

Supplemental Appendix

Spillovers in State Capacity Building: Evidence from the Digitization of Land

Records in Pakistan

Shan Aman-Rana and Clement Minaudier

A Appendix tables

Table A.1: Effect of Digitization Reform on Tax Collection: Robustness to Defining Phase Onset at 50% Village Digitization.

Dependent variable:	Tax collection per cultivated acre	
	TWFE	
	OLS (1)	Median (2)
Digitization of land records	-8.784** (3.986) [3.89]	-6.799*** (2.288) [2.55]
Dep. var. mean	14.7	14.7
District fixed effects	Yes	Yes
Fiscal year fixed effects	Yes	Yes
Observations	212	212

Notes: The unit of observation is a district-fiscal year. ‘Digitization of land records’ is a dummy equal to one for phase 1 and 2 districts starting in FY2013 and zero otherwise. Unlike [Table 1](#), we do not report stacked DID results, as the simultaneous start of the intervention in treated districts makes them redundant. ‘Tax collection per cultivated acre’ divides tax collected in thousands of Pakistani Rupees by average district-level cultivated acres (in thousands) at baseline. Dependent variable mean is the average tax collected per acre across all districts and all years from FY2006 to FY2012, prior to any district’s digitization. Standard errors clustered at district level are in parentheses. Clustered bootstrapped standard errors (with 1000 replications) are in square brackets. Significance levels are denoted as: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A.2: Effect of Digitization Reform on Tax Collection: Robustness to Defining Phase Onset at 1% and 2% of Villages Digitized.

Timing of phases: Dependent variable:	First 1% villages digitized				First 2% villages digitized			
	TWFE		Tax collection per cultivated acre		TWFE		Stacked DID	
	OLS (1)	Median (2)	OLS (3)	Median (4)	OLS (5)	Median (6)	OLS (7)	Median (8)
Digitization of land records	-3.565 (2.683) [2.78]	-2.742 (1.767) [1.60]	-3.925 (2.968) [3.06]	-3.388* (1.787) [1.63]	-6.173** (2.937) [2.95]	-2.932* (1.629) [1.66]	-6.376* (3.246) [3.25]	-4.674*** (1.343) [1.58]
Dep. var. mean	13.3	13.3	13.3	13.3	13.3	13.3	13.3	13.3
District fixed effects	Yes	Yes	No	No	Yes	Yes	No	No
Fiscal year fixed effects	Yes	Yes	No	No	Yes	Yes	No	No
District-by-event fixed effects	No	No	Yes	Yes	No	No	Yes	Yes
Fiscal year-by-event fixed effects	No	No	Yes	Yes	No	No	Yes	Yes
Observations	178	178	330	330	212	212	376	376

Notes: The unit of observation is a district-fiscal year. 'Digitization of land records' is a dummy variable that takes value 1 for phase 1 and 2 districts starting in years FY2011 and FY2012 respectively, and remains zero otherwise. In columns (1)–(4), phase 3 districts become digitized in FY2013, while in columns (5)–(8), phase 3 districts become digitized in FY2014. Therefore, in Columns (1)–(4), we restrict the data to 2006–2012, as there is no available control group from FY2013 onward. 'Tax collection per cultivated acre' divides tax collected in thousands of Pakistani Rupees by average district-level cultivated acres (in thousands) at baseline. Dependent variable mean is the average tax collected per acre across all districts and all years from FY2006 to FY2010, prior to any district's digitization. Standard errors clustered at district level are in parentheses. Clustered bootstrapped standard errors (with 1000 replications) are in square brackets. Significance levels are denoted as: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A.3: Missing District-Year observations by Year and Treatment Phase

	2006	2007	2008	2009	2010	2011	2012	2013
Phase 1 (11 districts in total)	5	3	9	2	2	3	1	0
Phase 2 (13 districts in total)	4	4	7	0	1	0	2	0
Phase 3 (12 districts in total)	7	7	9	0	2	1	0	0

Notes. The table shows the number of districts for which we have no observation within each phase and each fiscal year.

Table A.4: Compliance with the planned rollout of digitization

	2010	2011	2012	2013	2014	2015
Phase 1	0	.02	.155	.611	.886	.91
Phase 2	0	0	.035	.614	.906	.934
Phase 3	0	0	0	.012	.583	.885

Notes: The variable displayed is the cumulative proportion of villages in a district that are digitized in a given fiscal year for each phase.

Table A.5: Effect of Digitization Reform on Tax Collection: Robustness to Controlling for Bureaucrats' Managerial Ability

Dependent Variables:	District level tax		Performance of bureaucrats			
	Tax collection per cultivated acre		$\frac{\text{Tax collected}}{\text{Tax demand}}$ (%)	Whether at least 50% tax demand was collected	Whether at least 75% tax demand was collected	Share of months with zero collection
	OLS (1)	Median (2)	OLS (3)	OLS (4)	OLS (5)	OLS (6)
Digitization of land records	-6.423* (3.676)	-5.083** (2.420)	-36.27*** (10.44)	-0.403*** (0.117)	-0.427*** (0.114)	0.274** (0.105)
	[3.791]	[1.934]	[11.04]	[0.126]	[0.121]	[0.112]
Avg. ability of managers	-1.355 (6.691)	-0.814 (3.800)	29.50** (12.61)	0.313** (0.151)	0.287* (0.150)	-0.301** (0.133)
	[6.925]	[3.929]	[14.58]	[0.173]	[0.179]	[0.154]
Dep. var. mean	14.2	14.2	53.9	0.53	0.43	0.19
District fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Fiscal year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Observations	210	210	301	301	301	301

Notes: In Columns (1) and (2), the unit of observation is a district-fiscal year, while in Columns (3) through (6) it is a bureaucrat-revenue circles-fiscal year. 'Digitization of land records' is a dummy variable that takes value 1 for phase 1 and 2 districts in years after FY2012 and FY2013 respectively, and remains zero otherwise. 'Tax collection per cultivated acre' divides tax collected in thousands of Pakistani Rupees by average district-level cultivated acres (in thousands) at baseline. The outcome variable in Column (3) is the ratio of tax collected to tax demand issued. The outcome variables in Columns (4) and (5) are dummy variables that take values 1 if at least 50% (75%) of the annual tax demand was collected, and remain zero otherwise. The outcome variable in column (6) is the share of months in the fiscal year in which no tax was collected by bureaucrats. For each column, dependent variable mean is the average of the respective outcome variable across all districts (Columns (1)-(2)) or bureaucrats-revenue circles (Columns (3)-(6)) and all years from FY2006 to FY2011, prior to any district's digitization. Standard errors clustered at district level are in parentheses. Clustered bootstrapped standard errors (with 1000 replications) are in square brackets. Significance levels are denoted as: * p<0.1, ** p<0.05, *** p<0.01.

Table A.6: Effect of Digitization Reform on Tax Collection: Placebo Test Using Alternative Digitization Start Years

Dependent variable:	Tax collection per cultivated acre			
	Phase 1 2006 Phase 2 2007		Phase 1 2009 Phase 2 2010	
Placebo:	OLS	Median	OLS	Median
	(1)	(2)	(3)	(4)
Digitization of land records	-3.633 (8.321) [8.281]	-0.230 (4.509) [6.769]	-6.814 (4.394) [4.495]	-1.587 (2.538) [2.495]
Dep. var. mean	11.0	11.0	11.6	11.6
District fixed effects	Yes	Yes	Yes	Yes
Fiscal year fixed effects	Yes	Yes	Yes	Yes
Observations	212	212	212	212

Notes: The unit of observation is a district-fiscal year. The placebo is based on defining the ‘digitization of land records’ variable as a dummy that takes value 1 for phase 1 and 2 districts from years FY2006 (FY2009) and FY2007 (FY2010) respectively (instead of FY2012 and FY2013 in the main specification). ‘Tax collection per cultivated acre’ divides tax collected in thousands of Pakistani Rupees by average district-level cultivated acres (in thousands) at baseline. Dependent variable mean is the average tax collected per acre across all districts in FY2006 for columns (1) and (2) and all districts and all years from to FY2006 to FY2008 for columns (3) and (4). Standard errors clustered at district level are in parentheses. Clustered bootstrapped standard errors (with 1000 replications) are in square brackets. Significance levels are denoted as: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A.7: Robustness to [Callaway and Sant'Anna \(2021\)](#) Estimator

Dependent variable:	Tax collection per cultivated acre Callaway & Sant'Anna (2021)	
	TWFE (1)	(2)
Digitization of land records	-6.57* (3.69)	-7.78** (3.68)
Dep. var. mean	14.2	14.2
District fixed effects	Yes	Yes
Fiscal year fixed effects	Yes	Yes
Observations	212	212

Notes: This table is based on unstacked data. The unit of observation is a district-fiscal year. 'Digitization of land records' is a dummy variable that takes value 1 for phase 1 and 2 districts in every year from FY2012 and FY2013 respectively, and remains zero otherwise. 'Tax collection per cultivated acre' divides tax collected in thousands of Pakistani Rupees by average district-level cultivated acres (in thousands) at baseline. Dependent variable mean is the average tax collected per acre across all districts and all years from FY2006 to FY2011, prior to any district's digitization. Standard errors clustered at district level are in parentheses. Significance levels are denoted as: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A.8: Effect of Digitization Reform on Tax Collection: Event-by-event effects

Dependent variable:	Tax collection per cultivated acre			
	Event 1	Event 2	Event 1	Event 2
	Phase 1	Phase 2	Phase 1	Phase 2
	vs. Control	vs. Control	vs. Control	vs. Control
	OLS	OLS	Median	Median
	(1)	(2)	(3)	(4)
Digitization of land records	-3.365 (5.399) [5.662]	-11.62** (4.932) [4.824]	-6.293*** (2.242) [3.345]	-4.714 (2.976) [4.610]
Dep. var. mean	14.2	14.2	14.2	14.2
District fixed effects	Yes	Yes	Yes	Yes
Fiscal year fixed effects	Yes	Yes	Yes	Yes
Observations	199	195	199	195

Notes: Data is at the district-fiscal year level. 'Tax collection per cultivated acre' divides tax collected in thousands of Pakistani Rupees by average district-level cultivated acres (in thousands) at baseline. Phase 1 is a dummy that takes the value one for all the districts that were planned to be digitized in phase 1 in any years after FY2012, and remains zero otherwise. Phase 2 is a dummy that takes the value one for all the districts that were planned to be digitized in phase 2 in any years after FY2013, and remains zero otherwise. Phase 3 are districts that are never treated in the sample period 2006-2013. Dependent variable mean is the average tax collected per acre across all districts and all years from FY2006 to FY2011, prior to any district's digitization. Standard errors clustered at district level are in parentheses. Clustered bootstrapped standard errors (with 1000 replications) are in square brackets. Significance levels are denoted as: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A.9: Effect of Digitization Reform on Tax Collection: Randomization Inference p -value

Dependent variable:	Tax collection per cultivated acre			
	TWFE		Stacked DID	
	OLS	Median	OLS	Median
	(1)	(2)	(3)	(4)
Digitization_d $\times$ post_t	-7.146*	-5.768**	-6.746*	-4.940**
	(3.970)	(2.570)	(3.935)	(1.969)
	[4.011]	[1.892]	[3.974]	[1.942]
Randomization inference p -val	0.082	0.004	0.028	0.001
Dep. var. mean	14.2	14.2	14.2	14.2
District fixed effects	Yes	Yes	Yes	Yes
Fiscal year fixed effects	Yes	Yes	Yes	Yes
Observations	212	212	394	394

Notes: The unit of observation is a district-fiscal year. ‘Digitization of land records’ is a dummy variable that takes value 1 for phase 1 and 2 districts, and remains zero otherwise. Post is a dummy variable that takes the value 1 in years after FY2012, and remains zero otherwise. ‘Tax collection per cultivated acre’ divides tax collected in thousands of Pakistani Rupees by average district-level cultivated acres (in thousands) at baseline. Randomization inference p -values (at the bottom of the table) are from permutation tests similar to the randomization based inference test (Athey and Imbens, 2017; Young, 2019). We re-assign digitization over districts 1000 times and compute the estimates under the null hypothesis that the treatment has no effect. We then locate the point estimates coming from our real data in the distribution of the 1000 treatment assignment simulations. The p -value is based on the share of estimates from the 1000 reassignments that are higher in absolute value than our point estimates in Table 1. Because we only carry out the reassignment over districts, we create separate dummy variables “Digitization” and “post”. As a result, we can only use a single year to capture the start of digitization, unlike the staggered adoption over two years in our main specification. Dependent variable mean is the average tax collected per acre across all districts and all years from FY2006 to FY2011, prior to any district’s digitization. Standard errors clustered at district level are in parentheses. Clustered bootstrapped standard errors (with 1000 replications) are in square brackets. Significance levels are denoted as: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A.10: Effect of Digitization Reform on Agricultural Tax Base: Robustness to using stacked data

	Farm profit per acre (1)	Satellite vegetation cover index (2)	Whether agri land irrigated? (3)	Log agricultural land owned (4)
Digitization of land records	4.906 (3.329) [3.546]	0.00846 (0.00548) [0.00538]	-0.0000514 (0.0490) [0.0443]	0.0635 (0.0444) [0.0432]
Dep. var. mean	23.4	0.53	0.12	7.69
District-by-event fixed effects	Yes	Yes	Yes	Yes
Fiscal year-by-event fixed effects	Yes	Yes	Yes	Yes
Observations	10,916	541	314,704	314,767

Notes: Unit of observation is a household-survey wave in column (1), a district-fiscal year in column (2) and a citizen-survey wave in columns (3) and (4). ‘Farm profit per acre’ is the difference between value of output and total expenses per acre, based on HIES data sourced from [Beg \(2022a\)](#) (restricted to cultivating households), across survey waves 2005, 2007, 2011, and 2013. For this measure, ‘Digitization of land records’ is a dummy variable equal to 1 for phase 1 and 2 districts in the 2013 wave, and 0 otherwise, and dep. var. mean is average profit per acre across all districts and across waves 2005, 2007, and 2011. ‘Satellite vegetation cover index’ is the Normalized Difference Vegetation Index (NDVI) (ranging from -1 to 1), obtained from NASA’s MODIS land products. For this measure, ‘Digitization of land records’ is a dummy variable equal to 1 for phase 1 and 2 districts in every year from FY2012 and FY2013 respectively, and zero otherwise, and dep. var. mean is the average value of the index across all districts and all years from FY2006 to FY2011, prior to any district’s digitization. ‘Whether agricultural land irrigated’ is a dummy variable equal to 1 when the household’s agricultural land is irrigated, based on PSLM survey data. ‘Agricultural land owned’ measures the acres of agricultural land owned by households based on PSLM survey data. We use the 2006, 2008, 2010 and 2012 waves of the survey. For these two measures, ‘Digitization of land records’ is a dummy variable equal to 1 for phase 1 districts in the 2012 wave and 0 otherwise. Dep. var. mean are the respective average of each outcome variables (in levels, not in logs for both columns) across all districts and across waves 2006, 2008 and 2010. Standard errors clustered at district level are in parentheses. Clustered bootstrapped standard errors (with 1000 replications) are in square brackets. Significance levels are denoted as: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A.11: Bureaucrats' assessments of the tax base (district-level): Robustness to using stacked data

	Log assessed cultivated area (1)	Log admin tax demands (2)
Digitization of land records	-0.106*** (0.0351) [0.0339]	-0.568*** (0.195) [0.198]
Dep. var. mean	1069.2	28685.6
District-by-event fixed effects	Yes	Yes
Fiscal year-by-event fixed effects	Yes	Yes
Observations	393	376

Notes: The unit of observation is a district-fiscal year. 'Digitization of land records' is a dummy variable that takes value 1 for phase 1 and 2 districts in every year from FY2012 and FY2013 respectively, and remains zero otherwise. The reported cultivated area is measured in thousands of acres, while the administrative tax targets is in thousands of Pakistani Rupees. Dependent variable mean is the average assessed cultivated area and tax demand (in levels, not logs) across all districts and all years from FY2006 to FY2011, prior to any district's digitization. Standard errors clustered at district level are in parentheses. Clustered bootstrapped standard errors (with 1000 replications) are in square brackets. Significance levels are denoted as: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A.12: Did the digitization reform affect the performance of bureaucrats: Robustness to using bureaucrat FE

Dependent Variables:	Performance of bureaucrats			Assessment
	$\frac{\text{Tax collected (\%)}}{\text{Tax demand}}$	Whether at least 50% tax demand was collected	Whether at least 75% tax demand was collected	Share of months with zero collection
	(1)	(2)	(3)	(4)
				(5)
Digitization of land records	-29.54** (14.09) [12.21]	-0.338** (0.158) [0.138]	-0.394** (0.152) [0.135]	0.225 (0.139) [0.119]
Dep. var. mean	53.9	0.53	0.43	0.19
Bureaucrat fixed effects	Yes	Yes	Yes	Yes
Fiscal year fixed effects	Yes	Yes	Yes	Yes
Observations	304	304	304	304
				875.2 Yes Yes 301

Notes: The unit of observation is a bureaucrat-fiscal year. 'Digitization of land records' is a dummy variable that takes value 1 for phase 1 and 2 districts in every year from FY2012 and FY2013 respectively, and remains zero otherwise. The first measure is the ratio of the tax they collected to the tax demand they issued. The second and third measure are dummy variables that take values 1 if at least 50% (75%) of the annual tax demand was achieved, and remains zero otherwise. The fourth measure is the share of months in the fiscal year in which no tax was collected. The final column presents results for tax assessments conducted by bureaucrats. It reports the total tax demands (in 000 Pakistani Rupees) issued by bureaucrats within each revenue circle. For each column, dependent variable mean is the average of the respective outcome variable across all bureaucrats and all years from FY2006 to FY2011, prior to any district's digitization. Standard errors clustered at district level are in parentheses. Clustered bootstrapped standard errors (with 1000 replications) are in square brackets. Significance levels are denoted as: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A.13: Did the digitization reform affect the performance of bureaucrats?
Robustness to using stacked data

Dependent Variables:	Performance of bureaucrats			
	$\frac{\text{Tax collected}}{\text{Tax demand}} (\%)$	Whether at least 50% tax demand was collected	Whether at least 75% tax demand was collected	Share of months with zero collection
	(1)	(2)	(3)	(4)
Digitization of land records	-36.70*** (9.355) [9.749]	-0.405*** (0.107) [0.115]	-0.430*** (0.100) [0.107]	0.241** (0.100) [0.103]
Dep. var. mean	53.9	0.53	0.43	0.19
District-by-event fixed effects	Yes	Yes	Yes	Yes
Fiscal year-by-event fixed effects	Yes	Yes	Yes	Yes
Observations	549	549	549	549

Notes: The unit of observation is a bureaucrat-fiscal year. 'Digitization of land records' is a dummy variable that takes value 1 for phase 1 and 2 districts in every year from FY2012 and FY2013 respectively, and remains zero otherwise. The first measure is the ratio of the tax they collected to the tax demand they issued. The second and third measure are dummy variables that take values 1 if at least 50% (75%) of the annual tax demand was achieved, and remains zero otherwise. The final measure is the share of months in the fiscal year in which no tax was collected. For each column, dependent variable mean is the average of the respective outcome variable across all bureaucrats and all years from FY2006 to FY2011, prior to any district's digitization. Standard errors clustered at district level are in parentheses. Clustered bootstrapped standard errors (with 1000 replications) are in square brackets. Significance levels are denoted as: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A.14: Effect of Digitization Reform on Tax Collection: Testing the Information Mechanism

Dependent Variables:	Performance of bureaucrats			
	$\frac{\text{Tax collected}}{\text{Tax demand}}(\%)$	Whether at least 50% tax demand was collected	Whether at least 75% tax demand was collected	Share of months with zero collection
	(1)	(2)	(3)	(4)
Digitization	-38.52*** (12.80) [13.28]	-0.427*** (0.142) [0.149]	-0.486*** (0.132) [0.143]	0.300** (0.135) [0.140]
Digitization × Variance in tax demands	8.706 (11.99) [13.62]	0.0904 (0.153) [0.168]	0.194 (0.151) [0.174]	-0.103 (0.143) [0.152]
Dep. var. mean	53.9	0.53	0.43	0.19
District fixed effects	Yes	Yes	Yes	Yes
Fiscal year fixed effects	Yes	Yes	Yes	Yes
Observations	304	304	304	304

Notes: The unit of observation is a bureaucrat-fiscal year. ‘Digitization of land records’ is a dummy variable that takes value 1 for phase 1 and 2 districts in every year from FY2012 and FY2013 respectively, and remains zero otherwise. ‘Variance in tax demands’ is a dummy variable equal to one if a district’s baseline variance in tax demands is above the median. The first outcome measure is the ratio of the tax they collected to the tax demand they issued. The second and third measure are dummy variables that take values 1 if at least 50% (75%) of the annual tax demand was achieved, and remains zero otherwise. The final measure is the share of months in the fiscal year in which no tax was collected. For each column, dependent variable mean is the average of the respective outcome variable across all bureaucrats and all years from FY2006 to FY2011, prior to any district’s digitization. Standard errors clustered at district level are in parentheses. Clustered bootstrapped standard errors (with 1000 replications) are in square brackets. Significance levels are denoted as: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A.15: Effect of Digitization Reform on Tax Collection: Robustness to using Machado and Silva (2019)'s Quantile Regression

Dependent variable:	Tax collection per cultivated acre	
	TWFE Median	
	Unstacked (1)	Stacked (2)
Digitization of land records	-6.472* (3.477) [3.523]	-7.009** (3.535) [3.649]
Dep. var. mean	14.2	14.2
District fixed effects	Yes	No
Fiscal year fixed effects	Yes	No
District-by-event fixed effects	No	Yes
Fiscal year-by-event fixed effects	No	Yes
Observations	212	394

Notes: The unit of observation is a district-fiscal year. 'Digitization of land records' is a dummy variable that takes value 1 for phase 1 and 2 districts in every year from FY2012 and FY2013 respectively, and remains zero otherwise. 'Tax collection per cultivated acre' divides tax collected in thousands of Pakistani Rupees by average district-level cultivated acres (in thousands) at baseline. Dependent variable mean is the average tax collected per acre across all districts and all years from FY2006 to FY2011, prior to any district's digitization. Standard errors clustered at district level are in parentheses. Clustered bootstrapped standard errors (with 1000 replications) are in square brackets. Significance levels are denoted as: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

B Appendix: Figures

Figure B.1: An example of a manual land record

[illegible]

Notes: The source of the figure is *Figure 2* in [Adeel \(2010\)](#). Before the digitization reform such land records were maintained by the bureaucrats we study.

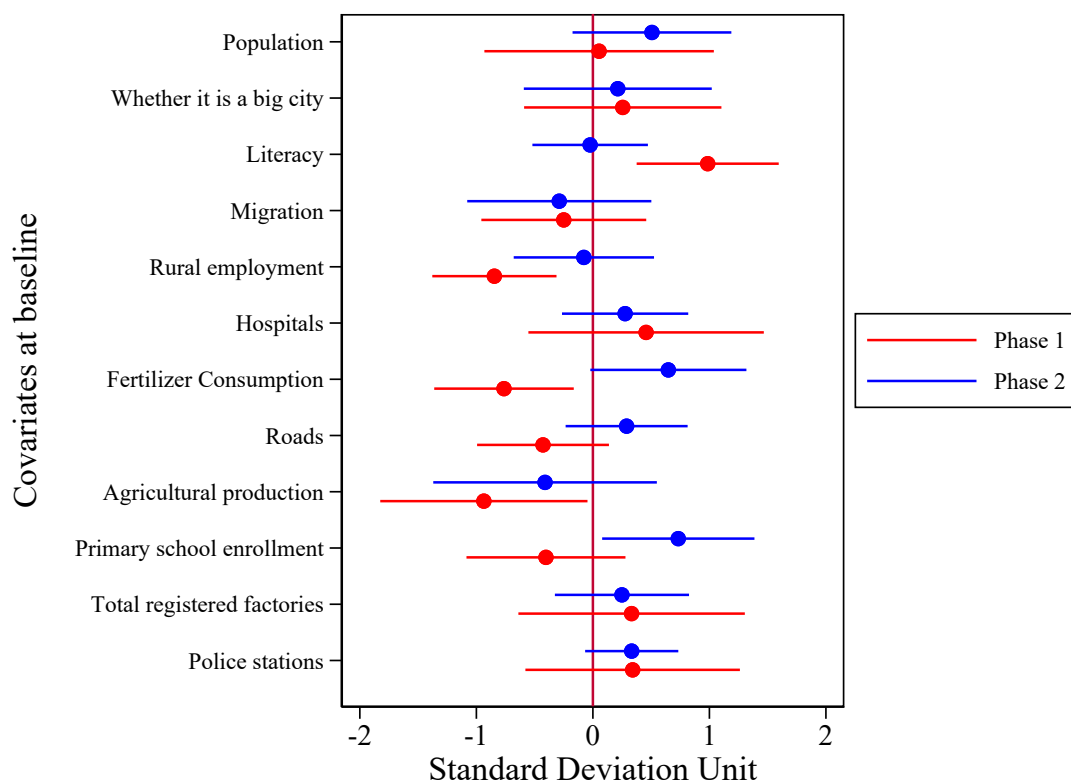
[illegible]

Figure B.3: A new bureaucracy set up to handle digitized land records



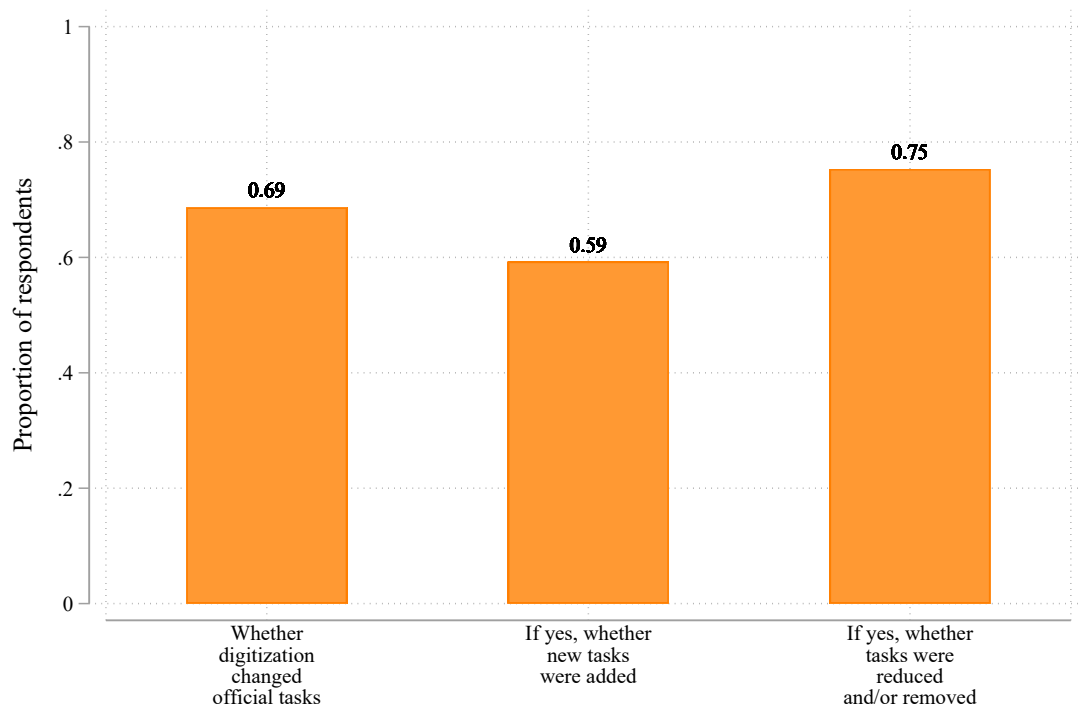
84

Figure B.4: Balance test on baseline characteristics of districts in phase 1 and 2 of the digitization reform



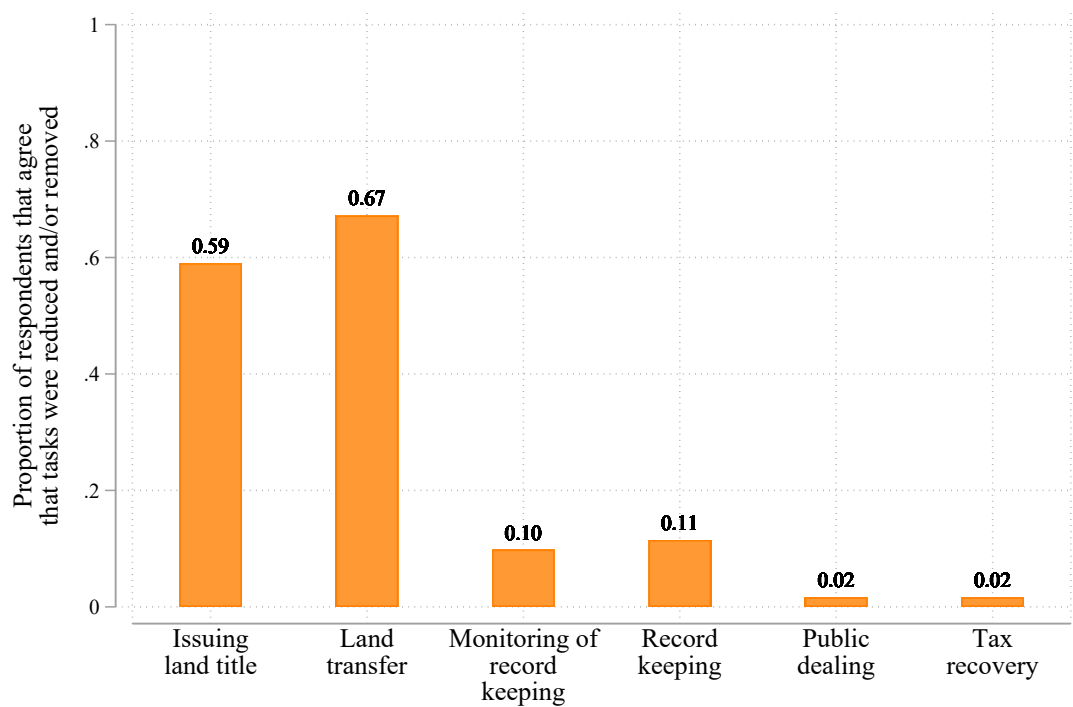
Notes: Data on baseline characteristics are from the Development Statistics of the Pakistan Bureau of Statistics 1997-2010. The point estimates are from a regression of the respective covariates on a dummy that takes value one if the district is in phase 1 or 2 of the digitization reform, and zero otherwise. The reference category are phase 3 districts. Intervals are 95% confidence intervals.

Figure B.5: Changes in the tasks of the bureaucrats after the digitization reform



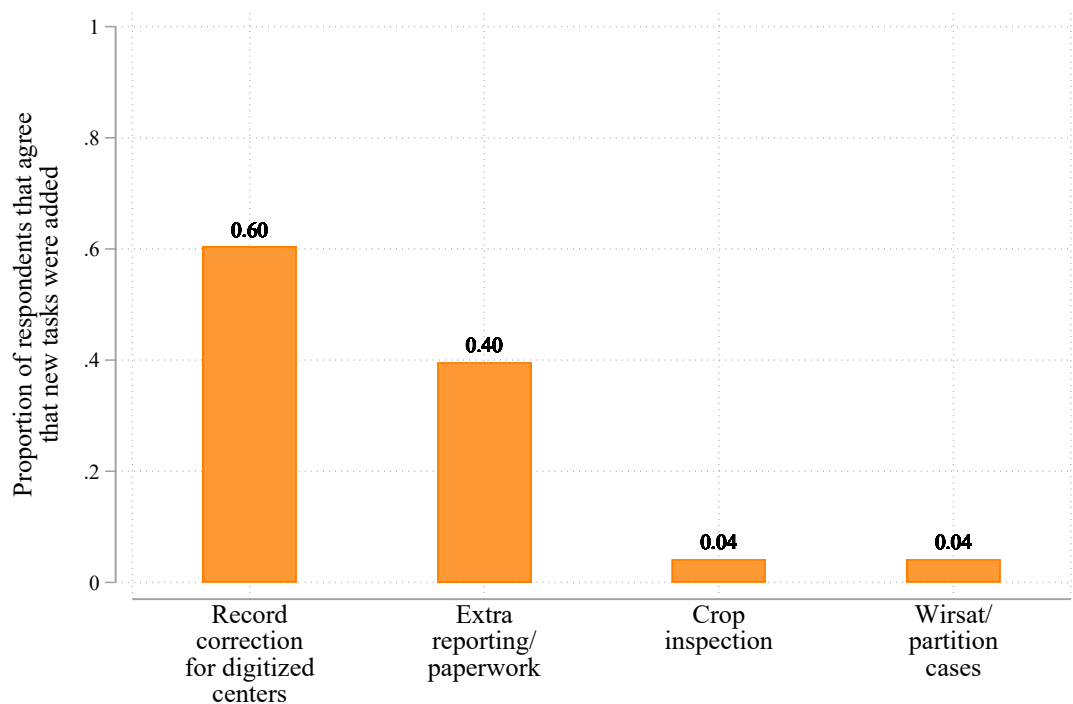
Notes: The figure is based on the bureaucrat survey restricted to the 118 bureaucrats who served as *Qanungo* between 2006-2013. The first bar plots the proportion of bureaucrats that agreed in the question “Do you think LRMIS (the digitization reform) changed the official tasks that you are supposed to do?”

Figure B.6: Bureaucrats' tasks reduced after the digitization reform



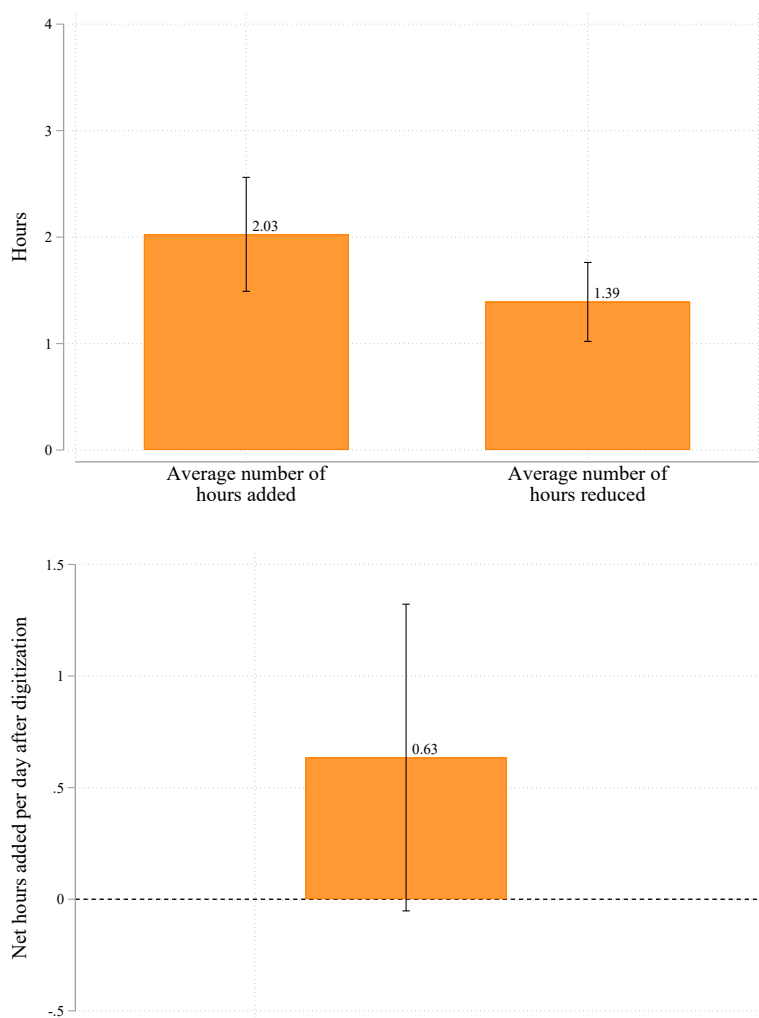
Notes: The figure is based on the bureaucrat survey restricted to the 81 bureaucrats who served as *Qanungo* between 2006-2013 and who agreed with the question “Do you think LRMIS (the digitization reform) changed the official tasks that you are supposed to do?”

Figure B.7: Bureaucrats' tasks added after the digitization reform



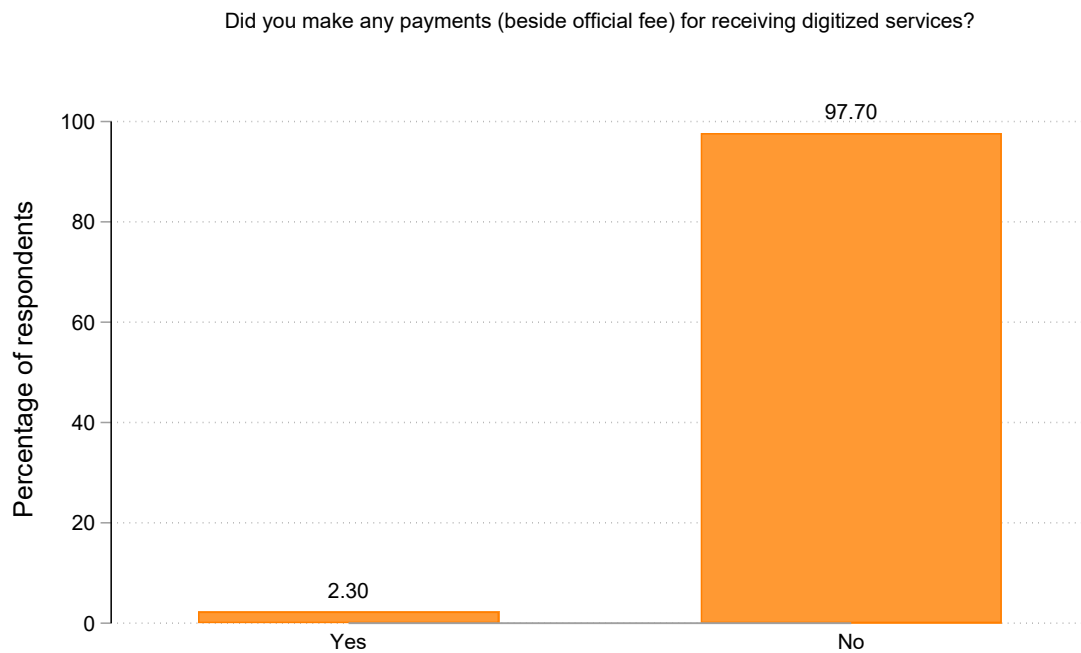
Notes: The figure is based on the bureaucrat survey restricted to the 81 bureaucrats who served as *Qanungo* between 2006-2013 and who agreed with the question “Do you think LRMIS (the digitization reform) changed the official tasks that you are supposed to do?”

Figure B.8: Changes in number of hours worked after the digitization reform



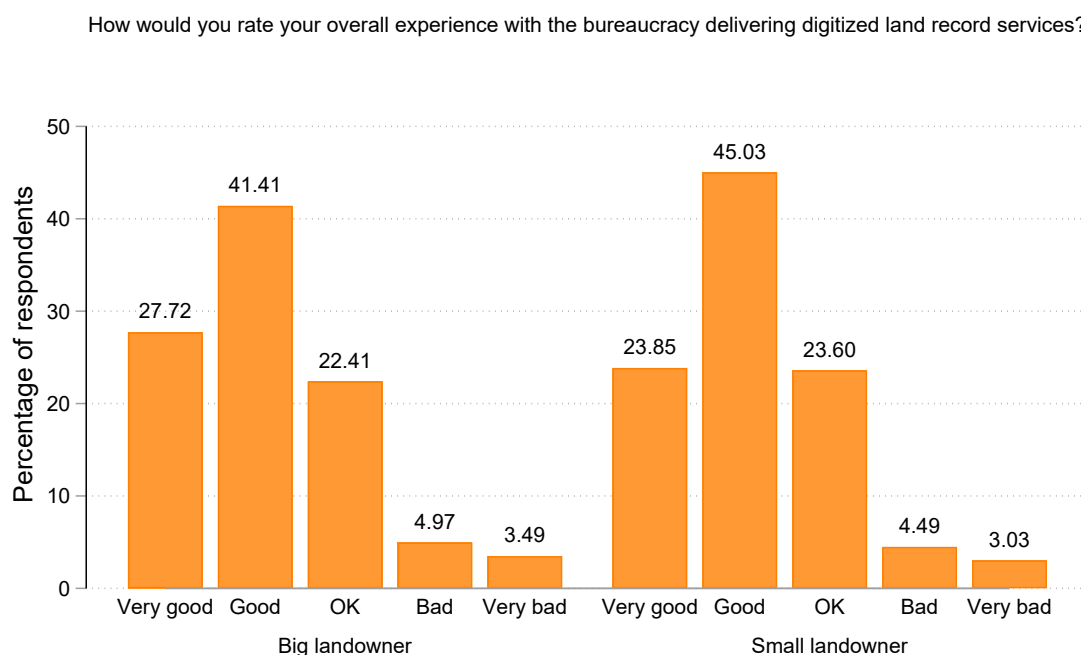
Notes: The figure is based on the bureaucrat survey restricted to the 118 bureaucrats who served as *Qanungo* between 2006-2013. The figure is based on responses to the question “Do you think LRMIS (the digitization reform) has changed the official tasks that you are supposed to do? If so, what is the number of hours per day that were added / reduced because of these changes?” Based on these answers, we calculate the average number of hours added and the average number of hours reduced across respondents. These are reported in the top panel. In the bottom panel we report the average net change in hours with 95% confidence intervals. The number is calculated by subtracting the number of hours reduced per day from the number of hours added per day, as reported by the bureaucrats.

Figure B.9: Bribe payments for getting land record services from the new bureaucracy



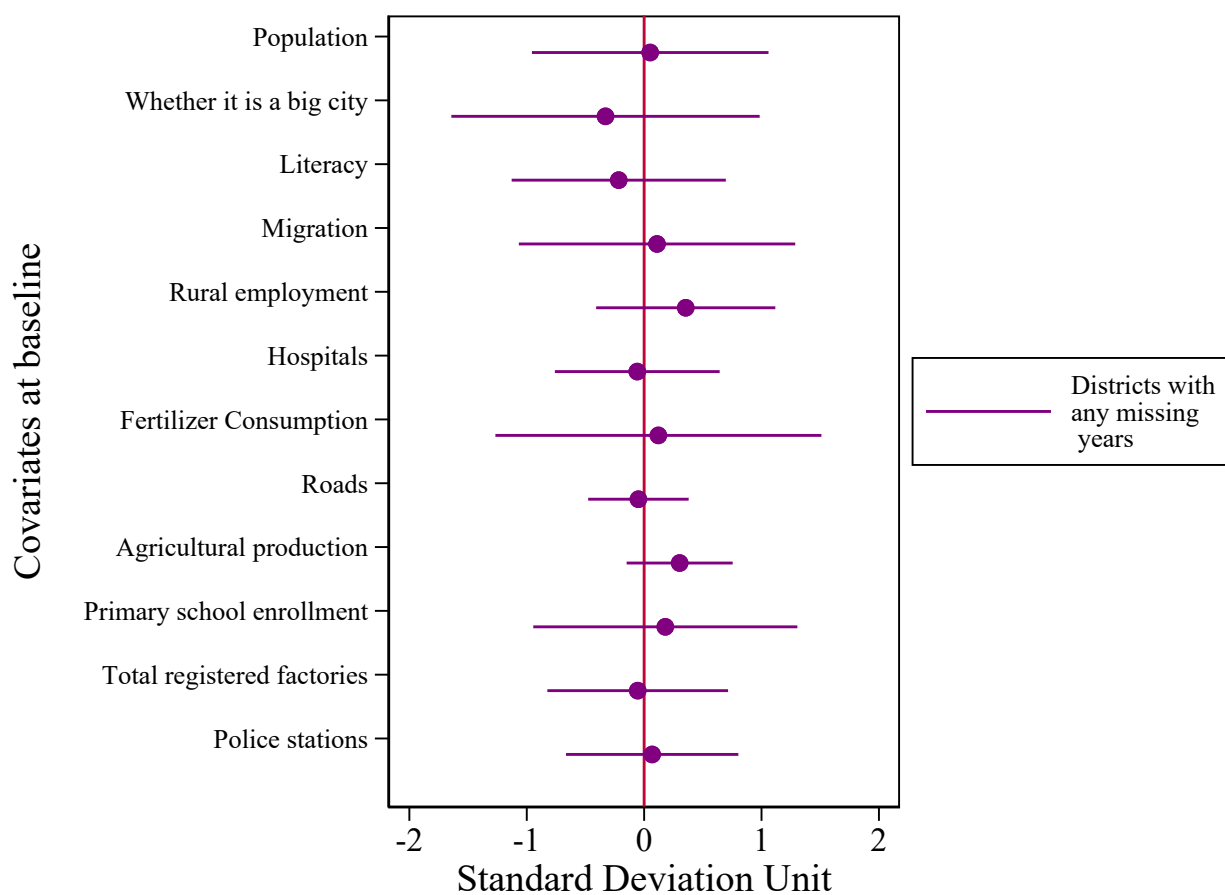
Notes: This data is from an endline citizen exit survey in 2016 on the services provided by the bureaucracy responsible for delivering digitized land record services.

Figure B.10: Citizens' satisfaction with the new bureaucracy delivering digitized land record services



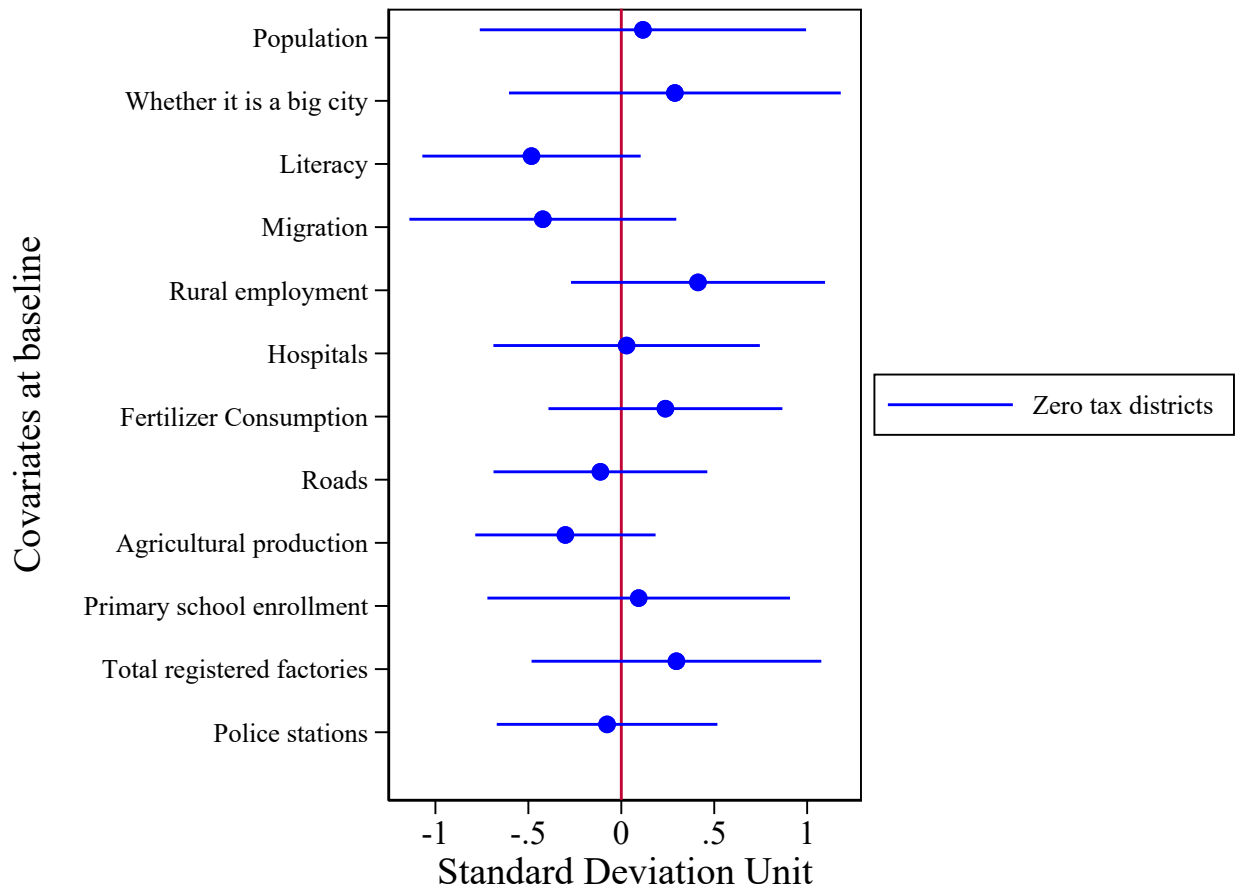
Notes: This data is from an endline citizen exit survey in 2016 on the services provided by the bureaucracy responsible for delivering digitized land record services. Big landlords are defined as those who own land acreage above the median.

Figure B.11: Balance on Baseline Characteristics of Districts with missing years



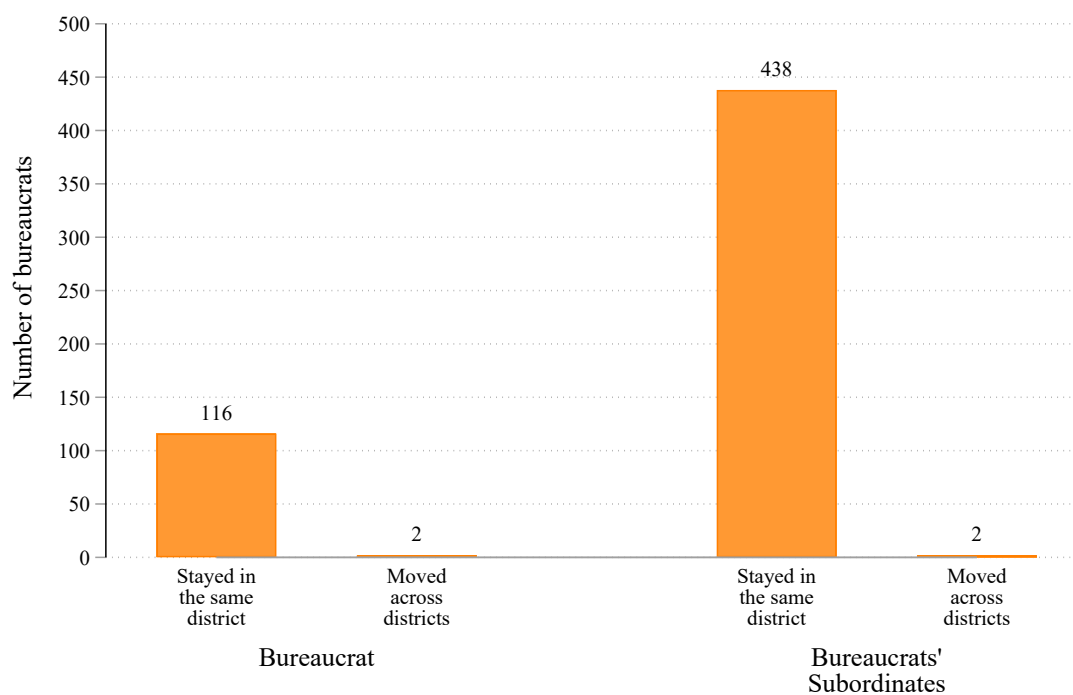
Notes. The point estimates are from a regression of the respective covariates on a dummy equal to one if a district has at least one year with missing tax collection. The districts of Chiniot and Nankana are excluded, as baseline data for these districts are unavailable prior to 2011.

Figure B.12: Balance on Baseline Characteristics of Districts with Zero Tax Collection



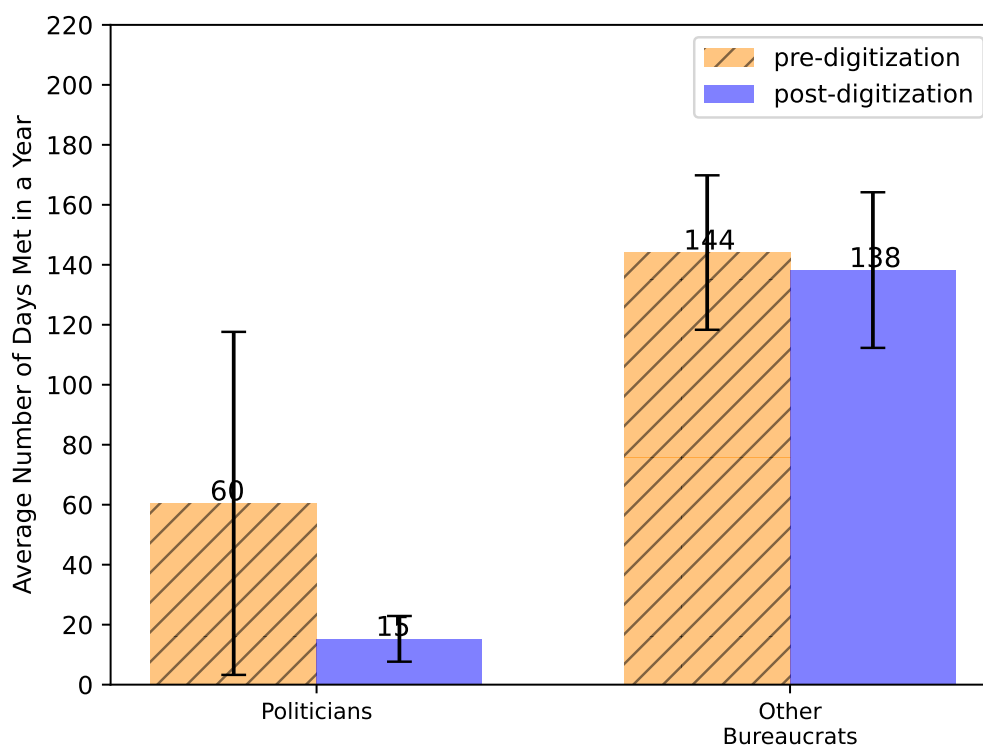
Notes. The point estimates are from a regression of the respective covariates on a dummy equal to one if a district has at least one year with zero tax collection. The districts of Chiniot and Nankana are excluded, as baseline data for these districts are unavailable prior to 2011.

Figure B.13: Movement of bureaucrats across districts



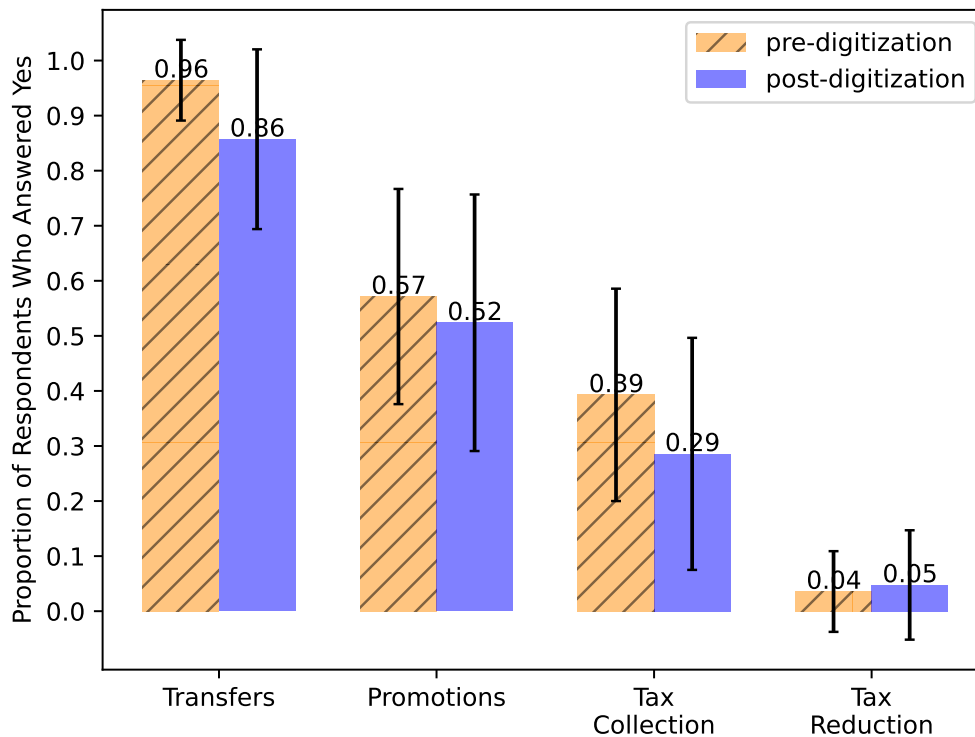
Notes: The figure is based on the bureaucrat survey. The left-hand side shows transfers among the bureaucrats that are the focus of this paper (Qanungos). The right-hand side shows transfers among their subordinates (Patwaris).

Figure B.14: Bureaucrats' social interactions with politicians and other bureaucrats before and after the digitization reform



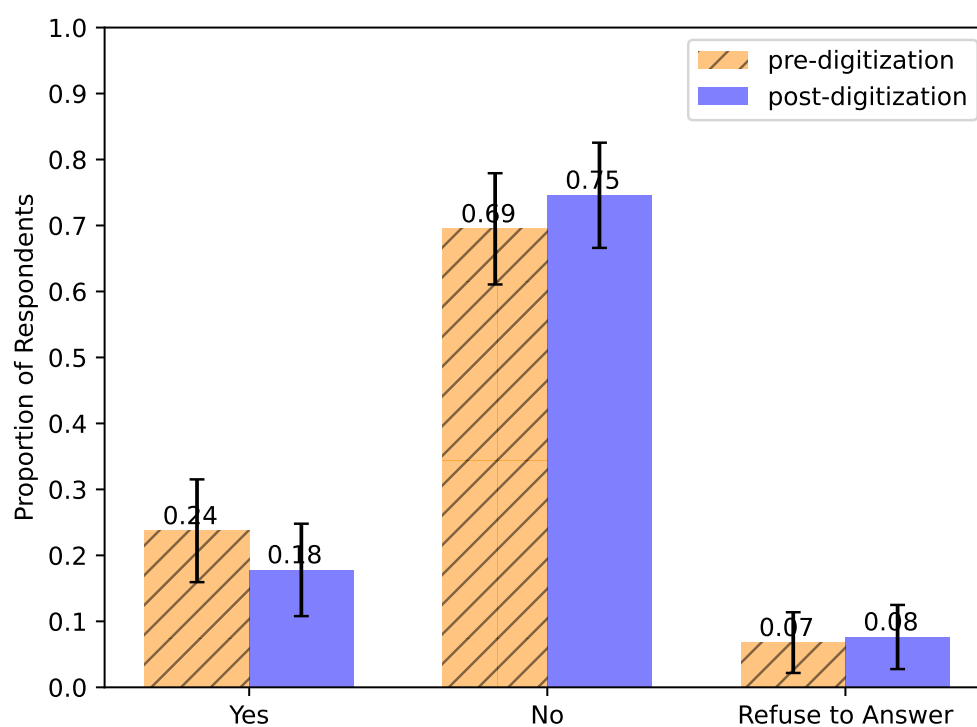
Notes: The figure is based on the bureaucrat survey restricted to the 118 bureaucrats who served as *Qanungo* between 2006-2013. We used a Likert scale to ask about the frequency of interactions between the respondent and politicians or other bureaucrats, before, and after, the reform. The Likert scale options were as follows: daily, twice a week, weekly, bi-monthly, monthly, quarterly, bi-annually, annually, less than once per year and never. We calculated the average number of days of interactions in a year for each bureaucrat based on these responses.

Figure B.15: Matters in which politicians interfere



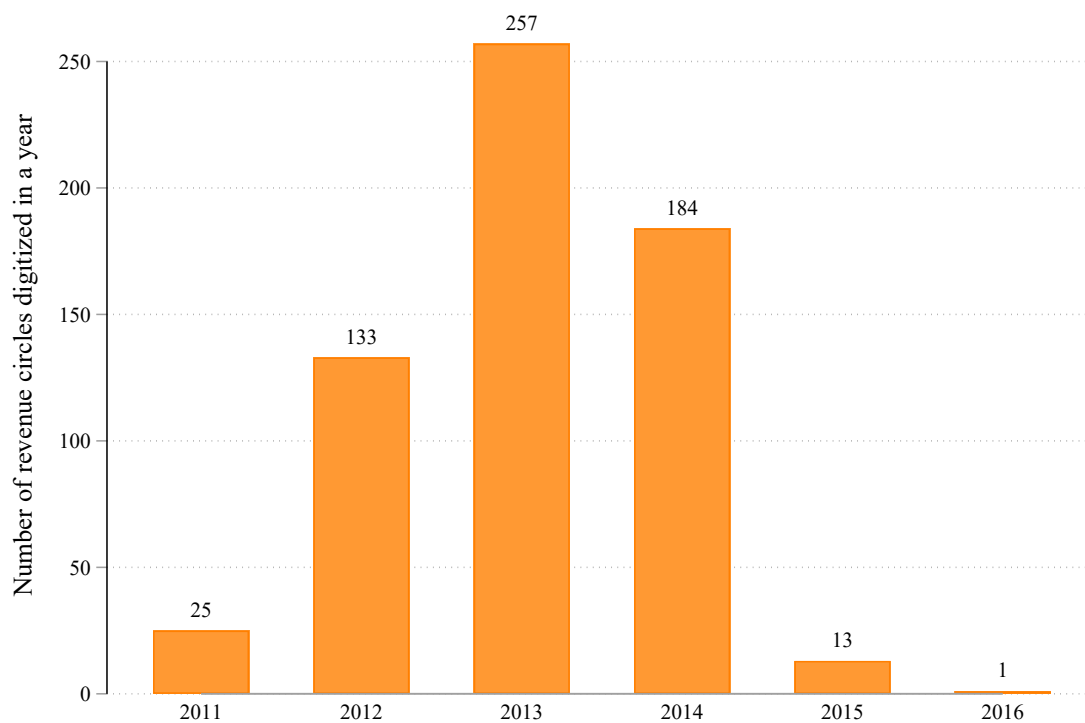
Notes: The figure is based on the bureaucrat survey restricted to the 118 bureaucrats who served as *Qanungo* between 2006-2013. Bureaucrats who responded 'yes' to the question: "In general, would you say that politicians interfere with the work of revenue officials in this revenue circle?" were further asked "On which matters politicians usually interfere with work?" The matters listed above were suggested by the research team along with the category of "others".

Figure B.16: Political interference in the work of bureaucrats



Notes: The figure is based on the bureaucrat survey restricted to the 118 bureaucrats who served as *Qanungo* between 2006-2013. The question was: "In general, would you say that politicians interfere with the work of revenue officials in this revenue circle?"

Figure B.17: Number of revenue circles digitized over the years



Notes: This figure displays the number of revenue circles digitized in a given year across all districts. A revenue circle is defined as digitized if at least one village in that revenue circle is digitized. In total, phase 1 covered 200 revenue circles, phase 2 covered 342 revenue circles, and phase 3 covered 275 revenue circles.

C Data Sources

Figure C.1: The Board Of Revenues' (BOR) record room



Figure C.2: The BOR tax collection pro forma

AGRICULTURAL INCOME TAX DISTRICT MUZAFFARGARH,
FOR THE MONTH OF September, 2007.
PREVIOUS A-I-T.

S. No.	Name of Tehsil	Demand	Suspension	Net Demand Recoverable	Previous Recovery	Current Recovery	Total Recovery	Balance	Percentage Month	Percentage Total
1-	M. Garh	17102682	—	17102682	76650	9300	85950	17016732	—	1%
2-	Kot Addu	28353571	—	28353571	87793	38100	125893	28227678	—	—
3-	Alipur	2079273	—	2079273	34150	44706	78856	2000417	2%	4%
4-	Taloti	18396542	—	18396542	50010	9500	59510	18337032	—	—
	Total A	65932068	—	65932068	248603	101606	350209	65581859	—	1%

CURRENT A-I-T.

1-	M. Garh	—	—	—	—	—	—	—	—	—
2-	Kot Addu	—	—	—	—	—	—	—	—	—
3-	Alipur	—	—	—	—	—	—	—	—	—
4-	Taloti	—	—	—	—	—	—	—	—	—
	Total B	—	—	—	—	—	—	—	—	—
	G.Total A+B	—	—	—	—	—	—	—	—	—

D Satellite vegetation cover data

We used NASA's MODIS land products to observe a satellite based vegetation cover index. MODIS vegetation indices provide consistent spatial and temporal comparisons of vegetation canopy greenness, a composite property of leaf area, chlorophyll and canopy structure. The normalized difference vegetation index (NDVI) are derived from atmospherically-corrected reflectance in the red, near-infrared, and blue wavebands. NDVI ranges from -1 to +1. If the NDVI values are negative it is highly likely that it's water. On the other hand, values close to +1 suggest that there is a high possibility that there are dense green leaves.³⁶

NASA's MODIS land products rely on the Sinusoidal Tile Grid System, which divides earth into 36×18 sinusoidal grids to locate a particular area on earth.³⁷ Since we are only interested in the Punjab province of Pakistan, our first step was to locate the tiles where Punjab is located. Given the shape file of districts of Punjab, we uniformly sample 10,000 points within each district and calculated their locations on the sinusoidal tile.³⁸ We find that all $36 \times 10,000$ points lie within three tiles, namely (horizontal 24, vertical 5), (horizontal 24, vertical 6), and (horizontal 23, vertical 5). There are a total of 4800×4800 pixels within each of the three tiles mentioned above, with each pixel having a $250 \text{ m} \times 250 \text{ m}$ size. Moreover, data for each year is divided into time intervals of 16 days. This results in 23 different time intervals in a given year. For each 16-day time interval in a year, and for each of the 36 districts of Punjab, we obtained the NDVI values of all pixels belonging to that particular district and took the average to get the NDVI value for that particular district in that particular 16-day interval. Since each year has 23, 16-day intervals, we end up having a list of 23 different NDVI values for a particular district in a particular year. Following the method used in Beg (2022a), we use the maximum value of that list as the NDVI value for that district for that year.

³⁶Details accessed at <https://gisgeography.com/ndvi-normalized-difference-vegetation-index/>

³⁷Accessed at: https://modis-land.gsfc.nasa.gov/MODLAND_grid.html

³⁸We decided to rely on this method rather than just consider the district's center and calculate their locations on the sinusoidal tile. This allows us a more holistic view of the vegetation cover of the district.

E Details on string matching revenue circles

We carried out an extensive string matching exercise to merge the tax, digitization, and bureaucrats' careers datasets. We took the following steps to merge the three sets of data:

- As a first step we manually checked each revenue circle, tehsil and district in the tax data against their counterparts recorded in the digitization data from the Punjab Land Records Authority (PLRA) and allocated a unique ID to each.³⁹ There were 1125 revenue circles in total, out of which 838 were given IDs using this process.
- The district, tehsil and revenue circle names in the bureaucrat survey data was manually cleaned. There were 690 unique revenue circles-tehsil-districts in this data, out of these we were able to give IDs to 458.
- Finally, we merged all three datasets on revenue circle, tehsil and district names.

We next checked the veracity of these data using further records from the government on details of revenue circle, tehsil and district names across Punjab. These were personally obtained from the government in 2020.

³⁹Digitization data from PLRA contained details of names of most of the tehsils and revenue circles except the following 19 (out of 141 in total) tehsils: Gujranwala Sadar, Kabirwala, Kharian, Shorkot, Khushab, Quaidabad, Jauharabad, Lahore city, Nishtar Town, Muzaffargarh, Depalpur, Renala khurd, Arifwala, Khanpur, Murree, Rawalpindi city, Rawalpindi Sadar, Rawalpindi Cantt and Daska.

F Sampling for the bureaucrat survey

The retrospective survey was carried out in 2020, with the main aim to rebuild career trajectories of bureaucrats between 2006-2013, the period for which tax data are available and reform effects can be estimated. Our sampling frame therefore, included people who were in charge of revenue circles (*Qanungo*) as well as people who in the recent past had worked as a *Qanungo*. These included bureaucrats that had risen through the ranks via promotions and were in-charge of the tehsils: (*Tehsildars* and *Naib-tehsildars*).

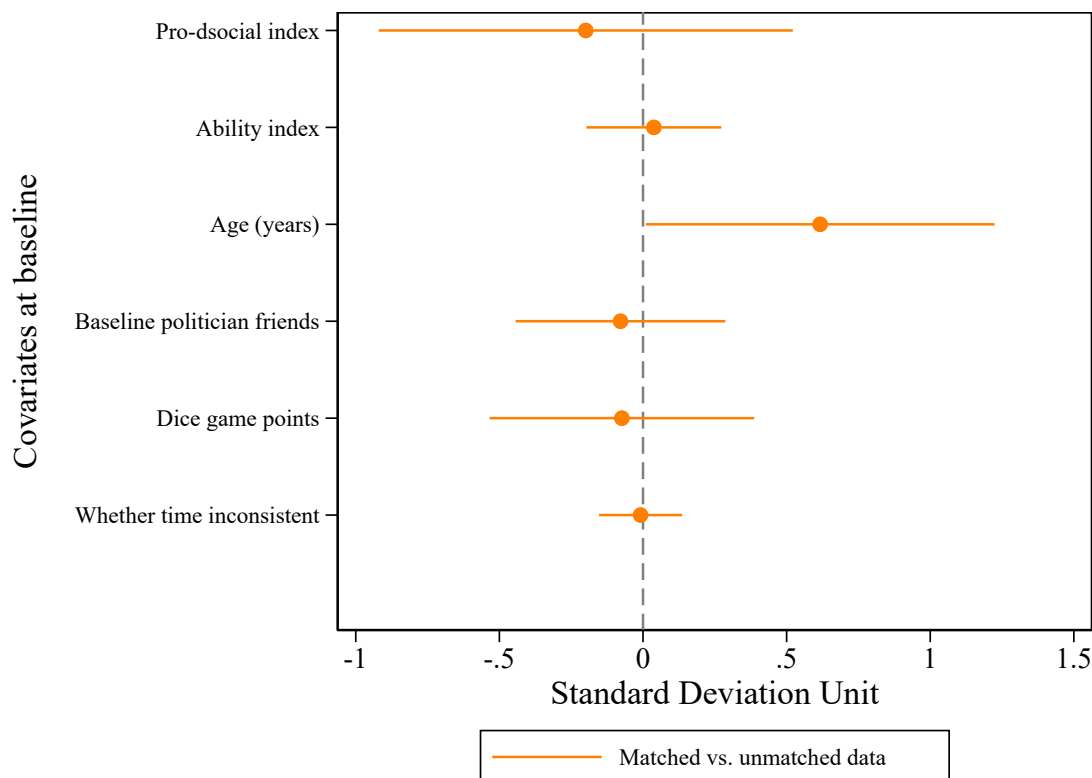
We stratified on districts and randomly sampled tehsils within each district.⁴⁰ We next created a sampling frame by contacting the local offices. Using that sampling frame we selected the universe of *Tehsildars* and *Naib-Tehsildars* working in the selected tehsils in Punjab. One *Qanungo* working with each of the *Naib-tehsildars* was randomly selected for the survey. We found 118 respondents who worked as *Qanungos* between 2006-2013.⁴¹

We could string-match the revenue circle names for 105 of those 118 respondents to match the survey data with the tax collection data. Of those the tax performance was missing for 27, so our final data set includes 78 respondents whose tax performance is observed between 2006-2013. In [Figure F.1](#) below, we examine the potential systematic differences between these bureaucrats and the broader sample across various characteristics, utilizing data gathered from the bureaucrat survey. The only covariate showing marginal statistical significance is age. The p -value resulting from a joint significance test of all covariates in the figure is 0.5290, providing evidence that the sample is not systematically selected based on characteristic of the bureaucrats.

⁴⁰Out of 141 tehsils in Punjab, we were able to survey bureaucrats from 138 tehsils. We were unable to survey the bureaucrats from the following three tehsils: Nishtar Town (Lahore districts), Shahkot (Nankana Sahib district), Ahmed Pur (Sheikhupura district).

⁴¹To find these, we started by surveying a total of 610 bureaucrats across different levels of hierarchy. Of those, 488 responded to the second round of telephonic survey about their career trajectory. The telephonic survey was used to recap the career paths of the bureaucrats, while their perceptions of digitization as well as their traits were measured in-person. In the pilot, the field team suggested this approach as the most effective way to achieve the maximum response rate, as the combined length of the two exceeded 1 hour and 30 minutes.

Figure F.1: Balance test on characteristics of the bureaucrats in the sample (that were matched with the tax data) and those who remained unmatched



Note: Data on bureaucrats' characteristics is from the bureaucrat survey. The point estimates are from a regression of the respective covariates on a dummy that takes the value of one if the bureaucrats in the survey data were matched with the tax data, and remains zero otherwise. Pro-sociality index is created from five measures: Inclusion of Others in Self scale ([Aron et al., 2004](#); [Ashraf et al., 2020](#)), whether they have donated blood, money donated in public good game, whether they do volunteer work and whether they give charity. Ability index is created from four measures: an incentivized matrix game and a memory game as in [Hanna and Wang \(2017\)](#), response to questions on general knowledge and revenue rules and regulations, respectively. Politician friends are the number of friends of the bureaucrats that are either federal or provincial politicians. Dice game points is a proxy for dishonesty and it is the total in an incentivized dice game as in [Hanna and Wang \(2017\)](#). Intervals are 95% confidence intervals.

G Marginal Value of Public Fund Calculations

The MVPF is calculated as $MVPF = \frac{\text{Benefits}}{\text{Net Govt Cost}} = \frac{\Delta W}{\Delta E - \Delta C}$. We explain the calculation behind each term in turn:

- **Benefits (ΔW):** The benefits included in the World Bank's cost-benefit analysis are "an increase in land value, reduced transaction costs, and revenue generation" (World Bank, 2017, p. 18). The World Bank incorporates this benefit as the Net Present Value (NPV) of the increase in land value from 2014 to 2023. Using the 10% discount rate from the World Bank report, the NPV of this benefit is equal to Rs. 34,560 million (World Bank, 2017, Table 5, Annex 3).⁴²
- **Cost of the reform (ΔE):** we calculate the cost of the reform as the net present value of the sum of the total capital costs and the total operating costs, as reported by the World Bank (World Bank, 2017), over the period 2007-2028.⁴³ The total capital cost of the reform (including software development, construction of the digitized centers, and hardware for the centers) amounts to Rs. 13,852 million (World Bank, 2017, Table 5, Annex 3), and its NPV is Rs. 5,837 million. The total operational cost of the reform (including staffing of the digitized centers, internet connection, and overheads) is Rs. 52,858 million (World Bank, 2017, Table 5, Annex 3), and its NPV is Rs. 13,125 million. The NPV of the total cost is therefore Rs. 18,962 million.
- **Long-run change in government cost (ΔC):** Our analysis suggests an additional effect of the reform on government costs due to lost government

⁴²Our analysis suggests an additional source of benefit: an increase in farmer's income. In our calculation, we have assumed that all benefits are reflected in land prices. However, if land markets are imperfect, the increase in income might not be fully reflected in land values. In this case, our MVPF calculation would understate the benefits of the reform. Extrapolating our own estimate of the increase in profits experienced by farmers of Rs. 4.9 per acre (Table 3) over the period 2012-2028 (from the start of the reform being implemented to the end of the World Bank's financial projections) and adding it to the benefit calculated by the World Bank would increase the total benefit of the reform to Rs. 35,587 million. This would increase the MVPFs to 1.88 (excluding the tax loss) and 1.75 (including the tax loss).

⁴³The time period for the costs begins from the first outlay of expenses towards the reform (which predates the start of the digitization as the software had to be developed and the centers built) and concludes at the end of the project's economic life, as determined in the World Bank's financial projections.

revenue. We calculate this as the loss in tax from our causal estimate, presented in [Table 1](#). We assume that the tax loss lasts throughout the period 2012-2028 in order to make it comparable to the benefit and direct cost.⁴⁴ Using the -6.74 loss in tax per acre estimate ([Table 1](#)), multiplied by the number of acres per districts and extrapolated across the time period, we obtain an NPV of Rs. 1,413 million.

- The MVPFs are therefore: $MVPF = \frac{34,560}{18,962} = 1.82$ excluding the tax loss, and $\frac{34,560}{18,962+1,413} = 1.70$ including the tax loss.

⁴⁴While the government could, in principle, offset tax losses during this period, in the absence of evidence of such interventions, we estimate the effect assuming no policy response.