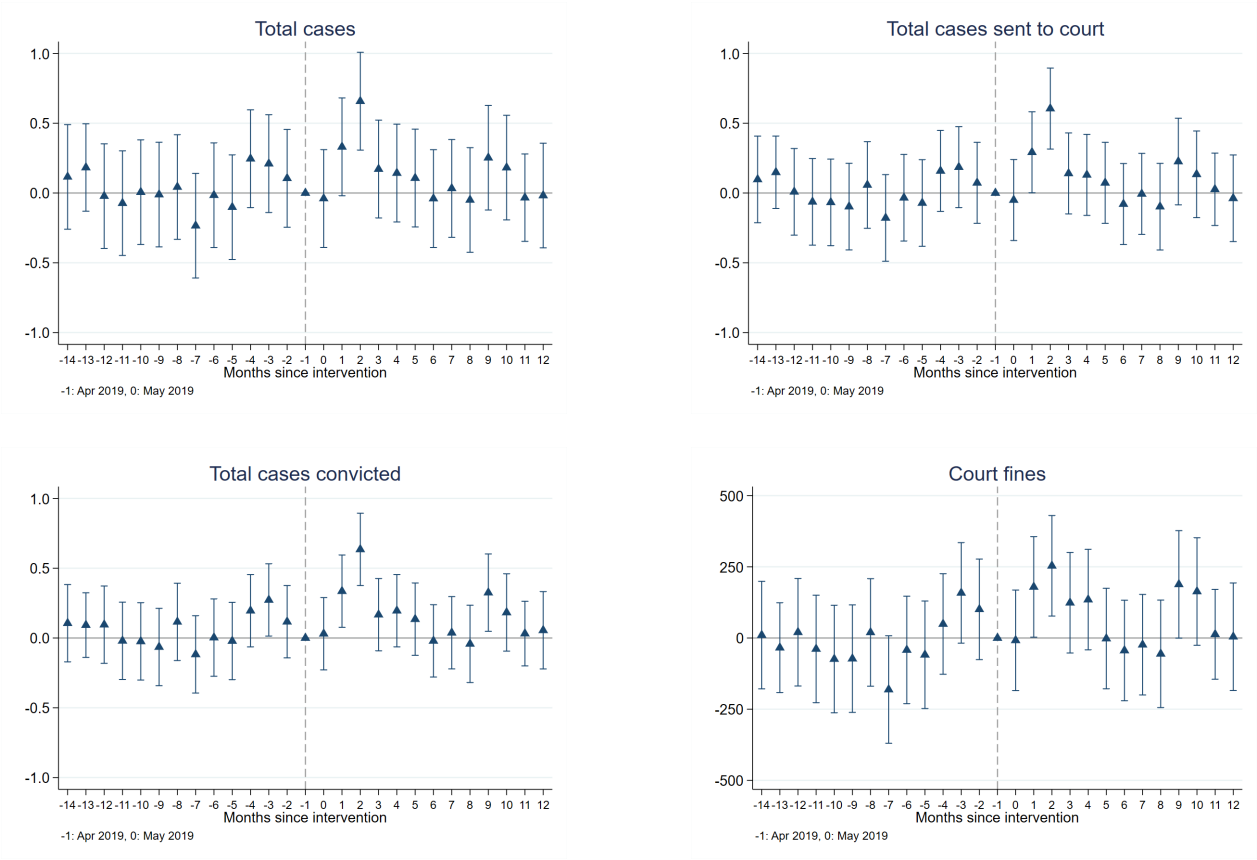
Notes: The figure shows the means and 95% confidence intervals of four survey-generated indexes, for trained and untrained officers in treatment police districts. The aggregated indexes are generated by survey answers of police officers during the endline phone interviews. The indexes are calculated 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 A1 in Appendix reports the individual questions constituting each index.

Figure 2: Officers' behavior in the incentivized cheating game



Notes: The figure shows the distribution of the number of tails reported by officers in the Control districts and by untrained and trained officers in Treatment (T) districts in the incentivized Coin Toss Game conducted at endline. The officers had to flip the coin 4 times, hence the number of reported tails ranges from 0 to 4. In gray, we display the theoretical binomial distribution of coin flips.

Figure 3: 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, for each of the 14 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. The figure on the bottom left displays the estimated impact of the training on the number of convicted cases on averaged per officer. The figure on the bottom right displays the estimated impact of the training on the fines issued to drivers on average per officer, in Ghanaian Shilling.

Table 1: Descriptive Statistics

	Mean C	Mean T	p-value (1-2)	Mean Untrained (T)	Mean Trained (T)	p-value (4-5)
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A - Demographics						
District number of officers at baseline	18.636	19.714	0.818	–	–	–
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	0.810	1936.280	1987.065	0.492
Intrinsic motivation to serve the community y/n (baseline)	0.631	0.594	0.470	0.598	0.590	0.887
Panel B - Outcome measures at baseline						
Education most effective against corruption y/n (baseline)	0.121	0.140	0.582	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	0.556	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	0.176	2.795	2.993	0.090*
Witness bribe 1/4 (baseline)	2.390	2.668	0.072*	2.576	2.755	0.171
Police as Service Provider y/n (baseline)	0.333	0.275	0.267	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	0.524	3.788	3.626	0.170
Panel C - Behavior in the baseline cheating game						
Mind Game: Matched number	0.621	0.609	0.805	0.606	0.612	0.927
Observations (max)	141	271		132	139	
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	0.123	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		64	61	

Notes: 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 they led 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)
Trained in T	0.339***	0.330***	0.264***	0.107	0.148	0.128	0.207	0.101	0.235**	0.243**
	(0.102)	(0.113)	(0.080)	(0.079)	(0.123)	(0.139)	(0.153)	(0.148)	(0.109)	(0.099)
	[0.030]	[0.063]	[0.030]	[0.717]	[0.731]	[0.934]	[0.634]	[0.972]	[0.184]	[0.117]
Untrained in T	-0.007	-0.001	0.034	-0.053	0.097	0.072	0.044	0.010	0.118	0.082
	(0.102)	(0.102)	(0.122)	(0.115)	(0.121)	(0.139)	(0.188)	(0.161)	(0.105)	(0.088)
	[0.992]	[0.997]	[0.992]	[0.990]	[0.908]	[0.990]	[0.992]	[0.997]	[0.769]	[0.934]
Mean Dependent Var. (Control)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
H_0: Trained=Untrained (p-val)	0.001	0.005	0.057	0.170	0.723	0.713	0.443	0.628	0.183	0.043
Observations	412	412	412	412	248	248	412	412	412	412
R-squared	0.026	0.084	0.014	0.111	0.004	0.062	0.007	0.141	0.008	0.060
Controls	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes

Notes: The table 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). All indexes are standardized around their Control mean. Therefore, estimates are expressed in standard deviations from the control mean. "Trained in T" is an indicator equal to 1 for trained officers in Treatment districts, and 0 otherwise. "Untrained in T" is an indicator equal to 1 for untrained officers in Treatment districts, and 0 otherwise. The excluded variable is an indicator equal to 1 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 (8 hypotheses across Tables 2 and 3). Controls include gender, age, wage and officer rank at baseline, officer graduation year, number of officers in the district, intrinsic motivation to serve the community when he/she joined the police, as measured at baseline, and, for each outcome, the corresponding variables measured at baseline.

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)
Trained in T	-0.215	-0.191	-0.200**	-0.183**	-0.130**	-0.125**
	(0.140)	(0.145)	(0.077)	(0.074)	(0.051)	(0.047)
	[0.521]	[0.734]	[0.098]	[0.117]	[0.104]	[0.095]
Untrained in T	0.030	0.039	-0.034	-0.024	-0.037	-0.025
	(0.132)	(0.128)	(0.067)	(0.066)	(0.049)	(0.050)
	[0.992]	[0.990]	[0.961]	[0.990]	[0.908]	[0.990]
Mean Dependent Var. (Control)	2.568	2.568	0.593	0.593	0.443	0.443
H_0: Trained=Untrained (p-value)	0.050	0.074	0.003	0.008	0.061	0.060
Observations	412	411	412	411	412	411
R-squared	0.013	0.032	0.031	0.054	0.013	0.029
Controls	No	Yes	No	Yes	No	Yes

Notes: The table 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. In columns 5 and 6, the dependent variable is a dummy equal to 1 if the officer reported 3 tails. "Trained in T" is an indicator equal to 1 for trained officers in Treatment districts, and 0 otherwise. "Untrained in T" is an indicator equal to 1 for untrained officers in Treatment districts, and 0 otherwise. The excluded variable is an indicator equal to 1 for officers in Control districts. Standard errors clustered at police district level are reported in parentheses. Multiple hypothesis corrected p-values using the Romano-Wolf (2005) procedure are reported in square brackets (8 hypotheses across Tables 2 and 3). Controls include gender, age, wage and officer rank at baseline, officer graduation year, number of officers in the district, intrinsic motivation to serve the community when he/she joined the police, as measured at baseline, and behavior in the baseline mind game.

Table 4: Robustness: Treatment effects on survey-generated outcomes, excluding the top 10% improvement in survey scores

	Values Index		Reporting Index		Monitoring Index		Perceptions Index		Relationship Index	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Trained in T	0.308***	0.304**	0.241**	0.081	0.135	0.117	0.293*	0.141	0.252**	0.253**
	(0.109)	(0.122)	(0.089)	(0.086)	(0.149)	(0.162)	(0.158)	(0.152)	(0.117)	(0.113)
	[0.131]	[0.131]	[0.916]	[0.916]	[0.969]	[0.969]	[0.916]	[0.916]	[0.192]	[0.192]
Untrained in T	-0.028	-0.029	0.021	-0.051	0.083	0.059	0.137	0.068	0.208**	0.162*
	(0.103)	(0.099)	(0.126)	(0.125)	(0.143)	(0.153)	(0.171)	(0.146)	(0.096)	(0.084)
	[0.994]	[0.994]	[0.994]	[0.994]	[0.994]	[0.994]	[0.994]	[0.994]	[0.314]	[0.314]
Mean Dependent Var. (Control)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
H_0: Trained=Untrained (p-val)	0.000	0.003	0.066	0.276	0.777	0.754	0.459	0.684	0.676	0.315
Observations	370	370	370	370	219	219	370	370	370	370
R-squared	0.024	0.096	0.012	0.110	0.003	0.062	0.012	0.180	0.011	0.046
Controls	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes

Notes: The table displays OLS estimates where the dependent variables are aggregated indexes generated by survey answers of police officers during the endline phone interviews. The sample excludes the officers who were above the 90th percentile in terms of the overall improvement in survey scores between baseline and endline. 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. "Trained in T" is an indicator equal to 1 for trained officers in Treatment districts, and 0 otherwise. "Untrained in T" is an indicator equal to 1 for untrained officers in Treatment districts, and 0 otherwise. The excluded variable is an indicator equal to 1 for officers in Control districts. Standard errors clustered at the police district level are reported in parentheses. Controls include gender, age, wage and officer rank at baseline, officer graduation year, number of officers in the district, intrinsic motivation to serve the community when he/she joined the police, as measured at baseline, and, for each outcome, the corresponding variables measured at baseline.

Table 5: Robustness: Treatment effects on game-generated outcomes, excluding the top 10% improvement in survey scores

	Coin flip:					
	Number of tails		N > 2 tails (y/n)		3 tails (y/n)	
	(1)	(2)	(3)	(4)	(5)	(6)
Trained in T	-0.214	-0.182	-0.221**	-0.198**	-0.143**	-0.132**
	(0.150)	(0.157)	(0.081)	(0.079)	(0.060)	(0.056)
	[0.818]	[0.818]	[0.127]	[0.127]	[0.173]	[0.173]
Untrained in T	0.039	0.055	-0.032	-0.015	-0.022	-0.002
	(0.133)	(0.132)	(0.063)	(0.067)	(0.059)	(0.062)
	[0.994]	[0.994]	[0.994]	[0.994]	[0.994]	[0.994]
Mean Dependent Var. (Control)	2.548	2.548	0.616	0.616	0.466	0.466
H_0: Trained=Untrained (p-value)	0.085	0.129	0.009	0.020	0.052	0.048
Observations	370	370	370	370	370	370
R-squared	0.013	0.031	0.038	0.060	0.016	0.039
Controls	No	Yes	No	Yes	No	Yes

Notes: The table displays OLS estimates where the dependent variables are the outcomes reported by police officers in the incentivized 4-time coin toss game. The sample excludes the officers who were above the 90th percentile in terms of the overall improvement in survey scores between baseline and endline. 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. In columns 5 and 6, the dependent variable is a dummy equal to 1 if the officer reported 3 tails. "Trained in T" is an indicator equal to 1 for trained officers in Treatment districts, and 0 otherwise. "Untrained in T" is an indicator equal to 1 for untrained officers in Treatment districts, and 0 otherwise. The excluded variable is an indicator equal to 1 for officers in Control districts. Standard errors clustered at police district level are reported in parentheses. Multiple hypothesis corrected p-values using the Romano-Wolf (2005) procedure are reported in square brackets (8 hypotheses across Tables 2 and 3). Controls include gender, age, wage and officer rank at baseline, officer graduation year, number of officers in the district, intrinsic motivation to serve the community when he/she joined the police, as measured at baseline, and behavior in the baseline mind game.

Table 6: Comparing Baseline Officer Characteristics in Districts with and without *Clean* Administrative Data

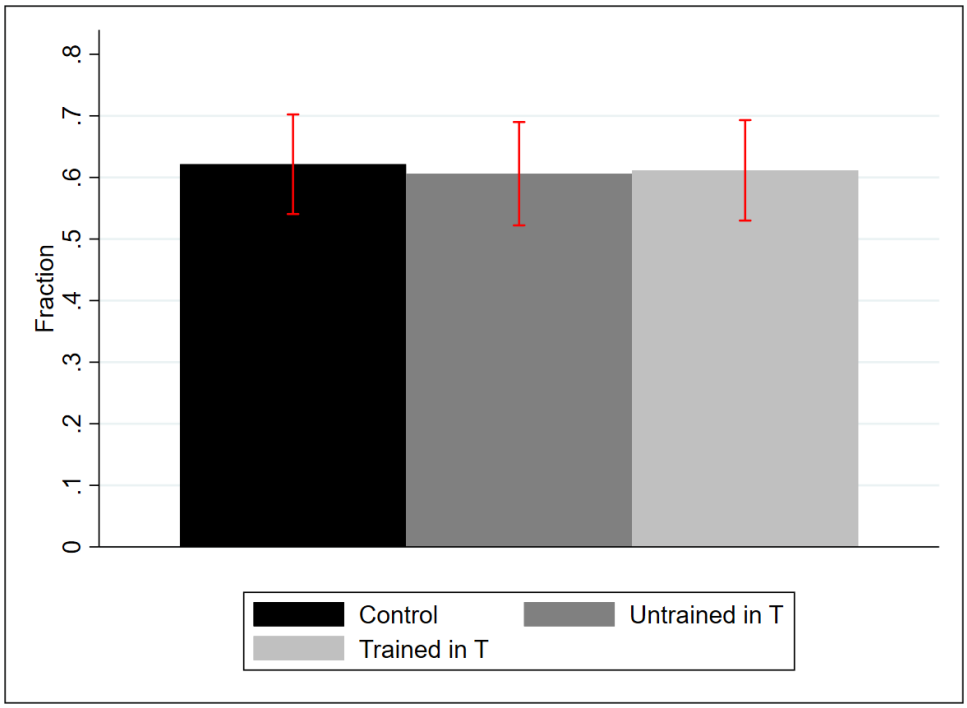
	No Admin Data	Admin Data	p-value
District size at baseline	22.60	16.47	0.163
Panel A - Demographics			
Female	0.215	0.240	0.520
Age (1/5: 3=35-44 years)	3.039	3.155	0.161
Year graduated (baseline)	2002.728	2001.885	0.280
Officer rank (baseline)	3.695	3.840	0.340
Wage (baseline)	1932.093	2016.750	0.143
Intrinsic motivation to serve the community y/n (baseline)	0.638	0.575	0.164
Panel B - Outcome Variables at Baseline			
Education most effective against corruption y/n (baseline)	0.154	0.115	0.221
Comfortable reporting corruption 1/5 (baseline)	3.240	3.305	0.628
Know where to report corruption 1/5 (baseline)	4.384	4.320	0.348
Ever reported unethical behaviour y/n (baseline)	0.294	0.320	0.542
Police corrupt 1/5 (baseline)	2.416	2.590	0.135
Witness unethical behaviour 1/4 (baseline)	2.781	2.835	0.549
Witness bribe 1/4 (baseline)	2.487	2.615	0.217
Police as Service Provider y/n (baseline)	0.290	0.310	0.643
Aggressive more useful than curteous 1/5 (baseline)	3.183	3.230	0.717
Overall workplace satisfaction 1/5 (baseline)	3.699	3.755	0.519
Mind Game: Number matched	0.604	0.630	0.570
How often monitor juniors 1/4 (baseline)	3.631	3.634	0.973
Caught juniors' unethical behaviour 1/4 (baseline)	2.532	2.594	0.642
Disciplined juniors 1/4 (baseline)	2.225	2.020	0.154

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.

APPENDIX FIGURES AND TABLES

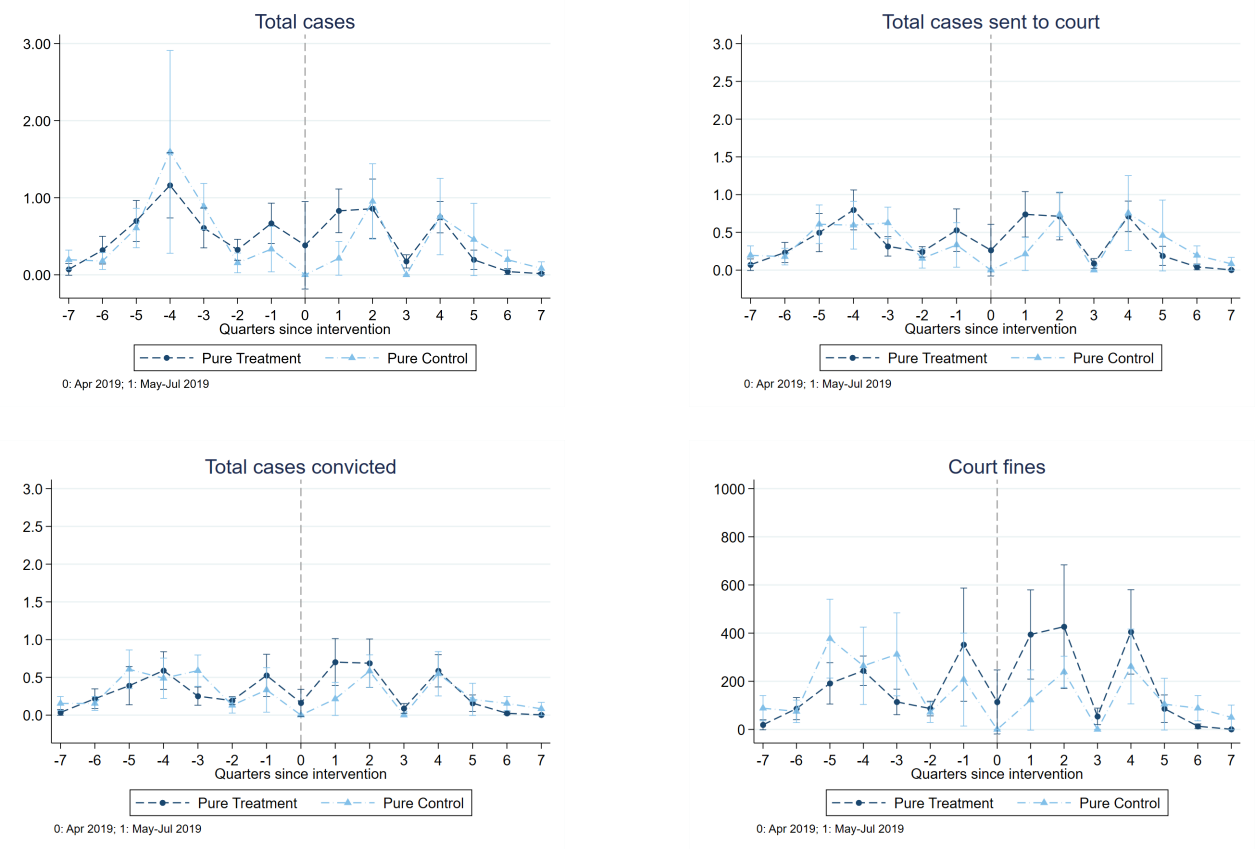
Figure A1: Agent of Change pin delivered to trained officers at the award ceremony

Figure A2: Behavior in the baseline Mind Game (Oct 2018)



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 percentage of trained and untrained officers in Treatment districts and of officers in Control districts who reported a match between the number they thought of and the number they rolled. The Mind Dice Game was conducted during the Baseline data collection, in October 2018, about 5 months prior to the training intervention.

Figure A3: District-level Administrative Data on Field Behavior



Notes: The figures report quarterly district-level administrative data, from January 2017 to December 2021, averaged by the number of officers in a district (at baseline, in October 2018). Clean data are available 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 C" and "Pure T" in the figure. "Total cases" in the top left figure refers to the quarterly numbers of cases filed in a district, averaged over the number of officers in the district. "Total cases sent to court" refers to the the quarterly numbers of cases sent to court in a district, averaged over the number of officers in the district. In our settings, traffic infractions lead to court orders, and drivers need to physically go to court to pay their fines. "Total cases convicted" refer to the quarterly numbers of cases that were convicted, averaged over the number of officers in the district. "Court fines" refers to the quarterly amount of fines, in Ghanaian Shilling, issued to drivers in a district, averaged over the number of officers in the district.

Table A1: Aggregate survey-based indexes

Index	Underlying survey questions
Values and Beliefs Index	Q1. What are the most important qualities of a police officers? Answered "being honest and professional;" Q2. Can organizational norms be changed? Answered "yes;" Q3. Do you see yourself more as a crime preventer, a law enforcer or a service provider for the citizen? Answered "service provider;" Q4. What do you think would be the most effective way of tackling corruption? Answered "Sensitize/educate the people on evils of corruption."
Reporting Index	Q5. I know where to report unethical behavior (5 point Likert agree/disagree scale); Q6. Have you ever reported unethical behavior? Answered "Yes"; Q7. How often do you talk about unethical behavior with colleagues? (4-point Likert scale)
Monitoring Index . (Officers who manage lower-rank officers)	Q8. How often do you monitor juniors? (4-point Likert scale); Q9. Over the past 24 months, have you caught officers behaving unethically?(4-point Likert scale); Q10. Over the past 24 months, have you disciplined/punished junior officers for behaving unethically? (4-point Likert scale).
Perceptions Index	Q11. 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?; Q12. There have been several reports that brand the police as corrupt. To what extent do you agree with that assessment? (5-point Likert scale); Q13. 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	Q14. In your opinion, how is the relationship between police officers and citizens this year compared to last year? (5-point Likert scale); Q15. In certain situations, it is more useful for an officer to be aggressive than to be courteous (5 point Likert scale).

Table A2: Intent-to-Treat 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)
Treated district	0.171*	0.169*	0.152*	0.028	0.123	0.101	0.127	0.056	0.178*	0.165*
	(0.096)	(0.096)	(0.085)	(0.080)	(0.099)	(0.116)	(0.132)	(0.121)	(0.098)	(0.086)
	[0.243]	[0.232]	[0.243]	[0.825]	[0.384]	[0.608]	[0.432]	[0.825]	[0.243]	[0.224]
Mean Dependent Var. (Control)	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
R-squared	0.006	0.067	0.005	0.107	0.004	0.061	0.003	0.140	0.006	0.056
Controls	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes

Notes: The table displays OLS estimates where the dependent variables are aggregate indexes based on survey answers of police officers during 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 coefficient of the "T district" variable is the intent-to-treat estimate. Standard errors clustered at police district level are reported in parentheses. Multiple hypothesis corrected p-values using the Romano-Wolf (2005) procedure are reported in square brackets (8 hypotheses). Controls include gender, age, wage and officer rank at baseline, officer graduation year, number of officers in the district, intrinsic motivation to serve the community when he/she joined the police, as measured at baseline, and, for each outcome, the corresponding variables measured at baseline.

Table A3: Intent-to-Treat effects on game-generated outcomes

	Coin flip:					
	Number of tails		N > 2 tails (y/n)		3 tails (y/n)	
	(1)	(2)	(3)	(4)	(5)	(6)
Treated district	-0.096	-0.079	-0.119*	-0.106	-0.085*	-0.076*
	(0.123)	(0.122)	(0.068)	(0.064)	(0.045)	(0.042)
	[0.432]	[0.818]	[0.243]	[0.255]	[0.225]	[0.232]
Mean Dependent Var. (Control)	2.568	2.568	0.593	0.593	0.443	0.443
Observations	412	411	412	411	412	411
R-squared	0.002	0.023	0.013	0.038	0.007	0.023
Controls	No	Yes	No	Yes	No	Yes

Notes: The table displays OLS estimates of model (1) where the dependent variable 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. In columns 5 and 6, the dependent variable is a dummy equal to 1 if the officer reported 3 tails. Standard errors clustered at police district level are reported in parentheses. Multiple hypothesis corrected p-values using the Romano-Wolf (2005) procedure are reported in square brackets (8 hypotheses). Controls include gender, age, wage and officer rank at baseline, officer graduation year, number of officers in the district, intrinsic motivation to serve the community when he/she joined the police, as measured at baseline, and behavior in the mind game at baseline.

Table A4: LATE (2SLS) on survey-generated outcomes

	Values Index		Reporting Index		Monitoring Index		Perceptions Index		Relationship Index	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Trained	0.371*	0.367*	0.315*	0.055	0.261	0.209	0.270	0.118	0.381*	0.351**
	(0.194)	(0.195)	(0.180)	(0.168)	(0.206)	(0.233)	(0.285)	(0.257)	(0.208)	(0.177)
	[0.195]	[0.211]	[0.195]	[0.836]	[0.371]	[0.605]	[0.448]	[0.836]	[0.195]	[0.186]
Mean Dependent Var. (Control)	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
R-squared	0.034	0.094	0.008	0.108		0.057	0.004	0.141		0.052
Controls	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes

Notes: The table displays IV regression estimates where the dependent variables are aggregate indexes based on survey answers of police officers during 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 instrument is the random assignment into treatment vs. control districts. Standard errors clustered at police district level are reported in parentheses. Multiple hypothesis corrected p-values using the Romano-Wolf (2005) procedure are reported in square brackets (8 hypotheses). Controls include gender, age, wage and officer rank at baseline, officer graduation year, number of officers in the district, intrinsic motivation to serve the community when he/she joined the police, as measured at baseline, and, for each outcome, the corresponding variables measured at baseline.

Table A5: LATE (2SLS) on game-generated outcomes

	Coin flip: number of tails		Coin flip: > 2 tails (y/n)		Coin flip: 3 tails (y/n)	
	(1)	(2)	(3)	(4)	(5)	(6)
Trained	-0.203	-0.168	-0.253*	-0.223*	-0.180**	-0.161*
	(0.256)	(0.251)	(0.139)	(0.132)	(0.090)	(0.085)
	[0.448]	[0.808]	0.195]	[0.246]	[0.185]	[0.211]
Mean Dependent Var. (Control)	2.568	2.568	0.593	0.593	0.443	0.443
Observations	412	411	412	411	412	411
R-squared	0.015	0.034	0.026	0.053	0.005	0.026
Controls	No	Yes	No	Yes	No	Yes

Notes: The table displays IV regression estimates for a model where the dependent variables are 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. In columns 5 and 6, the dependent variable is a dummy equal to 1 if the officer reported 3 tails. The instrument is the random assignment into treatment vs. control districts. Standard errors clustered at police district level are reported in parentheses. Multiple hypothesis corrected p-values using the Romano-Wolf (2005) procedure are reported in square brackets (8 hypotheses). Controls include gender, age, wage and officer rank at baseline, officer graduation year, number of officers in the district, intrinsic motivation to serve the community when he/she joined the police, as measured at baseline, and behavior in the mind game at baseline.

Table A6: Robustness: Treatment effects on survey-generated outcomes, excluding trained officers in C

	Values Index		Reporting Index		Monitoring Index		Perceptions Index		Relationship Index	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Trained	0.388***	0.389***	0.247***	0.093	0.148	0.126	0.205	0.094	0.230**	0.235**
	(0.115)	(0.124)	(0.077)	(0.077)	(0.124)	(0.141)	(0.153)	(0.150)	(0.109)	(0.099)
	[0.029]	[0.044]	0.036]	[0.805]	0.737]	[0.950]	0.652]	[0.988]	0.200]	[0.137]
Untrained in T	0.034	0.053	0.018	-0.069	0.098	0.069	0.040	-0.002	0.112	0.072
	(0.115)	(0.118)	(0.119)	(0.114)	(0.122)	(0.141)	(0.189)	(0.165)	(0.105)	(0.089)
	[0.992]	[0.988]	[0.992]	[0.988]	[0.909]	[0.988]	[0.992]	[0.988]	[0.797]	[0.968]
H_0: Trained=Untrained (p-value)	0.001	0.006	0.056	0.164	0.724	0.707	0.442	0.613	0.182	0.045
Observations	406	406	406	406	245	245	406	406	406	406
R-squared	0.030	0.096	0.013	0.107	0.004	0.062	0.007	0.141	0.008	0.058
Controls	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes

Notes: The table displays OLS estimates where the dependent variables are aggregate indexes based on survey answers of police officers during 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. In these regressions the 6 trained officers from Control districts are dropped. Standard errors clustered at police district level are reported in parentheses. Multiple hypothesis corrected p-values using the Romano-Wolf (2005) procedure are reported in square brackets (8 hypotheses). Controls include gender, age, wage and officer rank at baseline, officer graduation year, number of officers in the district, intrinsic motivation to serve the community when he/she joined the police, as measured at baseline, and, for each outcome, the corresponding variables measured at baseline.

Table A7: Robustness: Treatment effects on game-generated outcomes, excluding trained officers in C

	Number of tails		Coin flip: N > 2 tails (y/n)		3 tails (y/n)	
	(1)	(2)	(3)	(4)	(5)	(6)
Trained	-0.232*	-0.208	-0.205***	-0.189***	-0.128***	-0.124***
	(0.133)	(0.137)	(0.071)	(0.067)	(0.045)	(0.042)
	[0.401]	[0.598]	[0.064]	[0.068]	[0.066]	[0.054]
Untrained in T	0.013	0.018	-0.039	-0.030	-0.035	-0.024
	(0.125)	(0.119)	(0.059)	(0.059)	(0.043)	(0.045)
	[0.992]	[0.988]	[0.909]	[0.988]	[0.909]	[0.988]
H_0: Trained=Untrained (p-value)	0.050	0.077	0.003	0.008	0.061	0.060
Observations	406	405	406	405	406	405
R-squared	0.014	0.036	0.032	0.058	0.012	0.031
Controls	No	Yes	No	Yes	No	Yes

Notes: The table displays OLS estimates where the dependent variables are 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. In columns 5 and 6, the dependent variable is a dummy equal to 1 if the officer reported 3 tails. In these regressions, the 6 trained officers from Control districts are dropped. Standard errors clustered at police district level are reported in parentheses. Multiple hypothesis corrected p-values using the Romano-Wolf (2005) procedure are reported in square brackets (8 hypotheses). Controls include gender, age, wage and officer rank at baseline, officer graduation year, number of officers in the district, intrinsic motivation to serve the community when he/she joined the police, as measured at baseline, and behavior in the mind game at baseline.

Table A10: Balance Tests - Control vs Central district officers

	Mean Control	Mean Central district	p-value
Panel A: Demographics			
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	0.672
Officer rank (baseline)	3.589	3.346	0.542
Wage (baseline)	1939.582	1863.767	0.782
District employment (baseline)	23.000	273.000	0.000***
Intrinsic motivation (serve the community) y/n (baseline)	0.631	0.617	0.808
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 courteous 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: Matched number	0.621	0.549	0.250
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	0.207
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 found at baseline and at endline.

Table A11: Robustness - Treatment effects on survey-generated outcomes, including Central district officers

	Values Index		Reporting Index		Monitoring Index		Perceptions Index		Relationship Index	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Trained in T	0.136	0.300***	0.240***	0.184**	0.171	0.151	0.170	0.104	0.166*	0.239**
	(0.118)	(0.103)	(0.081)	(0.087)	(0.107)	(0.139)	(0.117)	(0.140)	(0.095)	(0.095)
	[0.780]	[0.058]	[0.106]	[0.225]	[0.582]	[0.864]	[0.609]	[0.959]	[0.501]	[0.117]
Untrained in T	-0.182	-0.013	-0.008	0.013	0.117	0.091	0.017	0.053	0.048	0.086
	(0.117)	(0.090)	(0.129)	(0.127)	(0.103)	(0.136)	(0.157)	(0.148)	(0.090)	(0.079)
	[0.589]	[0.998]	[0.985]	[0.998]	[0.780]	[0.960	[0.985]	[0.985]	[0.938]	[0.864]
H_0: Trained=Untrained (p-value)	0.001	0.003	0.058	0.190	0.711	0.699	0.450	0.779	0.188	0.052
Observations	545	545	545	545	315	315	545	545	545	545
R-squared	0.014	0.059	0.010	0.099	0.006	0.061	0.005	0.136	0.004	0.059
Controls	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes

Notes: The table displays OLS estimates where the dependent variables are aggregate indexes based on survey answers of police officers during 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 control group sample includes the Central district. Standard errors clustered at police district level are reported in parentheses. Multiple hypothesis corrected p-values using the Romano-Wolf (2005) procedure are reported in square brackets (8 hypotheses). Controls include gender, age, wage and officer rank at baseline, officer graduation year, number of officers in the district, intrinsic motivation to serve the community when he/she joined the police, as measured at baseline, and, for each outcome, the corresponding variables measured at baseline.

Table A12: Robustness - Treatment effects on game-generated outcomes, including Central district officers

	Coin flip:					
	Number of tails		N > 2 tails (y/n)		3 tails (y/n)	
	(1)	(2)	(3)	(4)	(5)	(6)
Trained in T	-0.319**	-0.223	-0.210***	-0.204**	-0.118***	-0.135**
	(0.121)	(0.141)	(0.061)	(0.076)	(0.042)	(0.050)
	[0.156]	[0.512]	[0.041]	[0.087]	[0.119]	[0.085]
Untrained in T	-0.074	0.003	-0.045	-0.045	-0.025	-0.033
	(0.112)	(0.126)	(0.046)	(0.063)	(0.039)	(0.050)
	[0.938]	[0.998]	[0.795]	[0.959]	[0.938]	[0.960]
H_0: Trained=Untrained (p-value)	0.049	0.075	0.003	0.007	0.060	0.048
Observations	545	544	545	544	545	544
R-squared	0.020	0.042	0.031	0.045	0.010	0.021
Controls	No	Yes	No	Yes	No	Yes

Notes: The table displays OLS estimates of model (1) where the dependent variables are 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. In columns 5 and 6, the dependent variable is a dummy equal to 1 if the officer reported 3 tails. The control group sample includes the Central district. Standard errors clustered at police district level are reported in parentheses. Multiple hypothesis corrected p-values using the Romano-Wolf (2005) procedure are reported in square brackets (8 hypotheses). Controls include gender, age, wage and officer rank at baseline, officer graduation year, number of officers in the district, intrinsic motivation to serve the community when he/she joined the police, as measured at baseline, and behavior in the baseline mind game.

Table A13: Heterogeneous effects by initial intrinsic motivations: Survey-generated outcomes

[H]

	Values Index		Reporting Index		Monitoring Index		Perceptions Index		Citizen Relationship Index	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Trained in T	0.330***	0.200	0.107	0.164	0.128	-0.139	0.101	-0.180	0.243**	0.554***
	(0.113)	(0.174)	(0.079)	(0.179)	(0.139)	(0.214)	(0.148)	(0.198)	(0.099)	(0.197)
	[0.075]	[0.922]	[0.818]	[0.976]	[0.976]	[0.985]	[0.998]	[0.976]	[0.149]	[0.181]
Untrained in T	-0.001	-0.148	-0.053	-0.038	0.072	-0.151	0.010	-0.124	0.082	0.450**
	(0.102)	(0.221)	(0.115)	(0.179)	(0.139)	(0.178)	(0.161)	(0.237)	(0.088)	(0.171)
	[1.000]	[0.985]	[1.000]	[0.999]	[1.000]	[0.976]	[1.000]	[0.992]	[0.976]	[0.210]
Trained in T x		0.207		-0.092		0.455		0.458*		-0.494*
Intrinsic Motivations		(0.233)		(0.249)		(0.278)		(0.269)		(0.285)
		[0.976]		[0.996]		[0.688]		[0.648]		[0.644]
Untrained in T x		0.235		-0.021		0.371		0.211		-0.589**
Intrinsic Motivations		(0.251)		(0.224)		(0.254)		(0.300)		(0.232)
		[0.971]		[1.000]		[0.785]		[0.985]		[0.237]
Intrinsic Motivations when	-0.123	-0.269*	0.161*	0.199	0.072	-0.196	0.067	-0.158	0.022	0.381***
joined the police	(0.098)	(0.142)	(0.093)	(0.164)	(0.126)	(0.193)	(0.124)	(0.180)	(0.147)	(0.136)
	[0.854]	[0.523]	[0.536]	[0.903]	[1.000]	[0.961]	[1.000]	[0.976]	[1.000]	[0.181]
H_0: Trained+										
Trained*Intrinsic Motivations=0		0.011		0.542		0.084		0.169		0.700
Observations	412	412	412	412	248	248	412	412	412	412
R-squared	0.084	0.086	0.111	0.111	0.062	0.072	0.141	0.148	0.060	0.074
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

.

Notes: The table 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). All indexes are standardized around their Control mean. Therefore, estimates are expressed in standard deviations from the control mean. "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?". "Trained in T" is an indicator equal to 1 for trained officers in Treatment districts, and 0 otherwise. "Untrained in T" is an indicator equal to 1 for untrained officers in Treatment districts, and 0 otherwise. The excluded variable is an indicator equal to 1 for officers in Control districts. Standard errors clustered at the police district level are reported in parentheses. Controls include gender, age, wage and officer rank at baseline, officer graduation year, number of officers in the district, and, for each outcome, the corresponding variables measured at baseline.

Table A14: Heterogeneous effects by initial intrinsic motivations: Game-generated outcomes

	Number of tails		Coin flip: N > 2 tails (y/n)		3 tails (y/n)	
	(1)	(2)	(3)	(4)	(5)	(6)
Trained in T	-0.191	0.007	-0.183**	-0.156	-0.125**	-0.118
	(0.145)	(0.253)	(0.074)	(0.120)	(0.047)	(0.081)
	[0.833]	[1.000]	[0.147]	[0.869]	[0.118]	[0.785]
Untrained in T	0.039	0.285	-0.024	0.062	-0.025	0.002
	(0.128)	(0.259)	(0.066)	(0.132)	(0.050)	(0.087)
	[1.000]	[0.940]	[1.000]	[0.992]	[1.000]	[1.000]
Trained in T x		-0.315		-0.042		-0.010
Intrinsic Motivations		(0.261)		(0.133)		(0.119)
		[0.903]		[0.996]		[1.000]
Untrained in T x		-0.396		-0.138		-0.043
Intrinsic Motivations		(0.289)		(0.150)		(0.117)
		[0.834]		[0.974]		[0.996]
Intrinsic Motivations when	0.030	0.265	0.042	0.101	0.051	0.068
joined the police	(0.118)	(0.250)	(0.066)	(0.110)	(0.055)	(0.087)
	[1.000]	[0.951]	[0.998]	[0.976]	[0.976]	[0.977]
H_0: Trained+						
Trained*Intrinsic Motivations=0		0.048		0.025		0.078
Observations	411	411	411	411	411	411
R-squared	0.032	0.040	0.054	0.057	0.029	0.030
Controls	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The table displays OLS estimates where the dependent variables are the outcomes reported by police officers in the incentivized 4-time coin toss game. The table 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. In columns 5 and 6, the dependent variable is a dummy equal to 1 if the officer reported 3 tails. "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?". "Trained in T" is an indicator equal to 1 for trained officers in Treatment districts, and 0 otherwise. "Untrained in T" is an indicator equal to 1 for untrained officers in Treatment districts, and 0 otherwise. Controls include gender, age, wage and officer rank at baseline, officer graduation year, number of officers in the district, and behavior in the mind game at baseline.

Table A15: Heterogeneous effects by officer rank: Survey-generated outcomes

	Values Index		Reporting Index		Monitoring Index		Perceptions Index		Citizen Relationship Index	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Trained in T	0.332***	0.499***	0.109	0.099	0.148	0.418**	0.102	0.131	0.252**	0.400***
	(0.114)	(0.162)	(0.079)	(0.165)	(0.140)	(0.190)	(0.149)	(0.253)	(0.101)	(0.122)
	[0.073]	[0.141]	[0.815]	[0.999]	[0.948]	[0.430]	[0.996]	[0.999]	[0.150]	[0.108]
Untrained in T	0.001	0.111	-0.052	-0.301	0.106	0.141	-0.002	-0.265	0.094	0.106
	(0.102)	(0.144)	(0.114)	(0.190)	(0.145)	(0.210)	(0.167)	(0.200)	(0.089)	(0.150)
	[1.000]	[0.995]	[0.998]	[0.768]	[0.996]	[0.999]	[1.000]	[0.876]	[0.948]	[0.998]
High Rank	0.032	0.169	0.016	-0.093	0.330***	0.481**	-0.102	-0.193	0.176	0.263
	(0.099)	(0.185)	(0.082)	(0.141)	(0.119)	(0.205)	(0.145)	(0.227)	(0.135)	(0.169)
	[0.999]	[0.984]	[0.999]	[0.999]	[0.089]	[0.355]	[0.996]	[0.987]	[0.847]	[0.768]
Trained x High Rank		-0.250		0.001		-0.388*		-0.062		-0.227
		(0.211)		(0.208)		(0.214)		(0.311)		(0.209)
		[0.923]		[1.000]		[0.641]		[1.000]		[0.947]
Untrained x High Rank		-0.157		0.365*		-0.036		0.384		-0.015
		(0.230)		(0.196)		(0.239)		(0.266)		(0.251)
		[0.999]		[0.603]		[1.000]		[0.826]		[1.000]
H_0: Trained + Trained*High Rank=0		0.096		0.320		0.856		0.708		0.252
Observations	412	412	412	412	248	248	412	412	412	412
R-squared	0.084	0.086	0.111	0.117	0.055	0.062	0.136	0.144	0.063	0.065
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

.

Notes: The table 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). All indexes are standardized around their Control mean. Therefore, estimates are expressed in standard deviations from the control mean. "Trained in T" is an indicator equal to 1 for trained officers in Treatment districts, and 0 otherwise. "Untrained in T" is an indicator equal to 1 for untrained officers in Treatment districts, and 0 otherwise. "High rank" is an indicator for officers who, at baseline, had police ranks above the two bottom ranks of Constable and Lance Corporal. 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 (8 hypotheses). Controls include gender, age, wage and officer graduation year, number of officers in the district, intrinsic motivation to serve the community intrinsic motivation to serve the community when he/she joined the police, as measured at baseline, and, for each outcome, the corresponding variables measured at baseline.

Table A16: Heterogeneous effects by officer rank: Game-generated outcomes

	Coin flip:					
	Number of tails		N > 2 tails (y/n)		3 tails (y/n)	
	(1)	(2)	(3)	(4)	(5)	(6)
Trained in T	-0.194	-0.258	-0.184**	-0.244**	-0.124**	-0.106
	(0.147)	(0.265)	(0.074)	(0.103)	(0.046)	(0.091)
	[0.847]	[0.979]	[0.150]	[0.354]	[0.104]	[0.931]
Untrained in T	0.034	0.165	-0.026	0.008	-0.022	0.032
	(0.132)	(0.255)	(0.066)	(0.099)	(0.050)	(0.074)
	[0.999]	[0.999]	[0.998]	[1.000]	[0.998]	[1.000]
High Rank	-0.061	-0.041	-0.028	-0.046	0.031	0.064
	(0.098)	(0.221)	(0.044)	(0.070)	(0.057)	(0.100)
	[0.997]	[1.000]	[0.996]	[0.999]	[0.997]	[0.999]
Trained x High Rank		0.107		0.093		-0.025
		(0.260)		(0.102)		(0.124)
		[1.000]		[0.984]		[1.000]
Untrained x High Rank		-0.194		-0.051		-0.078
		(0.263)		(0.107)		(0.110)
		[0.998]		[0.999]		[0.998]
H_0: Trained + Trained*High Rank=0		0.297		0.069		0.050
Observations	411	411	411	411	411	411
R-squared	0.033	0.037	0.054	0.057	0.029	0.030
Controls	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The table 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. In columns 5 and 6, the dependent variable is a dummy equal to 1 if the officer reported 3 tails. "Trained in T" is an indicator equal to 1 for trained officers in Treatment districts, and 0 otherwise. "Untrained in T" is an indicator equal to 1 for untrained officers in Treatment districts, and 0 otherwise. "High rank" is an indicator for officers who, at baseline, had police ranks above the two bottom ranks of Constable and Lance Corporal. 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 (8 hypotheses). Controls include gender, age, wage and officer graduation year, number of officers in the district, intrinsic motivation to serve the community when he/she joined the police, as measured at baseline, and behavior in the mind game at baseline.

Table A17: Heterogeneous effects by number of trained officers: Survey-generated outcomes (treatment districts only)

	Values Index			Reporting Index			Monitoring Index			Perceptions Index			Citizen Relationship Index		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
Trained	0.303**	0.303**	0.512***	0.178*	0.184*	0.144	0.040	0.079	-0.237	0.083	0.146	0.184	0.137*	0.131	0.291**
	(0.108)	(0.118)	(0.178)	(0.099)	(0.103)	(0.194)	(0.174)	(0.185)	(0.279)	(0.162)	(0.168)	(0.248)	(0.079)	(0.086)	(0.123)
	[0.048]	[0.190]	[0.146]	[0.201]	[0.462]	[0.966]	[0.788]	[0.986]	[0.959]	[0.788]	[0.897]	[0.966]	[0.201]	[0.576]	[0.295]
N trained (-i)		-0.000	0.009		-0.004	-0.006		-0.026	-0.041**		-0.037**	-0.035*		0.004	0.010
		(0.013)	(0.013)		(0.008)	(0.012)		(0.015)	(0.019)		(0.016)	(0.019)		(0.014)	(0.012)
		[0.986]	[0.977]		[0.986]	[0.994]		[0.474]	[0.392]		[0.238]	[0.535]		[0.986]	[0.959]
Trained x N trained (-i)			-0.017*			0.003			0.028			-0.003			-0.013
			(0.010)			(0.013)			(0.019)			(0.015)			(0.008)
			[0.562]			[1.000]			[0.710]			[1.000]			[0.562]
Observations	271	271	271	271	271	271	163	163	163	271	271	271	271	271	271
R-squared	0.094	0.094	0.097	0.124	0.124	0.124	0.071	0.080	0.088	0.153	0.172	0.172	0.050	0.050	0.052
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The table 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). All indexes are standardized around their Control mean. Therefore, estimates are expressed in standard deviations from the control mean. The analysis is restricted to treatment districts. "Trained in T" is an indicator equal to 1 for trained officers in Treatment districts, and 0 otherwise. "N trained (-i)" is the number of trained officers in a district, excluding officer i. 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 (8 hypotheses). Controls include gender, age, wage and officer rank at baseline, officer graduation year, number of officers in the district, intrinsic motivation to serve the community when he/she joined the police, as measured at baseline, and behavior in the baseline mind game.

Table A18: Heterogeneous effects by number of trained officers: Game-generated outcomes (treatment districts only)

	Number of tails			Coin flip: N > 2 tails (y/n)			3 tails (y/n)		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Trained	-0.243*	-0.254*	-0.518**	-0.168**	-0.163**	-0.179	-0.101*	-0.089	0.017
	(0.134)	(0.144)	(0.219)	(0.061)	(0.065)	(0.109)	(0.055)	(0.060)	(0.114)
	[0.201]	[0.468]	[0.296]	[0.049]	[0.200]	[0.611]	[0.201]	[0.576]	[1.000]
N trained (-i)		0.006	-0.005		-0.003	-0.004		-0.007	-0.002
		(0.017)	(0.018)		(0.010)	(0.010)		(0.008)	(0.008)
		[0.986]	[1.000]		[0.986]	[0.999]		[0.897]	[1.000]
Trained x N trained (-i)			0.022			0.001			-0.009
			(0.016)			(0.006)			(0.007)
			[0.765]			[1.000]			[0.765]
Observations	271	271	271	271	271	271	271	271	271
R-squared	0.050	0.051	0.057	0.058	0.059	0.059	0.030	0.032	0.036
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The table 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 to 3, the dependent variable is the number or reported tails, ranging from 0 to 4. In columns 4 to 6, the dependent variable is a dummy equal to 1 if the officers reported more than 2 tails. In columns 7 to 9, the dependent variable is a dummy equal to 1 if the officer reported 3 tails. The analysis is restricted to treatment districts. "Trained in T" is an indicator equal to 1 for trained officers in Treatment districts, and 0 otherwise. "N trained (-i)" is the number of trained officers in a district (excluding officer i). Standard errors clustered at the police district level are reported in parentheses (8 hypotheses). Multiple hypothesis corrected p-values using the Romano-Wolf (2005) procedure are reported in square brackets. Controls include gender, age, wage and officer rank at baseline, officer graduation year, number of officers in the district, intrinsic motivation to serve the community when he/she joined the police, as measured at baseline, and behavior in the mind game at baseline.

Table A19: 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	0.218	0.248	0.658
Age (1/5: 3=35-44 years)	3.055	3.193	0.356
Year graduated (baseline)	2002.509	2001.648	0.529
Officer rank (baseline)	3.818	3.848	0.910
Wage (baseline)	2052.073	2003.352	0.600
Intrinsic motivation to serve the community y/n	0.609	0.573	0.674
Intrinsic motivation to serve the community y/n (baseline)	0.582	0.572	0.905
Education most effective against corruption y/n (baseline)	0.091	0.124	0.513
Comfortable reporting corruption 1/5 (baseline)	3.273	3.317	0.848
Know where to report corruption 1/5 (baseline)	4.236	4.352	0.349
Ever reported unethical behaviour y/n (baseline)	0.255	0.345	0.224
Police corrupt 1/5 (baseline)	2.527	2.614	0.677
Witness unethical behaviour 1/4 (baseline)	2.836	2.834	0.990
Witness bribe 1/4 (baseline)	2.564	2.634	0.692
Police as Service Provider y/n (baseline)	0.382	0.283	0.178
Aggressive more useful than curteous 1/5 (baseline)	3.091	3.283	0.391
Overall workplace satisfaction 1/5 (baseline)	3.709	3.772	0.652
Mind Game: Number matched	0.655	0.621	0.660
How often monitor juniors 1/4 (baseline)	3.586	3.653	0.641
Caught juniors' unethical behaviour 1/4 (baseline)	2.379	2.681	0.173
Disciplined juniors 1/4 (baseline)	1.621	2.181	0.020

Notes: We obtained *clean* monthly district-level administrative data for 17 districts, 4 in the Control and 13 in the Treatment group, from January 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 A20: Treatment effects on District-Level Administrative Data (per officer per month)

	Total filed cases	Total cases sent to court	Total cases convicted	Court fines
Panel A				
Post	0.092	0.087	0.074	42.334
	(0.071)	(0.059)	(0.054)	(36.465)
Post × Training	0.018	0.020	0.029	10.456
	(0.051)	(0.043)	(0.039)	(26.308)
	[0.713]	[0.639]	[0.466]	[0.744]
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 × Training	0.256**	0.242**	0.237***	126.382**
	(0.119)	(0.099)	(0.089)	(60.904)
	[0.066]	[0.082]	[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 × Training	0.075	0.087	0.107	122.917**
	(0.119)	(0.099)	(0.089)	(60.904)
	[0.519]	[0.377]	[0.256]	[0.071]
After 6 months	-0.061	-0.064	-0.074	-45.333
	(0.093)	(0.077)	(0.070)	(47.651)
After 6 months × Training	-0.016	-0.013	-0.004	-15.897
	(0.053)	(0.044)	(0.040)	(27.305)
	[0.750]	[0.761]	[0.922]	[0.667]
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. "Total cases sent to court" refers to the monthly cases sent to court on average per officer. "Total cases convicted" refers to the number of convicted cases on average per officer. "Court fines" refers to the monthly amount of issued fines (in Ghanaian Cedi) on average per officer. Standard errors reported in parentheses, Wild Clustered Bootstrap (Restricted) p-values reported in square brackets (as per Roodman et al. (2019) to account for the small number of districts