Supplemental Appendix

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

By Donna Harris, Oana Borcan, Bruno Schettini, Danila Serra, Henry Telli, and Stefan Dercon

APPENDIX A: FIGURES AND TABLES



FIGURE A1. AGENT OF CHANGE PIN DELIVERED TO TRAINED OFFICERS AT THE AWARD CEREMONY

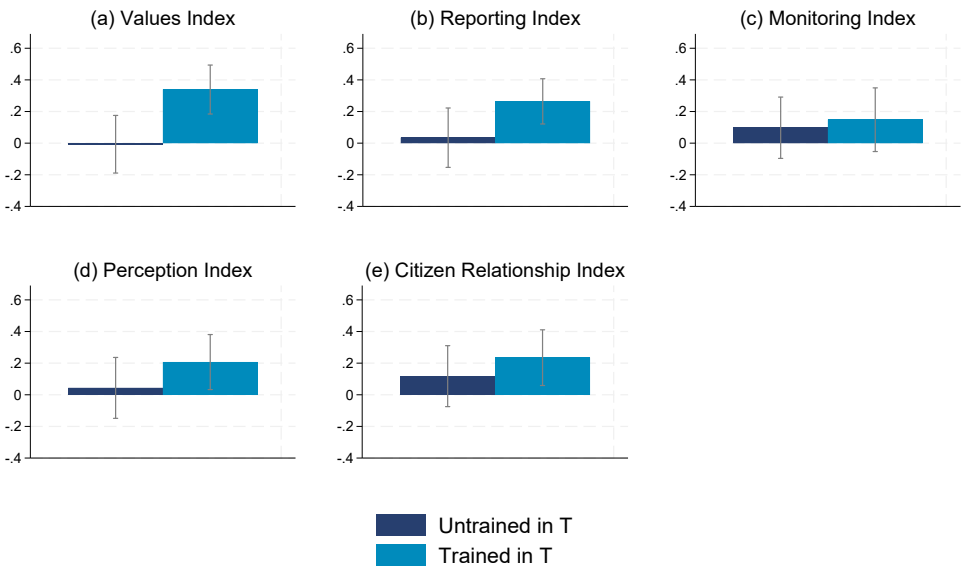


FIGURE A2. SURVEY-BASED INDEXES

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

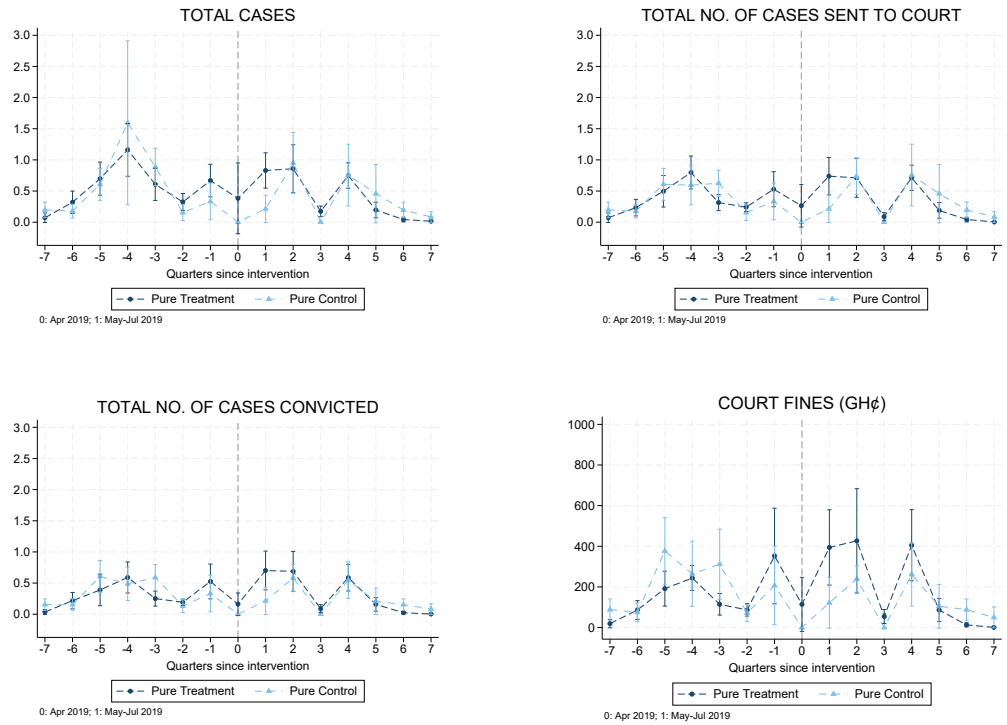
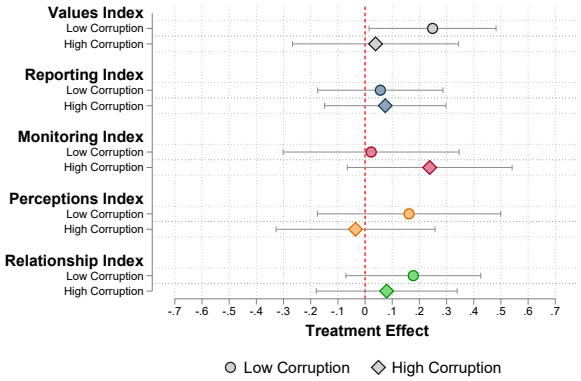


FIGURE A3. DISTRICT-LEVEL ADMINISTRATIVE DATA ON FIELD BEHAVIOR

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

(a) Survey Outcomes



(b) Cheating Game Outcomes

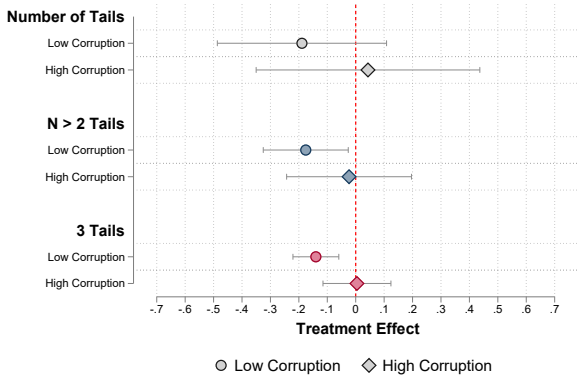


FIGURE A4. HETEROGENEOUS IMPACTS BY DISTRICT-LEVEL CORRUPTION AT BASELINE

Notes: The figures display estimated intent-to-treat effects and 90% confidence intervals from regressions interacting the Treatment indicator with a dummy for above-median district-level corruption at baseline. Baseline corruption is measured using officers' reports of how often they witnessed or heard of unethical behavior or bribery in their district, averaged across questions and districts, and dichotomized at the median. All indexes are standardized around their Control mean; estimates are expressed in standard deviations from the control mean. Panel (b) displays the estimated treatment impacts on the number reported in the cheating game, the likelihood of reporting a number greater than 2, and the likelihood of reporting the number 3. For each outcome, we display ITT estimates. Standard errors are clustered at the police district level, and the specifications generating the reported estimates include Double Lasso controls.

TABLE A1—LIKELIHOOD OF ATTRITION IN TREATMENT VERSUS CONTROL DISTRICTS

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

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

Table A2—Aggregate survey-based indexes

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

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

Table A3—Treatment Effects on Job Satisfaction

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

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

TABLE A4—RANDOMIZATION INFERENCE

	Values Index	Reporting Index	Monitoring Index	Perceptions Index	Relationship Index	Number of tails	N> 2 tails (y/n)	3 tails (y/n)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: ITT estimates								
Treatment District	0.145	0.101	0.146	0.068	0.159	-0.078	-0.109	-0.080
	(0.091)	(0.086)	(0.105)	(0.117)	(0.082)	(0.122)	(0.066)	(0.041)
	[0.144]	[0.299]	[0.177]	[0.617]	[0.105]	[0.551]	[0.156]	[0.108]
Control Mean								
Observations	412	412	248	412	412	412	412	412
LASSO Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Panel B: Treatment Effects								
Trained in T	0.273	0.135	0.154	0.127	0.238	-0.162	-0.169	-0.116
	(0.100)	(0.091)	(0.129)	(0.146)	(0.103)	(0.145)	(0.073)	(0.046)
	[0.015]	[0.161]	[0.205]	[0.243]	[0.041]	[0.112]	[0.001]	[0.031]
Untrained in T	-0.032	-0.074	0.140	0.026	0.050	0.092	0.005	-0.006
	(0.105)	(0.119)	(0.133)	(0.155)	(0.094)	(0.136)	(0.069)	(0.051)
	[0.764]	[0.472]	[0.299]	[0.794]	[0.648]	[0.370]	[0.919]	[0.896]
Control Mean	0.000	0.000	-0.000	-0.000	0.000	2.553	0.610	0.461
Trained=Untrained (p-val)	0.005	0.060	0.925	0.585	0.035	0.030	0.002	0.030
Observations	412	412	248	412	412	412	412	412
LASSO Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

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

Table A5—Balance Tests - Control vs Central district officers

	Mean Control	Mean Central	p-value
Panel A: Demographics			
District size at baseline	23	273	0.000
Female	0.277	0.128	0.356
Age (1/5: 3=35-44 years)	3.014	2.880	0.462
Year graduated (baseline)	2003.454	2004.489	0.673
Officer rank (baseline)	3.589	3.346	0.542
Wage (baseline)	1939.582	1863.767	0.781
Intrinsic motivation to serve the community y/n (baseline)	0.631	0.617	0.807
Panel B: Outcome measures at baseline			
Education most effective against corruption y/n (baseline)	0.121	0.158	0.389
agreement	3.170	3.421	0.629
agreement	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 (endline)	0.333	0.406	0.239
Aggressive more useful than curteous 1/5 (baseline)	3.220	3.421	0.728
Overall workplace satisfaction 1/5 (baseline)	3.638	3.594	0.829
Panel C: Behavior in the baseline cheating game			
Mind Game: Number matched	0.624	0.549	0.233
Panel D: Outcome measures at baseline (restricted sample)			
How often monitor juniors 1/4 (baseline)	3.566	3.566	0.998
Caught juniors' unethical behaviour 1/4 (baseline)	2.453	2.618	0.379
Disciplined juniors 1/4 (baseline)	1.774	2.026	0.201
Observations (max)	141	133	

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

TABLE A6—ROBUSTNESS - TREATMENT EFFECTS, INCLUDING CENTRAL DISTRICT OFFICERS

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

Notes: The table reports estimates obtained when including officers working in the Central district office. Standard errors clustered at the police district level are reported in parentheses. Multiple hypothesis corrected p-values using the Romano-Wolf (2005) procedure are reported in square brackets. Due to small number of clusters (32), we report Wild-bootstrapped p-values in curly brackets. The controls are selected using the Double Lasso procedure (Belloni, Chernozhukov and Hansen, 2014; Urminsky, Hansen and Chernozhukov, 2016).

TABLE A7—ROBUSTNESS - TREATMENT EFFECTS, EXCLUDING TOP 10% IMPROVEMENT IN SURVEY SCORES

	Values Index	Reporting Index	Monitoring Index	Perceptions Index	Relationship Index	Number of tails	N > 2 tails (y/n)	3 tails (y/n)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: ITT estimates								
Treatment District	0.112	0.067	0.148	0.143	0.202	-0.066	-0.113	-0.074
	(0.097)	(0.094)	(0.116)	(0.113)	(0.082)	(0.122)	(0.063)	(0.048)
	[0.394]	[0.394]	[0.394]	[0.394]	[0.038]	[0.492]	[0.081]	[0.115]
	{0.278}	{0.497}	{0.211}	{0.250}	{0.031}	{0.622}	{0.110}	{0.160}
Control Mean	0.000	0.000	0.000	-0.000	-0.000	2.548	0.619	0.460
Observations	370	370	219	370	370	370	370	370
R-squared	0.075	0.096	0.088	0.158	0.060	0.012	0.029	0.028
LASSO Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Panel B: LATE estimates								
Treatment District	0.244	0.150	0.333	0.310	0.434	-0.142	-0.245	-0.161
	(0.198)	(0.206)	(0.259)	(0.257)	(0.171)	(0.257)	(0.131)	(0.100)
	[0.412]	[0.412]	[0.412]	[0.412]	[0.037]	[0.499]	[0.082]	[0.099]
	{0.255}	{0.499}	{0.213}	{0.274}	{0.026}	{0.620}	{0.099}	{0.150}
Control Mean	0.000	0.000	0.000	-0.000	-0.000	2.548	0.619	0.460
Observations	370	370	219	370	370	370	370	370
R-squared	0.096	0.098	0.069	0.148	0.042	0.022	0.051	0.038
LASSO Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Panel C: Treatment effects								
Trained in T	0.317	0.116	0.141	0.123	0.218	-0.223	-0.209	-0.132
	(0.102)	(0.096)	(0.151)	(0.141)	(0.102)	(0.140)	(0.069)	(0.047)
	[0.029]	[0.673]	[0.829]	[0.829]	[0.133]	[0.178]	[0.014]	[0.022]
	{0.013}	{0.266}	{0.357}	{0.458}	{0.052}	{0.143}	{0.012}	{0.016}
Untrained in T	-0.007	-0.032	0.139	0.082	0.128	0.048	-0.017	-0.001
	(0.097)	(0.130)	(0.149)	(0.125)	(0.084)	(0.127)	(0.058)	(0.055)
	[0.938]	[0.938]	[0.829]	[0.829]	[0.400]	[0.888]	[0.904]	[0.986]
	{0.944}	{0.813}	{0.375}	{0.529}	{0.142}	{0.728}	{0.783}	{0.986}
Control Mean	0.000	0.000	0.000	-0.000	-0.000	2.548	0.619	0.460
Trained=Untrained (p-val)	0.001	0.251	0.992	0.794	0.305	0.065	0.007	0.038
Observations	370	370	219	370	370	370	370	370
R-squared	0.092	0.096	0.082	0.155	0.060	0.025	0.053	0.039
LASSO Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The table reports estimates obtained when excluding officers who are in the top 10% of the distribution based on the extent to which their survey outcome measures improved over their baseline measures. Standard errors clustered at the police district level are reported in parentheses. Multiple hypothesis corrected p-values using the Romano-Wolf (2005) procedure are reported in square brackets. Due to small number of clusters (32), we report Wild-bootstrapped p-values in curly brackets. The controls are selected using the Double Lasso procedure (Belloni, Chernozhukov and Hansen, 2014; Urminsky, Hansen and Chernozhukov, 2016).

TABLE A8—HETEROGENEOUS EFFECTS BY INITIAL INTRINSIC MOTIVATIONS

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

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

TABLE A9—HETEROGENEOUS EFFECTS BY OFFICER RANK

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

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

TABLE A10—HETEROGENEOUS EFFECTS BY NUMBER OF TRAINED OFFICERS

	Values Index	Reporting Index	Monitoring Index	Perceptions Index	Relationship Index	Number of tails	N > 2 tails (y/n)	3 tails (y/n)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Trained in T	0.518	0.134	-0.149	0.093	0.321	-0.514	-0.175	0.030
	(0.180)	(0.193)	(0.320)	(0.252)	(0.120)	(0.211)	(0.106)	(0.100)
	[0.120]	[0.933]	[0.971]	[0.971]	[0.155]	[0.099]	[0.291]	[0.962]
	{0.022}	{0.520}	{0.649}	{0.718}	{0.038}	{0.026}	{0.104}	{0.785}
N trained (-i)	0.011	0.009	-0.012	0.000	0.015	-0.023	-0.010	0.001
	(0.013)	(0.013)	(0.011)	(0.017)	(0.006)	(0.010)	(0.005)	(0.005)
	[0.920]	[0.933]	[0.822]	[1.000]	[0.191]	[0.118]	[0.226]	[0.962]
	{0.461}	{0.538}	{0.363}	{0.980}	{0.013}	{0.086}	{0.159}	{0.882}
Trained in T x	-0.018	0.002	0.015	-0.001	-0.015	0.023	0.001	-0.011
N trained (-i)	(0.010)	(0.014)	(0.020)	(0.017)	(0.010)	(0.016)	(0.006)	(0.006)
	[0.426]	[0.996]	[0.931]	[1.000]	[0.550]	[0.398]	[0.962]	[0.226]
	{0.139}	{0.883}	{0.479}	{0.972}	{0.366}	{0.147}	{0.834}	{0.064}
Observations	271	271	163	271	271	271	271	271
LASSO Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

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

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

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

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

TABLE A12—TREATMENT EFFECTS ON DISTRICT-LEVEL ADMINISTRATIVE DATA (PER OFFICER PER MONTH)

	Total filed cases	Total cases sent to court	Total cases convicted	Court fines
	(1)	(2)	(3)	(4)
Post	0.126	0.117	0.100	59.588
	(0.068)	(0.057)	(0.052)	(35.216)
Post x Treatment	0.021	0.023	0.033	12.282
	(0.051)	(0.043)	(0.039)	(26.206)
	[0.671]	[0.590]	[0.407]	[0.713]
Pre-Control Mean	0.18	0.14	0.11	62.91
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: The table reports estimates from a standard two-period difference-in-differences (DiD) model. "Post" refers to all months after May 2019. "Treatment" refers to the district assignment to treatment (1) or control group (0). We include time (month and year) fixed effects, as well as district fixed effects. Since data for 10 districts are aggregated into 4 administrative units, we conduct the analysis at the administrative unit level, i.e., we analyze data for 11 administrative units for 60 months, leading to 660 observations. The four dependent variables are monthly district-level outcomes, averaged over the number of officers in the district (or administrative unit). Given that the data are aggregated at the district level, and averaged by the number of officers, N, in a district, yet districts differ in N, we use analytical weights where, for each district, the weight is the number of officers in a district. "Total filed cases" refers to the monthly number cases filed on average per officer in a given district or administrative unit. "Total cases sent to court" refers to the monthly cases sent to court on average per officer in a given district or administrative unit. "Total cases convicted" refers to the number of convicted cases on average per officer in a given district or administrative unit. "Court fines" refers to the monthly amount of issued fines (in Ghanaian Cedi) on average per officer in a given district/administrative unit. Standard errors reported in parentheses. Wild Clustered Bootstrap (Restricted) p-values are reported in square brackets (as per Roodman et al. (2019)) to account for the small number of districts or administrative unit.

TABLE A13—TREATMENT EFFECTS ON SURVEY-GENERATED OUTCOMES WITHOUT CONTROLS

	Values Index	Reporting Index	Monitoring Index	Perceptions Index	Relationship Index
	(1)	(2)	(3)	(4)	(5)
Panel A: ITT estimates					
Treatment District	0.171	0.152	0.123	0.127	0.178
	(0.096)	(0.085)	(0.099)	(0.132)	(0.098)
	[0.194]	[0.194]	[0.272]	[0.272]	[0.194]
	{0.118}	{0.097}	{0.243}	{0.402}	{0.126}
Control Mean	0.000	0.000	-0.000	-0.000	0.000
Observations	412	412	248	412	412
LASSO Controls	No	No	No	No	No
Panel B: LATE estimates					
Treatment District	0.363	0.324	0.263	0.271	0.378
	(0.190)	(0.181)	(0.208)	(0.283)	(0.207)
	[0.155]	[0.155]	[0.260]	[0.265]	[0.155]
	{0.096}	{0.098}	{0.238}	{0.412}	{0.122}
Control Mean	0.000	0.000	0.000	-0.000	-0.000
Observations	412	412	248	412	412
LASSO Controls	No	No	No	No	No

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

APPENDIX B: THE TRAINING PROGRAM

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

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

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

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

The final component of Part 1 focused on effective communication with citizens and public perceptions of the police. The primary message was that each interaction with a citizen should be treated as the most important interaction, because

it only takes one *non-value-driven action* for public perceptions of the police *as a whole* to turn negative. The officers also learned about the principles of effective communication i.e., the 3 Vs – visual (i.e., body language), vocal (i.e., speech), and verbal (i.e., content) – and participated in a role-play exercise where some officers played in the role of citizens and others in the role of police officers. Each group took turns to enact the role play. A class discussion on communication strategies followed.

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

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

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

An important component of Day 2 was to show evidence that change is possible. To this end, officers watched a 12 minute documentary video about Inspector Schettini's own journey at the Federal Highway Police in Brazil. The main purpose of the documentary was to show a story and narratives that the officers could identify with, i.e., the experience of another police transport department, in which corruption was the norm, but where, by working together, officers managed to change the prevailing norms and permanently shift officers' mindsets. In the documentary, Inspector Schettini and his colleagues talked about how things were within their Department, e.g., how corruption was the norm, how change

started when a new Head of the Department took office and gained the support of a group of like-minded officers who wanted to change, and how the new norms are maintained through rigorous training programs.[1] 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.[2] The idea behind having officers come up with a group dance and chant was to reinforce a sense of unity between the officers participating in the training. The dance was performed by the full group at the end of the day, and served to cement their sense of belonging to the new group.

Post-training Reinforcements

We implemented two post-training reinforcements. First, two months after the training, in early July 2019, we created WhatsApp groups consisting of officers who had been trained together. The training facilitator moderated the groups and, over the following six months, regularly shared inspirational messages and reminders reinforcing the content of the training. Second, on December 23rd 2019, nearly eight months after the training, the Police held an award ceremony where all trained officers received a certificate and an "Agent of Change" lapel pin. These initiatives were designed not only to reinforce the training content, but also to carry symbolic and signaling value by visibly recognizing officers associated with the program and its professional standards.

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

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

APPENDIX C: Deviations from the Pre-Analysis Plan

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

Timing of data collection

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

Outcome Measures

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

With respect to the game-generated outcomes, we deviated from the PAP by implementing a different cheating game at endline. We had originally planned to

use the same "mind game" employed at baseline. However, this game required the rolling of a dice. Since the pandemic forced us to conduct a phone survey at endline, we opted for a cheating game that required the tossing of a coin. The assumption was that all respondents would be able to find a coin in their pockets or in their home, to be able to participate in the game over the phone.

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

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

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

Estimation strategy

Our estimation strategy closely follows our pre-analysis plan (PAP). We pre-specified the estimation of Intent-to-Treat (ITT) effects (Equation 1 in the PAP) and Local Average Treatment Effects (LATE) (Equation 2 in the PAP) as our

primary specifications. The resulting estimates are reported in Panels A (ITT) and B (LATE) of our primary regression tables.

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

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

REFERENCES

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014. "Inference on treatment effects after selection among high-dimensional controls." *Review of Economic Studies*, 81(2): 608–650.

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

Roodman, David, Morten Ørregaard Nielsen, James G. MacKinnon, and Matthew D. Webb. 2019. "Fast and wild: Bootstrap inference in Stata using boottest." *The Stata Journal*, 19(1): 4–60.

Urminsky, Oleg, Christian Hansen, and Victor Chernozhukov. 2016. "Using double-lasso regression for principled variable selection." *Available at SSRN 2733374*.

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