Essays on Development and Political Economy

By

Muhammad Zia Mehmood

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Business Administration

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Edward Miguel, Co-chair
Professor Frederico Finan, Co-chair
Professor Guo Xu
Professor Steve Tadelis

Summer 2024

Essays on Development and Political Economy

Abstract

Essays on Development and Political Economy

by

Muhammad Zia Mehmood

Doctor of Philosophy in Business Administration

University of California, Berkeley

Professor Edward Miguel, Co-chair

Professor Frederico Finan, Co-chair

Low productivity in the private sector, especially amongst small businesses, and poor public sector service delivery are significant barriers to sustainable and equitable development in low-income countries. This dissertation studies these barriers through the lens of management challenges. It comprises of three chapters, where the first chapter examines the potential of SMS-based business trainings to address gaps in management skills amongst micro-entrepreneurs in Kenya, the second sheds light on the demand for SMS-based business trainings amongst these micro-entrepreneurs, and the third studies the effectiveness of a command and control management intervention implemented at scale in Punjab, Pakistan.

In my first chapter, titled *Short Messages Fall Short for Micro-Entrepreneurs: Experimental Evidence from Kenya*, I study the effectiveness of SMS-based business management trainings for improving outcomes for micro-entrepreneurs. SMS-based trainings are becoming a popular tool to remotely support micro-entrepreneurs in low-capacity contexts due to their scalability and low costs. However, little evidence exists on the effectiveness of such trainings to improve business outcomes. In this study, I evaluate a field experiment in which access to an SMS-based training was randomized across 4,700 micro-entrepreneurs in Kenya. After three months, I find positive effects on knowledge and adoption of best business practices. Younger entrepreneurs see stronger effects on sales, profits and business survival, driven by higher engagement with training content, more time spent on business, and getting larger loans. Contrary to predictions elicited from social scientists, I find that these positive effects disappear twelve months after the intervention, as all engagement with content ended within the first five months. Findings from this study suggest that, despite the promise and wide-spread use, SMS-based trainings are unlikely to be effective for micro-entrepreneurs in the long run. Results highlight the importance of lack of engagement as a major challenge limiting the potential of remotely provided information-based support.

My second chapter, titled *Demand for SMS-Based Business Trainings Amongst Kenyan Micro-Entrepreneurs*, studies the demand for SMS-based business management trainings in Kenya. I leverage two key components added to the field experiment from the first chapter to measure the demand; first, upon completion of the business management training, or prolonged disengagement from it, micro-entrepreneurs in the treatment group were given the opportunity to buy a second SMS training through Take-It-Or-Leave-It (TIOLI) offers where the asking price was randomized across three levels. Observing buying decisions across the randomized price levels allows me to study how demand changes with price, and also sheds light on correlations between individual and enterprise characteristics and demand. Second, I conduct an in-person demand elicitation activity with a select subset of the sample across the treatment and control groups, using a modified version of the Becker-DeGroot-Marschak (BDM) method. In the TIOLI sample, 70% of individuals chose to accept the additional training when it was offered for free, 68% accepted when the price was half the marginal cost faced by the service provider, and about 50% accepted when the price was double the marginal cost. In the BDM sample, the average willingness to pay for SMS trainings was five times the marginal cost, and almost a quarter of the respondents were willing to buy the training for ten times the marginal cost. Both methods of demand elicitation thus showed that micro-entrepreneurs were willing to pay a positive amount for SMS-based business management trainings. I also find correlational evidence suggesting demand for trainings was higher amongst individuals with more children in the household, those that recently applied for a loan, those with more knowledge of best practices, and those with higher education levels. Taken together with results from the first chapter, these findings suggest that engagement levels might not reflect true demand for SMS-based trainings, pointing towards possible behavioral explanations driving under utilization of the resource.

Finally, in my third chapter, titled *Command and Can't Control: Assessing Centralized Accountability in the Public Sector*, I study the potential and limitations of centralized management in the public sector, with Saad Gulzar, Juan Felipe Ladino, and Daniel Rogger. A long-established approach to management in government has been the transmission of information up a hierarchy, centralized decision-making by senior management, and corresponding centralized accountability; colloquially known as 'command and control'. We examine the effectiveness of a centralized management and accountability system implemented at scale in the public education sector bureaucracy of Punjab, Pakistan, for six years. The scheme automatically identified poorly performing schools and jurisdictions for the attention of central management. We find that flagging of schools and corresponding de facto punishments had no impact on school or student outcomes. We use detailed data on key elements of the education production function to show that command and control approaches to managing the general public sector do not induce bureaucratic action towards improvements in government performance.

Contents

List of Figures

List of Tables

بسم الله الرحمن الرحيم
الحمد لله رب العالمين

I dedicate this dissertation to my late grandmother, Salma Begum, who taught me to recognize my place and purpose in the universe - the greatest education one can receive in this world.

I also dedicate it to my parents, Samia Talat and Muhammad Talat Mehmood, who have always sacrificed their present for my future, and have been a source of unconditional and unwavering love and support throughout my life.

Acknowledgments

I am grateful for the guidance, mentorship and support provided by the members of my dissertation committee; Guo Xu was always willing to take out time to offer sage advice at every stage of my dissertation research. Steve Tadelis was a source of unwavering support as I powered through the ups and downs of my teaching, research and personal pursuits. Edward Miguel's commitment to rigorous, transparent, ethical, and impactful research has been a great source of inspiration for me. Frederico Finan's candid feedback always drove me to anticipate and pre-empt potential concerns in my research ideas, pushing me to be a better researcher.

I am extremely lucky to have had extraordinary mentors at different stages of my life; Syed Ali Hasanain took me under his wing when I was a clueless fresh graduate out of college, and inspired me to fix the world around me through research in development economics. He also introduced me to Muhammad Yasir Khan, Saad Gulzar, and Michael Callen, all of whom have been a consistent source of guidance and support ever since. This entire group also motivated me to apply for graduate studies in the US, and played an instrumental role in my journey to Harvard University and later to UC Berkeley.

After I completed my Masters in Public Policy at the Harvard Kennedy School, Saad introduced me to Daniel Rogger, who joined the list of extraordinary mentors who continue to illuminate my path and inspire me to continue challenging myself in the pursuit of actualizing my full potential.

I am also grateful for the invaluable feedback on my research projects provided by Ernesto Dal Bó, Ned Augenblick, Thiago Scot, Suraj Nair, Tianyu Han, Joel Ferguson, Jedediah Silver, Osman Siddiqi, and Mehmet Seflek. Jeff Ngugi provided excellent research assistance, and Carol Nekesa and her team at Remit-Kenya were an incredible source of support during fieldwork.

Lastly, I am blessed to have a wonderful family who has always supported my educational, professional and personal pursuits. My sisters, Saleha Talat, Shamsa Chaudhri, and Tayyaba Talat have always been a source of strength and compassion. I am forever indebted to my parents, Samia Talat and Muhammad Talat Mehmood; nothing I will ever accomplish in my life would be possible without their love, sacrifices, and prayers. I would have probably dropped out of the PhD program, if not for the unrelenting support of my wife, Stephanie Bonds - my best friend, my garment, my crown.

Co-author information: Chapter 3 of this dissertation is co-authored with Saad Gulzar from Princeton University, Juan Felipe Ladino from Stockholm University, and Daniel Rogger from the World Bank.

Chapter 1

Short Messages Fall Short for Micro-Entrepreneurs: Experimental Evidence from Kenya

1.1 Introduction

Employing 70% of the labor force world-wide and accounting for 40% of the GDP in emerging economies (ILO (2019); World Bank (2024)), small businesses form the economic backbone of society in low-income countries across the globe. Moreover, they are also crucial vehicles for female empowerment as at least one third of them are owned by women (World Bank (2020*b*)). Research aimed at exploring effective ways to address the challenges faced by small businesses is therefore key for poverty alleviation efforts.

Poor management practices is a major factor constraining firm productivity in low-income contexts (Bloom et al. (2010, 2013); Bloom and Van Reenen (2010*a*); Bruhn, Karlan and Schoar (2010); McKenzie and Woodruff (2017)). Business management trainings aimed at encouraging adoption of best practices are a popular tool employed to address this challenge,[1] and over $1 billion is spent annually to train 4-5 million entrepreneurs in low-income countries (McKenzie (2020)). However, most of these trainings are conventional in-person classroom-style trainings, which are expensive and hard to scale. Furthermore, most of them are conducted in or around large cities and often exclude entrepreneurs that are unable to take out time to participate in person, as well as those that are based in smaller cities and rural areas.

Phone-based trainings offer a potential solution to these challenges. In particular, SMS-based trainings are cheap, easy to scale, do not require in-person attendance or even internet

[1] The Start and Improve Your Business (SIYB) training program by the International Labor Organization has trained over 15 million entrepreneurs across the world (Mehtha (2017)), CEFE International has reached 13 million (Ramirez (2019)), International Finance Corporation's Business Edge training has reached over 100,000 entrepreneurs (*Business Edge : Status and Disposition* (2006)) etc.

access, and also allow targeted beneficiaries to move through the content at their own individual pace rather than having a single fixed pace for everyone. Due to this, SMS-based trainings are gaining popularity as a low-cost tool for information-based support across several low-income contexts.[2] However, despite the widespread use, little evidence exists on whether SMS-based trainings can be effective for improving outcomes, particularly for micro-entrepreneurs.

In this paper, I study the impact of SMS-based business trainings on business practices and outcomes for micro-entrepreneurs. To this end, I evaluate a field experiment whereby access to an SMS-based business training was randomized across 4,700 micro-entrepreneurs in Kenya. Data was collected by phone-based surveys conducted three months and twelve months after the intervention to estimate short and longer-run effects respectively. The main outcomes studied include knowledge and adoption of best practices, time spent on business and side jobs, labor employment decisions, credit outcomes, and business performance.

Kenya is an ideal setting for this study as Micro, Small, and Medium Enterprises (MSMEs) play a major role in the national economy; 7.4 million MSMEs engage over 90% of the active labor force in the country, and account for about a third of the GDP. Approximately 55% of these MSMEs are owned by women, and 98% are micro-enterprises[3]. The average education level for micro-entrepreneurs is approximately 11 years, yet adoption of basic best practices for business management is dismally low. Just under 80% of micro-entrepreneurs do not advertise any of their products in any way and almost 70% don't keep any type of business records to keep track of daily sales or expenses. Furthermore, less than 10% of micro-entrepreneurs account for prices of their competitors when setting prices for their own products and services. These statistics highlight a clear gap in management skills that could be constraining profitability, which can potentially be addressed through business management trainings, yet 90% of micro-entrepreneurs have never received any type of business training.[4]

The primary intervention in this study was aimed at addressing this gap through an SMS-based Business Education training course that used simply worded content to encourage micro-entrepreneurs to adopt business practices that have been shown to be highly correlated with profitability (Bloom and Van Reenen (2010*a*)). This training was developed in light of existing research on the importance of keeping training content simple in low-capacity contexts (Drexler, Fischer and Schoar (2014); Arráiz, Bhanot and Calero (2019)), by my implementation partner - a local firm that specializes in creation and dissemination of digital content. Available in English as well as Swahili, the content covered practices including marketing, advertising, pricing, record-keeping and stock management, and was divided into bite-sized chunks spanning approximately 150 text messages. These messages were pushed

[2] See Ulmann (2023); van Vark (2012); Haddad (2022); Hinrichsen and Ajadi (2020); *M-Shule SMS Learning & Training, Kenya — UIL* (2022), and work of TechnoServe (Regan-Sachs (2022)), and Arifu (*Arifu: WhatsApp Chatbot Provides Tips for Micro-Retailers* (N.d.)) etc.

[3] Less than 10 employees

[4] Statistics as of 2016, sourced from the country-wide *Micro, Small and Medium Enterprises Survey* (2016).

to micro-entrepreneurs through an interactive chat-bot in a fixed sequence, with a limited number of reminders being sent to those who stopped engaging at any point.

The sample of micro-entrepreneurs used in the study was sourced from a list of contacts maintained by my implementation partner in collaboration with a local microfinance institution. This list was compiled by them through fieldwork aimed at identifying micro-entrepreneurs to target for their products. Out of the 4,700 individuals recruited from this list for the study, 2,820 micro-entrepreneurs were randomly selected into the Treatment group and provided access to this SMS training course, while the remaining 1,880 - the Control group - received placebo messages aimed at reminding them about their business without conveying any substantive information about best business management practices. Approximately 300 individuals were surveyed three months after the intervention, while 2,780 individuals were surveyed after another nine months to estimate short and longer-run effects on business outcomes, respectively.

In order to determine if the main findings from this study depart from priors held by social science experts, I also conducted a survey through the Social Science Predictions Platform (Mehmood (2023)). In this survey, I described the study to social science researchers and elicited their predictions about how key outcomes will be affected twelve months after the intervention. This exercise allows me to shed light on whether the observed results from this study are expected and obvious, or surprising and informative.

Three months after the intervention, I find that assignment to treatment increased knowledge and adoption of best practices by 0.20 and 0.33 standard deviations, respectively. I also find large positive, but statistically insignificant, effects on business performance in the overall sample, and significant positive effects for younger (below-median) micro-entrepreneurs on sales (109% increase), profits (38% increase), and business survival (11.6 percentage points increase). These positive effects for younger entrepreneurs are driven by higher engagement with the content, and larger effects on time spent on business, and loan amounts applied for and received.

However, these positive results dissipate in the longer run; twelve months after the intervention, I see no effects on knowledge and adoption of best practices, as well as business sales, profits and survival. Additionally, the positive effects on business outcomes observed for younger entrepreneurs after three months, also disappear after twelve months. The time-trend of engagement reveals that the lack of longer-run effects is likely driven by micro-entrepreneurs abandoning all interactions with the content within the first few months of the training deployment, and well before the twelve-month follow-up.

I therefore conclude that, despite their growing popularity, SMS-based trainings on their own are unlikely to be effective for micro-entrepreneurs. Comparing results with the predictions elicited from social science researchers reveals that social scientists overestimate the potential of SMS-based trainings, thus the findings from this study are contrary to priors, and informative.

This study contributes to three strands of literature. First, building on the literature connecting management practices and firm profitability (Bloom et al. (2010, 2013); Bloom and Van Reenen (2010*a*); Bruhn, Karlan and Schoar (2010); McKenzie and Woodruff (2017);

Bruhn, Karlan and Schoar (2018)), I contribute to the large body of evidence on the impact of business trainings on adoption of best practices and business outcomes. Most of the studies in this literature focus on conventional classroom-style trainings that are expensive and hard to scale, with older studies finding little to no effects on sales and profits (Cho and Honorati (2014); Blattman and Ralston (2015); McKenzie and Woodruff (2014)), and more recent work finding positive effects (McKenzie (2020); Chioda et al. (2021)). The evidence on remotely delivered business trainings is still thin and mixed; Davies et al. (2023) find positive short-run effects of Zoom-based trainings for micro-entrepreneurs in Mexico that dissipate within six months of the intervention and Estefan et al. (2023) find significant effects of a mobile app-based training with virtual one-on-one consulting meetings on business outcomes for micro-entrepreneurs in Guatemala. Cole, Joshi and Schoar (2021) find weekly pre-recorded Interactive Voice Response (IVR) messages to be ineffective for improving business outcomes for micro-entrepreneurs. To the best of my knowledge, this paper presents the first rigorous evaluation of an SMS-based business training for micro-entrepreneurs.

Second, this paper adds to an emerging literature on the potential of modifying training content based on insights from psychology to make it easier to internalize. Campos et al. (2017) evaluate a training intervention with psychology-based personal initiative-oriented content in Togo and observe positive effects on business outcomes. Drexler, Fischer and Schoar (2014), and Arráiz, Bhanot and Calero (2019) find encouraging returns from simplifying the training content and focusing on easy to internalize heuristics. The training content used in this study was inspired by these approaches, and this is the first study that tests the effectiveness of a fully automated remote delivery of similarly simplified content.

Third, I contribute to the broader literature on the potential and limitation of information communication technologies for improving socio-economic outcomes in low-income contexts (Spielman et al. (2021); United Nations Conference on Trade and Development (2012); Otis et al. (2024)). This paper highlights lack of engagement as an important limitation of remote delivery of automated information-based support in low-income contexts, pointing towards the need for further research in this direction to fully harness the potential of ICT for development.

The remainder of the paper is organized as follows: Section 1.2 describes the context of the study, Section 1.3 outlines the research design, Section 1.4 discusses the data and timeline of the experiment, Section 1.5 presents the results, and Section 1.6 concludes.

1.2 Context

Home to over 47.6 million people, three-fourths of whom are under the age of 35, Kenya is the largest economy in Eastern and Central Africa (*Kenya Population and Housing Census* (2019)). Similar to other low-income countries, Micro, Small and Medium Enterprises (MSMEs) form the backbone of the economy in Kenya. According to statistics from the nation-wide *Micro, Small and Medium Enterprises Survey* (2016), 7.4 million MSMEs engage over 90% of the active labor force in the country, and contribute just above a third of the GDP. Approximately 55% of these MSMEs are owned by women, and 98% are micro enterprises.[5]

The average education level for micro-entrepreneurs is approximately 11 years, yet adoption of basic best practices for business management is dismally low. About 78.6% of micro-entrepreneurs do not advertise any of their products in any way and 69.8% don't keep any type of business records to keep track of daily sales or expenses. Furthermore, less than 10% of micro-entrepreneurs account for prices of their competitors when setting prices for their own products and services.

These statistics highlight a clear gap in management skills that could be constraining profitability, which can potentially be addressed through business management trainings. However, conventional business trainings are not affordable for micro-entrepreneurs and the scale of externally funded programs is very limited. Due to this, 90% of micro-entrepreneurs surveyed in *Micro, Small and Medium Enterprises Survey* (2016) had never received any type of business training, which highlights the need for an affordable and scalable solution that can cost-effectively extend this support to a wider population.

All these contextual features make Kenya a highly appropriate empirical setting to test the effectiveness of SMS-based business trainings.

1.3 Research Design

This section describes the primary intervention in the study and the randomization design.

1.3.1 The Intervention: SMS-based Business Training

The primary treatment consisted of an SMS-based Business Education training course, which was accessible through smartphones as well as feature-phones, and required no internet access. This course was developed by my primary implementing partner, a Kenyan education technology company that specializes in creation and dissemination of digital training content for audiences including small farmers and micro-entrepreneurs. For this project, I focus on their SMS-based business training course for micro-entrepreneurs. This course was developed in light of existing research on the importance of keeping training content simple in low-capacity contexts (Drexler, Fischer and Schoar (2014); Arráiz, Bhanot and Calero (2019)), and adapted to the local context through extensive qualitative piloting. Information about best practices was conveyed in an easy-to-internalize narrative format describing decision-making of hypothetical micro-entrepreneurs in different scenarios.

Available in English as well as Swahili (the two national languages of Kenya), the training covered practices including marketing, advertising, pricing, record-keeping, and stock management. The content was divided into bite-sized chunks spanning over approximately 150 text messages, and was pushed to users through a chat-bot. The chat-bot was interactive, and users had to keep engaging with it by replying to its messages to keep receiving

[5] Less than 10 employees.

more content. All text messages sent to the chat-bot were completely free, and users were informed about this up front. Figure 1.1 shows what engagement with the content looked like for users.[6]

The content was organized in a fixed sequence and users could go through the sequence at their own pace by only responding to the chat-bot when they wanted additional content. The entire training could be completed in four to six hours if one wanted to do it in one go. Users retained access to all content that they had engaged with up until any point, and could revisit it offline on their phones at will. Those who either did not start engaging with the training content, or started but subsequently abandoned engagement for at least a week, were sent an SMS reminder every week. The weekly reminders were halted if the user engaged at any time and would resume if engagement was abandoned for a week again. The reminders completely stopped after two consecutive months of no engagement.

This training is similar to other light-touch simplified content used for remotely supporting micro-entrepreneurs as well as small-scale agriculturists in low-income settings. Due to this, I expect results from this study to speak to the efficacy of this tool more generally instead of in the specific context of this experiment alone.

1.3.2 Randomization Design

Stratified by gender, the primary sample of 4,701 micro-entrepreneurs who had agreed to participate in the study was randomized at the level of the individual into two groups: (I) Treatment, and (II) Control. The Treatment group of 2,820 micro-entrepreneurs (60% of the study sample) was offered access to the SMS-based business training described in the preceding section.[7] The Control group of 1,881 micro-entrepreneurs (40% of the study sample) received placebo messages designed to remind them about their business without providing any substantive information on best practices. Comparing outcomes across these two groups will allow for the evaluation of the effectiveness of SMS-based business trainings.

1.4 Data and Timeline

There are three main sources of data for the project: (I) back-end data from the SMS platform, (II) Midline survey, and (III) Endline Survey. In addition to these, I use an online elicitation of predictions for treatment effects from social science researchers. In this section, I offer more details about these data sources and the timeline of research activities.

[6] The entire content of the training cannot be provided due to commercial reasons.

[7] The Treatment group was designed to be larger to accommodate outreach requirements from implementing partner.

1.4.1 Main Data Sources

The implementing partner provided the back-end engagement data from the SMS training platform. This data contains information about how engagement levels of each entrepreneur changed over the course of the study period.

The second main data source is the Midline survey conducted three months after the intervention. Approximately 700 randomly selected leads from the primary sample were approached for the phone-based data collection activity, resulting in 307 completed surveys.[8] Response rates in the Treatment and Control groups were 45% and 42%, respectively, with an overall response rate of 43.9%. In addition to demographic information, the Midline data consists of outcomes including measures of knowledge and adoption of best practices, time spent on business and side jobs in the last 30 days, labor hours employed in business in the last 30 days, loans applied for and received in the last 3 months, and business sales, profits and survival in the last 30 days.

The third data source is the Endline survey conducted twelve months after the intervention. The full sample of 4,701 leads was approached for the phone-based data collection activity, resulting in 2,780 completed surveys. The response rate in the treatment, control and overall sample was the same at 59%. This is higher than the response rate in the Midline since more time was spent on calling back leads for which the respondent could not be reached in the first attempt. In addition to the outcomes measured in the Midline survey, the Endline data also consists of knowledge and adoption of more advanced business practices, and sales and profits from all businesses combined in the last 30 days, and time spent on business as well as labor hours employed in the last 7 days. Compared to the Midline, the Endline thus covered more outcomes for a larger sample.

1.4.2 Predictions for Treatment Effects

In order to determine whether the main findings from this study depart from priors held by social science experts, I conducted a survey through the Social Science Predictions Platform (Mehmood (2023)). In this survey, I described the study to social science researchers and elicited their predictions about how key outcomes will be affected twelve months after the SMS training intervention. More specifically, I ask them about their expectations for (i) the extensive margin engagement - i.e. what proportion of those offered the SMS training will have started engaging with it, (ii) the intensive margin engagement - i.e. what proportion of the training content will the average individual in the treatment group will have engaged with, (iii) the effect of treatment assignment on knowledge about best practices, (iv) effect of treatment assignment on adoption of best practices, (v) the effect of treatment assignment on sales from primary business in the last 30 days, and (vi) the effect of treatment assignment on profits from primary business in the last 30 days.

[8] Those who had started engaging with content at the time were slightly over-sampled in the treatment group due to reporting requirements from implementing partner, but all analyses for this paper adjusts the weighting of observations to account for the sampling strategy.

In addition to improving interpretation of results by credibly highlighting whether the findings are unexpected and informative, comparing research findings with expert forecasts contributes to broader efforts aimed at improving accuracy of forecasts in the field, mitigating publication bias, and improving experimental designs in future work; see DellaVigna, Pope and Vivalt (2019) for a more detailed discussion of benefits.

1.4.3 Sample

The primary sample for the intervention came from a list of micro-entrepreneurs maintained by my primary implementation partner in collaboration with a local microfinance institution. This list was compiled through fieldwork they conducted all over Kenya with the aim of collecting contact information of micro-entrepreneurs to target their services to. Subjects were invited to participate in the study over SMS and offered an incentive of KES 100[9]. Those who accepted and signed on to the SMS platform were randomized into groups as detailed in Section 1.3.2. Figure 1.2 shows that the entrepreneurs in the sample were very widely spread out geographically across Kenya, which bolsters the external validity of the findings from this sample. Figure 1.3 shows that an overwhelming majority of businesses in the sample either fall into the category of retail or services.

No baseline survey could be conducted for the study due to timing and logistical constraints, and the only information available for each entrepreneur was their gender. I therefore draw on Midline and Endline data on covariates that are unlikely to change systematically across the randomization groups over the course of the study (e.g. years of education, age etc.), in addition to a limited set of retrospectively framed questions (e.g. did the respondent have a job in December 2021 etc.), for showing pre-intervention summary statistics, balance checks and heterogeneity analyses.

Table A.1.1 in Appendix A.1 presents summary statistics for these pre-intervention covariates for the Midline and Endline samples. Almost half the micro-entrepreneurs in the study sample are women and approximately 45% are based in rural areas. The average micro-entrepreneur has just under 12 years of education, and is between 35 and 36 years old. About 87% had an active business and 40% had an active loan at the time of the intervention deployment.

Table A.1.2 shows that pre-intervention covariates are largely balanced across Treatment and Control groups, for Midline as well as Endline.

1.4.4 Timeline

Below, I summarize the timeline of the experiment implementation and the main data collection activities.

[9] Roughly equal to 1 USD at the time.

Nov 2021 - Dec 2021	**Intervention:** Recruitment into Study and Intervention Deployment
March 2022	**Midline Survey:** Phone-based data collection on knowledge and adoption of best business practices, and business outcomes
Nov 2022 - Dec 2022	**Endline Survey:** Phone-based data collection on wider range of knowledge and adoption of best business practices, and business outcomes
Oct 2023 - Nov 2023	**Social Science Predictions Platform Survey:** Online elicitation of predictions for treatment effects from social science researchers

1.5 Results

This section reports the main results from the study. First, I report the main specifications used for all the analyses. Second, I present results from the Midline survey, detailing treatment effects on engagement, knowledge and adoption of best practices, business performance, and mechanisms. Third, I present results from the Endline survey in the same order. I then move on to comparisons of observed effects with predictions elicited from social scientists.

1.5.1 Specifications

I present results using two main specifications: the first uses OLS Intention-To-Treat (ITT) estimates of treatment assignment, the second uses Local Average Treatment Effect (LATE) estimates where engagement in the training is instrumented by treatment assignment. Both estimation strategies control for gender, which is the stratifying variable. I discuss each specification below.

The first estimation strategy produces ITT estimates of treatment group assignment on the outcomes of interest. The following equation represents the main specification:

$$Y_i = \beta_0 + \beta_1 treatment_i + X_i'\theta + \epsilon_i \tag{1.1}$$

where Y_i is the outcome of interest for individual i, $treatment_i$ is the treatment indicator, X_i is a vector of controls including gender and pre-intervention covariates (if included) with θ representing the associated coefficients, ϵ_i is the error term, and β_1 is the ITT estimate.

The second estimation strategy produces LATE estimates for the effect of treatment on the outcomes of interest. The following equations represent the main specification:

$$eng_i = \gamma_0 + \gamma_1 treatment_i + X_i'\psi + \eta_i \tag{1.2}$$

$$Y_i = \beta_0 + \beta_1 \hat{eng}_i + X_i'\theta + \epsilon_i \tag{1.3}$$

where Y_i is the outcome of interest for individual i, eng_i is a binary indicator for engagement in the training, $treatment_i$ is the treatment indicator (instrument), X_i is a vector of controls including gender and pre-intervention covariates (if included) with ψ and θ representing the associated coefficients, η_i and ϵ_i are error terms, and β_1 is the LATE estimate.

Heterogeneity analysis is based on ITT estimates (as in Equation 1) from the relevant subsamples (e.g. results using respondents of median age and above, and those using respondents of below median age etc.), with the difference in treatment effects across subsamples estimated using the following equation:

$$Y_i = \beta_0 + \beta_1 treatment_i * m_i + \beta_2 treatment_i + \beta_3 m_i + X_i'\theta + [m_i * X_i]'\zeta + \epsilon_i \tag{1.4}$$

where Y_i is the outcome of interest for individual i, $treatment_i$ is the treatment indicator, m_i is the binary covariate of interest for heterogeneity analysis, X_i is a vector of other controls including gender (except for the heterogeneity analyses for gender, where it is accounted for by m_i) and other pre-intervention covariates (if included) with ζ representing the associated coefficients, ϵ is the error term, and β_1 is the estimated difference in treatment effects across the subsamples.

As per my registered pre-analysis plan, I explore heterogeneity of treatment effects along four dimensions: (i) gender, (ii) age, (iii) rural/urban, and (iv) education.

1.5.2 Midline

In this section, I examine effects on key outcomes for the Midline sample, measured three months after the intervention deployment.

1.5.2.1 Engagement at Midline

Table 1.1 shows the treatment effect on engagement with the training content using four different measures of engagement. Column 1 shows the effect of treatment assignment on extensive margin engagement - i.e., a binary indicator for whether or not the individual started engaging with the content. Column 2 shows the effect on whether or not the individual engaged with at least 25% of the training content. Column 3 shows the treatment effect on intensive margin engagement, conditional on starting to engage - i.e. the proportion of training content engaged with given that the individual started engaging. Finally, Column 4 shows the treatment effect on unconditional intensive-margin engagement - i.e., the proportion of training content that the individual engaged with, including engagement of those who never engaged as zero. No controls are added to these regressions.

I find that roughly 30% of the treatment group had engaged with the training three months after the intervention. Overall, only 8.4% of treatment individuals completed at least one-fourth of the training, (Table 1.1 Column 2). Conditional on starting to engage, average percentage of training content completed was 23.4% (Table 1.1, Column 3), and the unconditional average percentage training content completed was 7% (Table 1.1, Column 4).

Taken together, these results suggest that engagement levels three months after the intervention were generally low; most treatment group individuals did not start engaging, and those that did, only covered a quarter of the training content on average. Heterogeneity results by age (above or below median age) show that younger entrepreneurs engaged significantly more with the content (Table A.2.1).

1.5.2.2 Knowledge and Adoption at Midline

Next, I examine whether engaging with the content affected knowledge and adoption of best business practices. Table 1.2 shows the OLS (Columns 1 and 3) and 2SLS (Columns 2, and 4) estimates on means effect indices of knowledge and adoption of best practices, respectively. Coefficients show effects in terms of control group standard deviations (SD). The endogenous variable in Columns (2) and (4) is whether or not the individual engaged with training content. Results show that the training led to a 0.198 SD increase in knowledge of best business practices, which was statistically significant at the 10% level (Table 1.2 Column 1). Conditional on engaging with any of the content (Table 1.2 Column 2) there is a 0.67 SD effect, which is also significant at the 10% level.

Turning to adoption, I see a positive and statistically significant effect of treatment assignment on adoption of best business practices. The OLS regression indicates that assignment to the treatment increases adoption of best business practices by 0.332 SD, with the effect being statistically significant at the 5% level. The LATE estimate is also statistically significant, with an effect of 1.115 SD. I find that this large and statistically significant increase in the adoption index is driven by an increase in advertising (putting up posters/flyers advertising products/services); Table A.2.2 shows that treatment individuals are 8.4 percentage points more likely to put up posters on average compared to those in the control group. Conditional on covering the module on advertising, the effect estimate is 81.2 percentage points and still highly statistically significant (Column 6).

1.5.2.3 Sales, Profits, and Business Survival at Midline

Finally, I examine business sales, profits, and survival at Midline in order to test whether the gain in adoption of best business practices led to meaningful improvements in business outcomes three months after the intervention. Table 1.3 shows the ITT and LATE estimates of the effect of SMS trainings on primary business sales and profits from last 30 days, and business survival. Coefficients in Columns (1) through (4) represent effects in terms of Kenyan Shillings, while those in columns (5) and (6) represent probability of the individual having an active business.

The effect of assignment to treatment was positive for sales (KES 5,721) and profits (KES 1,680.6) in the last 30 days, as well as business survival (4.14 percentage point higher survival) in the overall sample, however these effects are not statistically significant.

I do find larger and statistically significant impacts of trainings on business performance for micro-entrepreneurs of below-median age. Table A.2.3 shows that observable pre-intervention covariates are balanced across treatment and control, for younger as well as older entrepreneurs, which shows that the treatment and control groups within each of these age groups are comparable. Table A.2.4 shows the treatment effects; I find that for younger entrepreneurs, the training increased sales by KES 35,607 (a 109% increase), which is statistically significant at the 5% level. Profits increased by 38% but this increase in borderline insignificant at the 10% level with a p-value of 0.115. Finally, younger entrepreneurs in the treatment group saw a positive and statistically significant increase in business survival of 11.6 percentage points. Business sales and business survival are statistically significantly different across above- and below-median-aged entrepreneurs.

1.5.2.4 Midline Mechanisms

Results thus far show that the treatment induced some engagement with the content, particularly for younger entrepreneurs, and improved knowledge and adoption of best business practices. Sales, profits, and business survival increase overall, but only statistically significantly so for younger (below-median-age) micro-entrepreneurs. I examine three mechanisms to better understand these effects: (1) time spent on business, (2) labor employed in business, and (3) credit outcomes.

Results show that the treatment led business owners to work an additional 28.88 hours on their primary business in the last 30 days - an increase of 16 percent from the control mean of 178.6 hours (Table 1.4). Table A.2.5 shows a similar increase when I consider time spent on all businesses combined to account for cases where the entrepreneur had more than one business. To determine whether this was a result of reallocation of time away from leisure or other income generating activities, I also look at the effect of treatment assignment on time spent on side jobs; Table A.2.6 shows a small negative but statistically insignificant coefficient for the effect on time spent on side jobs, so I don't find evidence for a reallocation of time away from side jobs, suggesting that the additional time spent on business could be resulting from reallocation away from leisure.

Table 1.4 further shows that treatment assignment did not have a statistically significant impact on labor hours employed, loan amount applied for, nor loan amount received.

Table A.2.7 examines time spent on business by age, and finds a statistically significant difference for younger individuals. The training led below median-age individuals to spend 67 more hours on their primary business and 70.9 more hours on all of their businesses combined in the last 30 days, and the differences between above and below median age households is statistically significantly different at the 5% level. There is no statistically significant difference in labor hours employed by age (Table A.2.8), but I find that younger entrepreneurs applied for and received significantly larger loans (Table A.2.9).

Taken together, these results suggest that the increased engagement, time spent on business, and loan amount applied for and received may explain the improvement in business performance experienced by younger micro-entrepreneurs as a result of the SMS training.

1.5.2.5 Midline Summary of Results

To summarize the Midline results, I find that the treatment group engaged with the training content but engagement levels were low. Despite this, the treatment micro-entrepreneurs saw significant improvements in knowledge and adoption of best practices, and large positive but statistically insignificant effects on business sales, profits and survival. Furthermore, I find that younger entrepreneurs see large statistically significant increases in business performance, driven by higher engagement, more time spent on business, and applying for and receiving larger loans.

1.5.3 Endline

In this subsection, I examine longer run effects of the SMS business training using data from the Endline survey, which was conducted twelve months after the intervention deployment, and covered a wider set of outcomes compared to the Midline.

Since the scale of the Midline Survey was considerably smaller than that of the Endline (which targeted the full study sample), there can be a concern that the samples for the two rounds of data collection are systematically different and thus a comparison of treatment effects across Midline and Endline doesn't show how effects changed over time, but rather effects at different time periods for different samples. I argue that this is not likely the case as: (i) Table A.3.1 shows that within the Endline, the sample matched with the Midline[10] is very similar to the sample that is not matched with the Midline, in terms of observable pre-intervention covariates, (ii) the two samples are also very similar in terms of control group outcomes (Table A.3.2), and (iii) I check robustness of observed effects at Endline to restricting the analyses to the Endline sample that was also covered in the Midline, and find no meaningful difference in results.

The following subsections present the observed treatment effects at Endline, paralleling the Midline analyses.

1.5.3.1 Engagement at Endline

This subsection examines the engagement with the training content at Endline, measured twelve months after the launch of the intervention. Mirroring the format of Table 1.1, Column 1 of Table 1.5 shows the effect of treatment assignment on extensive margin engagement, Column 2 shows the effect on whether or not the individual engaged with at least 25% of the training content, Column 3 shows the effect on intensive margin engagement conditional on starting to engage, and Column 4 shows the effect on the unconditional intensive margin engagement.

I find that 28% of the treatment group had engaged with the training twelve months after deployment, with 8.2% covering at least 25% of the content (Column 2). Conditional

[10] Out of 307 entrepreneurs surveyed in the Midline, 227 were surveyed again in the Endline.

on starting to engage, the average engagement was 23.3% (Column 3), and the unconditional average engagement was 6.5% (Column 4).

These results are largely the same as the results at Midline, suggesting that there was little if any engagement beyond the first three months of the intervention.

1.5.3.2 Knowledge and Adoption at Endline

At Endline, in addition to testing knowledge and adoption of basic business practices (covered in Midline), I also test knowledge and adoption of more advanced business practices which were not explicitly covered by the training content. Table 1.6 shows the ITT and LATE estimates for the treatment effect of SMS trainings on means effect indices of basic and advanced knowledge and adoption. Results show no significant improvement in knowledge or adoption of best business practices - both for basic as well as advanced practices - twelve months after the intervention. Running the same analysis but with the Midline matched sample shows the same result of no effects for basic as well as advanced practices (Table A.4.2).

1.5.3.3 Sales, Profits, and Business Survival at Endline

Table 1.7 shows that the training had no effect on sales and profits from primary business in the last 30 days, as well as on business survival twelve months after the intervention. While the coefficient signs are negative, the magnitudes are very small and the p-values are very large. Table A.4.3 shows the results from running the same regressions with the Midline sample; I find the same takeaway of no significant effect of SMS trainings on business performance.

In the Endline, I measure business performance not just for the primary business, but also for all businesses combined, to account for any reallocation across businesses for those who have more than one. Table A.4.4 shows that including other businesses into the equation does not change the results, and Table A.4.5 confirms that the story stays the same when I restrict the analysis to the Midline-matched sample.

In the Midline, younger entrepreneurs saw significant effects on primary business performance, so I check for heterogeneity by age in the Endline as well; I find no effects for younger entrepreneurs (Table A.4.3), and this result does not change when I restrict the analysis to the Midline-matched sample (Table A.4.7). I also check for effects on business performance across younger and older entrepreneurs aggregating sales and profits across all businesses, but I see the same result in the full Endline sample (Table A.4.8), as well as the Midline-matched sample (Table A.4.9).

1.5.3.4 Endline Mechanisms

I examine the same potential mechanisms at Endline as I did at Midline. Table 1.8 indicates that there was a negative and statistically significant decrease in hours worked on primary business for treatment individuals, but the magnitude is small compared to positive effects observed at Midline (9.7 hours less vs. 29 hours more in the last 30 days), with this negative effect primarily driven by rural entrepreneurs. Table A.4.10 shows that the negative effect goes away when I restrict the analysis to the Midline-matched sample. Tables A.4.11 and A.4.12 show that the negative effect on hours worked on primary business is also not robust to aggregating time spent across all businesses, and when focusing on the Midline-matched sample for this analysis too. Taken together, I interpret these results as not showing evidence of a treatment effect on time spent on business twelve months after the intervention. Moreover, Tables A.4.13 and A.4.14 show that there is no significant effect on time spent on side jobs either.

Furthermore, I find that younger entrepreneurs no longer spend more time on business; whether I look at the last 30 days (Tables A.4.15 and A.4.16), or the last 7 days (Tables A.4.17, and A.4.18).

Tables A.4.19, A.4.20, A.4.21, and A.4.22 show that there are no effects of treatment assignment on labor hours employed as well.

Lastly, I observe that the positive treatment effects for younger entrepreneurs on loan amounts applied for and received also disappear (Tables A.4.23 and A.4.24).

Taken together, these results show that there was no meaningful improvement in any of the intermediate business outcomes twelve months after the launch of the training.

1.5.3.5 Endline Summary of Results

In summary, all positive effects observed at Midline were short-lived, and within twelve months of the intervention, there was no difference across treatment and control individuals on average in terms of any outcome of interest.

This lack of effects could potentially be explained by the fact that engagement levels looked very similar at Midline and Endline, suggesting that there wasn't much engagement beyond the three month follow-up. This is largely confirmed by the time trend of aggregate engagement levels measured by the SMS training platform. Figure 1.4 shows the survival curve of engagement of all approximately 30% of the treatment group that started engaging with the content during the course of the study. The figure reveals that all interactions with the SMS training platform ended within the first few months of the intervention, and well before the Endline. Despite the two month long reminder protocol described in Section 1.3, almost no one in the treatment group interacted with the SMS chat-bot after June 2022.

1.5.4 Predictions vs Observations

Are the main results I observe expected and obvious, or are they surprising and informative? Hindsight bias makes it hard to objectively answer this question once the results are revealed. I circumvent this problem by eliciting predictions for the Endline treatment effects from social science researchers without informing them about my findings, as detailed in Section 1.4.2. In this section, I present comparisons of predicted and observed treatment effects to argue that the findings in this study are indeed surprising and informative.

The predictions survey received 70 responses, with half of the sample consisting of PhD student researchers, and the other half consisting of researchers with a PhD who are at more advanced stages in their careers (academic faculty, post-docs, and researchers at think tanks and policy organizations). About 89% of the respondents listed Economics as one of their main disciplines, and 7.7% listed Political Science. Other social science disciplines represented, but in significantly smaller numbers, included Psychology and Sociology.

Figures 1.5, 1.6, and 1.7 illustrate how the predicted treatment effects for Endline compare with observed effects at Midline and Endline. In all three figures, the red circle shows the mean of the distribution of predictions for the treatment effect, while the red rectangle represents the inter-quartile range, with the line inside the rectangle representing the median. Observed treatment effects at Midline and Endline are represented by a gray rhombus and a black triangle, respectively, with error bars showing 90% confidence intervals.

Figure 1.5 shows how extensive and intensive margin engagement at Endline compares with observed levels. Respondents predicted that about 50% of the treatment group will have started engaging with the training content by Endline, and the average micro-entrepreneur in the treatment group will have covered approximately 40% of the training content. The actual engagement levels observed are much lower; only about 30% of the treatment group had started engaging by Midline, and the extensive margin engagement stood at the same level by the Endline. Additionally, the average micro-entrepreneur in the treatment group only covered approximately 7% of the training content by Midline, and this unconditional intensive margin engagement level stayed largely the same by the Endline.

Figure 1.6 shows respondent expectations for the effect of assignment to treatment on knowledge and adoption of best practices in terms of control group standard deviations. Respondents predicted that knowledge and adoption will increase by approximately 0.3 and 0.2 standard deviations, respectively, at the Endline. While I observe 0.2 and .33 standard deviations increases in knowledge and adoption at Midline, respectively, these positive effects disappear by the Endline. Hence, I find that respondents also overestimated the effect of treatment assignment on knowledge and adoption of best practices twelve months after the intervention.

Figure 1.7 shows a similar story for effects on sales and profits. Respondents predicted a 13% and 12% increase in sales and profits in the Endline, respectively. While observed effects in Midline are of similar magnitudes as the predictions, albeit statistically insignificant, the observed effects are close to zero for the Endline. I therefore find that respondents also overestimated the effect of treatment assignment on sales and profits twelve months after the intervention.

Lastly, while I find no meaningful difference between predictions given by PhD student researchers and more advanced PhD researchers, I observe that those that are more confident about their predictions overestimate treatment effects the most - Figure A.5.1 illustrates the correlation between predictions for treatment effects and the reported confidence in predictions reported by respondents.

In light of these results, I conclude that social science researchers overestimate the potential of SMS-based trainings to improve outcomes for micro-entrepreneurs, and the findings from this study are thus contrary to priors. Updating these priors is important as investment of resources into such remote information-based support programs by policy makers and practitioners are often informed by beliefs about impacts held by social scientists.

1.6 Conclusion and Policy Implications

This paper assesses the potential of SMS-based business trainings to improve business outcomes for micro-entrepreneurs via a field experiment in Kenya, and compares the findings with priors held by social science researchers. About 4,700 micro-entrepreneurs recruited over SMS were randomized into a Treatment and a Control group; the 2,820 entrepreneurs in the Treatment group were provided access to SMS-based training content, while the remaining 1,880 were sent placebo messages aimed at reminding them about their business.

Three months after the intervention, I find positive effects on knowledge and adoption of best practices, particularly for advertising, and large positive but statistically insignificant effects on business sales, profits and survival. I further find significant positive effects on business performance for younger entrepreneurs, driven by higher engagement with the training content, and larger increases in time spent on business, and amount of credit applied for and received.

However, these positive results dissipate in the longer run; twelve months after the intervention, I see no effects on knowledge and adoption of best practices, as well as business sales, profits and survival. Additionally, the positive effects on business outcomes observed for younger entrepreneurs after three months also disappear within twelve months. The time-trend of engagement further reveals that the lack of longer-run effects is likely driven by micro-entrepreneurs abandoning all interactions with the content within the first few months of the intervention.

I therefore conclude that, despite their growing popularity, SMS-based trainings on their own are unlikely to be effective for micro-entrepreneurs. Comparing results with elicited priors of social science researchers reveals that social scientists overestimate the potential of SMS-based trainings, thus the findings from this study are surprising and informative.

The takeaways from this study direct attention towards an interesting question that can be critical for maximizing impacts of remotely provided information-based support, and should be an important research direction going forward; why was engagement low? There can be two possible non-mutually exclusive components to blame for this - (i) the content, and/or (ii) the content delivery.

If entrepreneurs judge the content to not be worth engaging with, they are unlikely to invest the required time to finish it. It is unclear whether this factor played a major role in this study as 70% of the treatment group never started engaging with the content, and were thus not exposed to the training. Nevertheless, better content is an important research direction, with a growing body of evidence on the benefits of simplifying the message (Drexler, Fischer and Schoar (2014); Arráiz, Bhanot and Calero (2019)), adding psychology-based personal initiative oriented content (Campos et al. (2017)), and also customizing the content in light of the needs of the recipients (Fabregas et al. (2022)), and further research in this direction can help move the needle more on addressing the challenge of low engagement in remotely provided trainings.

The other reason for low engagement can be the nature of the content delivery. This component perhaps speaks to why, despite having access to basic as well as advanced knowledge about countless subjects through the internet[11], we still have to attend school and other classroom-type environments for acquiring knowledge and skills beyond foundational language and mathematics. Indeed, qualitative interviews with a small subset of individuals in the treatment group reveal that the reason micro-entrepreneurs did not engage with the content was that they were ”busy during the day with customers” and then ”forgot to engage after getting free”, suggesting behavioral constraints might be playing an important role in limiting engagement. I leave a deeper dive into possible behavioral drivers in this context to future work. Further research in this direction, possibly guided by insights from psychology and behavioral economics (as reflected in the likes of Della Vigna and Malmendier (2006); Bai et al. (2021); de Oliveira (2023)) can add to our understanding of what makes people engage with remotely provided content, which can greatly help unlock the full potential of digital technologies for remote learning.

[11] Through resources including Youtube, Khan Academy, Udemy, Coursera, Lynda, Skillshare, Udacity etc.

1.7 Main Tables and Figures

1.7.1 Main Tables

1.7.1.1 Midline

Table 1.1: Midline: Engagement

	(1)	(2)	(3)	(4)
	Engaged	Covered $\geq$ 25%	% Covered (cond.)	% Covered (uncond.)
Training	0.298***	0.0844***	0.234***	0.0698***
	(0.0306)	(0.0150)	(0.0229)	(0.00988)
Control Mean	0	0	0	0
Observations	307	307	229	307

Notes: This table shows the output from OLS regressions of four measures of engagement on treatment assignment at Midline, with no controls added. Column (1) shows effect of treatment assignment on extensive margin engagement - i.e. whether or not the individual started engaging with content. Column (2) shows the effect on whether or not the individual engaged with at least 25% of the training content. Column (3) shows the effect on percentage of training content that the individual engaged with conditional on starting to engage. The observations used for this regression exclude the 78 individuals in the treatment group who had not started engaging with the content, thus the number of observations is 307 - 78 = 229. Column (4) shows the unconditional effect on percentage of training content that the individual engaged with. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 1.2: Midline: Knowledge and Adoption of Best Practices

	Knowledge		**Adoption**	
	OLS	IV	OLS	IV
Training	.198*		.332**	
	(.118)		(.155)	
Engaged		.673*		1.115**
		(.404)		(.535)
Female	.0402	.00138	-.175	-.243
	(.117)	(.123)	(.183)	(.208)
P-value	.0953	.0957	.0330	.0371
Control Mean	0	0	0	0
Observations	307	307	297	297

Notes: This table shows the intent-to-treat and local average treatment effect estimates of SMS trainings on means effect indices of knowledge and adoption of best practices at Midline. Coefficients represent effects in terms of control group standard deviations. Columns (1) and (3) show output from OLS regressions, and columns (2) and (4) show output from 2SLS regressions where the endogenous variable is whether or not the individual engaged with training content. Robust standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table 1.3: Midline: Sales, Profits and Survival

	Sales		**Profits**		**Survival**	
	OLS	IV	OLS	IV	OLS	IV
Training	5721.0		1680.6		.0414	
	(10814.3)		(1601.9)		(.0326)	
Engaged		19112.7		5407.1		.141
		(35918.8)		(5134.1)		(.111)
Female	-34684.4***	-35888.4***	-7567.2***	-7926.3***	-.0350	-.0431
	(11785.7)	(12228.7)	(1674.6)	(1837.3)	(.0299)	(.0311)
P-value	.597	.595	.295	.292	.204	.207
Control Mean	47581.2	47581.2	10886.9	10886.9	.908	.908
Observations	290	290	294	294	307	307

Notes: This table shows the intent-to-treat and local average treatment effect estimates of SMS trainings on primary business sales and profits from last 30 days, and business survival at Midline. Coefficients in columns (1) thought (4) represent effects in terms of Kenyan Shillings, while those in columns (5) and (6) represent probability of individual having an active business. Columns (1), (3) and (5) show output from OLS regressions, and columns (2), (4) and (6) show output from 2SLS regressions where the endogenous variable is whether or not the individual engaged with training content. Robust standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table 1.4: Midline: Intermediate Outcomes

	Hrs. worked		**Lab. Hrs. employed**		**Loan Applied**		**Loan Received**	
	OLS	IV	OLS	IV	OLS	IV	OLS	IV
Training	28.88*		4.933		-365.6		987.1	
	(16.52)		(29.64)		(7629.7)		(5570.2)	
Engaged		108.4*		16.74		-1240.6		3349.5
		(63.22)		(100.0)		(25767.3)		(18792.9)
Female	-5.896	-12.26	-81.69***	-82.66***	-14817.1**	-14745.6**	-11173.2**	-11366.5**
	(16.94)	(18.25)	(26.02)	(25.56)	(6344.6)	(6715.8)	(5249.0)	(5669.1)
P-value	.0817	.0864	.868	.867	.962	.962	.859	.859
Control Mean	178.6	178.6	122.6	122.6	13818.3	13818.3	10104.6	10104.6
Observations	269	269	307	307	307	307	307	307

Notes: This table shows the intent-to-treat and local average treatment effect estimates of SMS trainings on hours worked and labor hours employed in the primary business in the last 30 days, and loan amounts applied for and received (in Kenyan Shillings) in the last 3 months at Midline. Columns (1), (3), (5), and (7) show output from OLS regressions, and columns (2), (4), (6), and (8) show output from 2SLS regressions where the endogenous variable is whether or not the individual engaged with training content. Robust standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

1.7.1.2 Endline

Table 1.5: Endline: Engagement

	(1)	(2)	(3)	(4)
	Engaged	Covered $\geq$ 25%	% Covered (cond.)	% Covered (uncond.)
Training	0.280***	0.0821***	0.233***	0.0651***
	(0.0110)	(0.00672)	(0.0115)	(0.00411)
Control Mean	0	0	0	0
Observations	2780	2780	1578	2780

Notes: This table shows the output from OLS regressions of four measures of engagement on treatment assignment at Endline, with no controls added. Column (1) shows effect of treatment assignment on extensive margin engagement - i.e. whether or not the individual started engaging with content. Column (2) shows the effect on whether or not the individual engaged with at least 25% of the training content. Column (3) shows the effect on percentage of training content that the individual engaged with conditional on starting to engage. Column (4) shows the unconditional effect on percentage of training content that the individual engaged with. Standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table 1.6: Endline: Knowledge and Adoption of Best Practices

	Basic Knowledge		**Basic Adoption**		**Advanced Knowledge**		**Advanced Adoption**	
	OLS	IV	OLS	IV	OLS	IV	OLS	IV
Training	.0248		-.0638		-.0355		-.0223	
	(.0384)		(.0399)		(.0387)		(.0410)	
Engaged		.0887		-.222		-.127		-.0777
		(.137)		(.139)		(.138)		(.143)
Female	.0127	.0113	-.287***	-.284***	-.0833**	-.0814**	-.0491	-.0479
	(.0376)	(.0376)	(.0390)	(.0391)	(.0382)	(.0383)	(.0405)	(.0405)
P-value	.518	.518	.110	.110	.359	.359	.586	.586
Control Mean	0	0	0	0	0	0	0	0
Observations	2780	2780	2563	2563	2780	2780	2563	2563

Notes: This table shows the intent-to-treat and local average treatment effect estimates of SMS trainings on means effect indices of basic and advanced knowledge and adoption of best practices at Endline. Basic knowledge and basic adoption indices are similar to the knowledge and adoption indices analysed for the Midline, while the advanced knowledge and adoption indices are based on best practices are a bit more advanced and not necessarily directly mentioned in the SMS-trainings. Coefficients represent effects in terms of control group standard deviations. Columns (1), (3), (5), and (7) show output from OLS regressions, and columns (2), (4), (6), and (8) show output from 2SLS regressions where the endogenous variable is whether or not the individual engaged with training content. Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 1.7: Endline: Sales, Profits and Survival

	Sales		**Profits**		**Survival**	
	OLS	IV	OLS	IV	OLS	IV
Training	-2206.5		-220.6		-.0159	
	(3534.1)		(1009.0)		(.0112)	
Engaged		-7891.3		-789.6		-.0568
		(12640.7)		(3610.8)		(.0400)
Female	-34537.5***	-34415.8***	-9444.3***	-9432.7***	.00903	.00990
	(3380.7)	(3389.6)	(982.3)	(982.6)	(.0111)	(.0111)
P-value	.532	.532	.827	.827	.154	.155
Control Mean	59356.0	59356.0	19453.4	19453.4	.915	.915
Observations	2772	2772	2770	2770	2779	2779

Notes: This table shows the intent-to-treat and local average treatment effect estimates of SMS trainings on primary business sales and profits from last 30 days, and business survival at Endline. Coefficients in columns (1) thought (4) represent effects in terms of Kenyan Shillings, while those in columns (5) and (6) represent probability of individual having an active business. Columns (1), (3) and (5) show output from OLS regressions, and columns (2), (4) and (6) show output from 2SLS regressions where the endogenous variable is whether or not the individual engaged with training content. Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

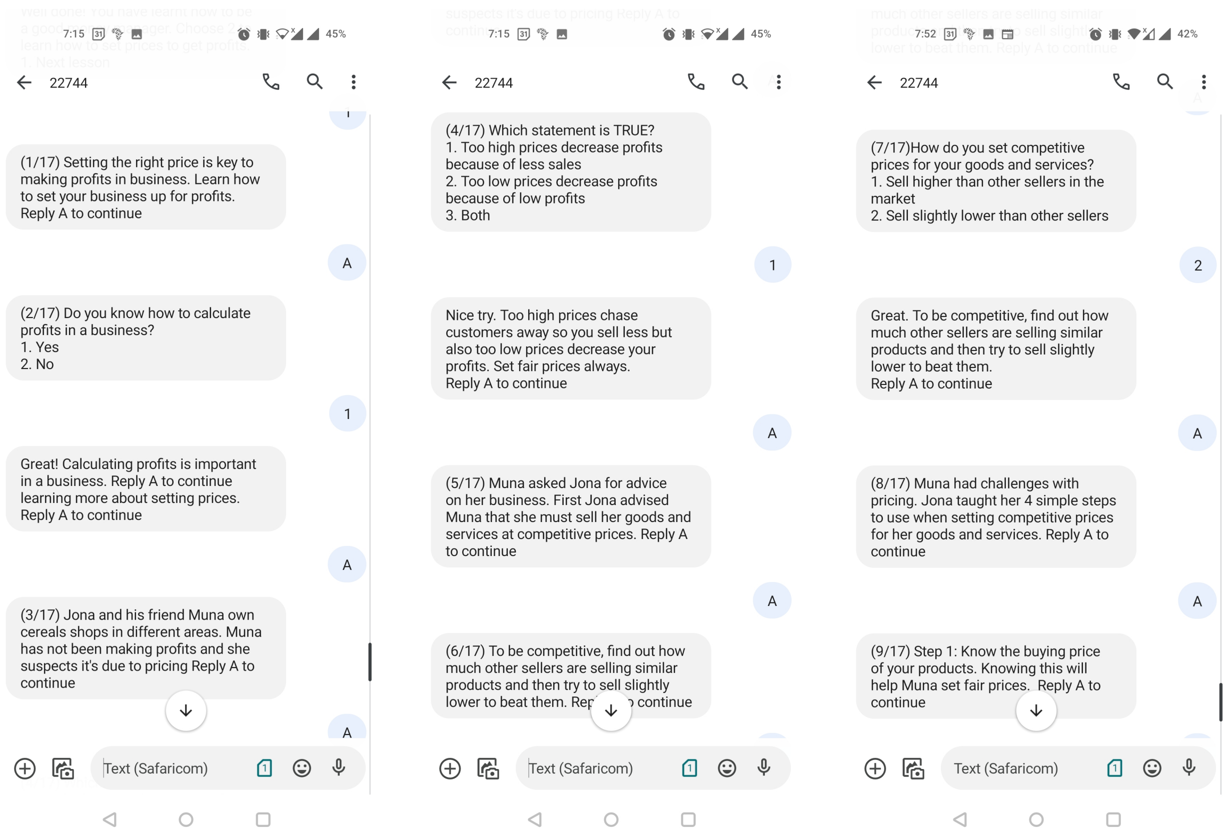
Table 1.8: Endline: Intermediate Outcomes

	Hrs. worked		**Lab. Hrs. employed**		**Loan Applied**		**Loan Received**	
	OLS	IV	OLS	IV	OLS	IV	OLS	IV
Training	-9.702**		-2.085		-667.9		-397.2	
	(4.459)		(6.554)		(2321.2)		(2071.2)	
Engaged		-34.67**		-7.438		-2384.6		-1418.2
		(15.99)		(23.38)		(8282.3)		(7390.0)
Female	-2.011	-1.466	-68.08***	-67.97***	-7467.3***	-7431.5***	-5751.0***	-5729.7***
	(4.343)	(4.364)	(6.368)	(6.415)	(2220.8)	(2217.8)	(1977.4)	(1973.8)
P-value	.0296	.0302	.750	.750	.774	.773	.848	.848
Control Mean	215.6	215.6	91.26	91.26	20392.3	20392.3	16813.0	16813.0
Observations	2777	2777	2778	2778	2780	2780	2780	2780

Notes: This table shows the intent-to-treat and local average treatment effect estimates of SMS trainings on hours worked and labor hours employed in the primary business in the last 30 days, and loan amounts applied for and received (in Kenyan Shillings) in the last 3 months at Endline. Columns (1), (3), (5), and (7) show output from OLS regressions, and columns (2), (4), (6), and (8) show output from 2SLS regressions where the endogenous variable is whether or not the individual engaged with training content. Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

1.7.2 Main Figures

Figure 1.1: SMS Business Training Content



(1/17) Setting the right price is key to making profits in business. Learn how to set your business up for profits. Reply A to continue

A

(2/17) Do you know how to calculate profits in a business?
1. Yes
2. No

1

Great! Calculating profits is important in a business. Reply A to continue learning more about setting prices. Reply A to continue

A

(3/17) Jona and his friend Muna own cereals shops in different areas. Muna has not been making profits and she suspects it's due to pricing Reply A to continue

A

(4/17) Which statement is TRUE?
1. Too high prices decrease profits because of less sales
2. Too low prices decrease profits because of low profits
3. Both

1

Nice try. Too high prices chase customers away so you sell less but also too low prices decrease your profits. Set fair prices always.
Reply A to continue

A

(5/17) Muna asked Jona for advice on her business. First Jona advised Muna that she must sell her goods and services at competitive prices. Reply A to continue

A

(6/17) To be competitive, find out how much other sellers are selling similar products and then try to sell slightly lower to beat them. Rep[…] continue

(7/17)How do you set competitive prices for your goods and services?
1. Sell higher than other sellers in the market
2. Sell slightly lower than other sellers

2

Great. To be competitive, find out how much other sellers are selling similar products and then try to sell slightly lower to beat them.
Reply A to continue

A

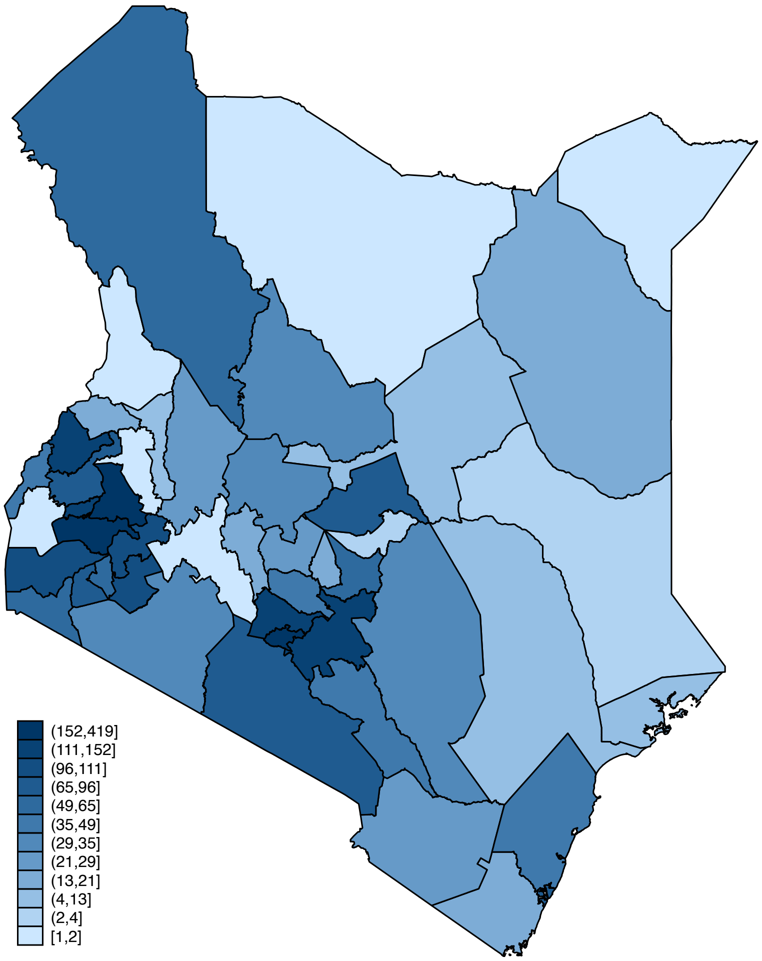
(8/17) Muna had challenges with pricing. Jona taught her 4 simple steps to use when setting competitive prices for her goods and services. Reply A to continue

A

(9/17) Step 1: Know the buying price of your products. Knowing this will help Muna set fair prices. Reply A to continue

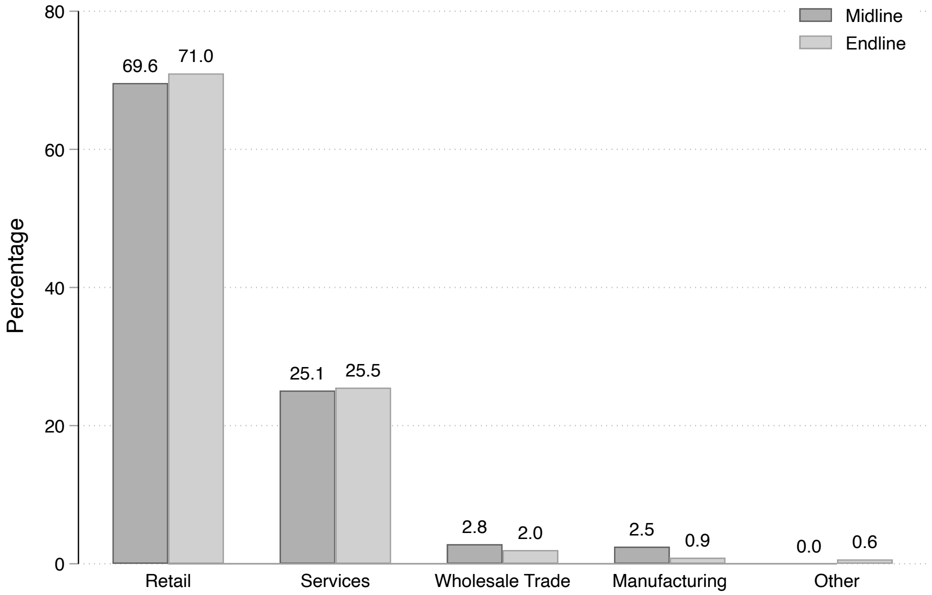
Notes: This figure shows screenshots of interactions with the SMS-based chatbot as it pushes out content to users. In this context, most micro-entrepreneurs set prices just based on their buying costs, without accounting for prices of their competitors, so the content pushes them to change their pricing strategy.

Figure 1.2: Geographical Spread of Study Sample



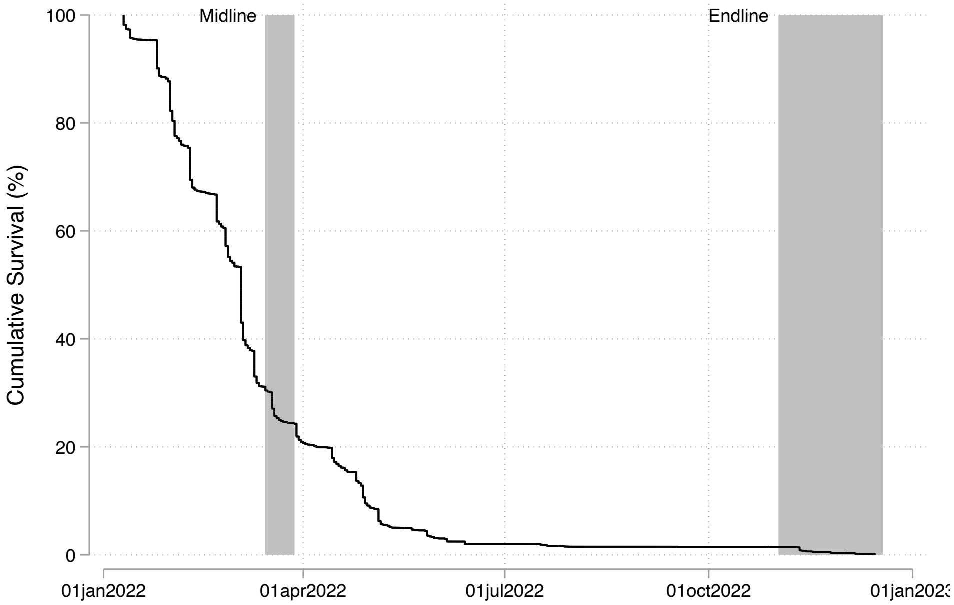
Notes: This figure shows the geographical distribution of micro-entrepreneurs in the study sample. The figure legend in the bottom-left assigns color-coding to number of micro-entrepreneurs based in each of the 47 counties of Kenya.

Figure 1.3: Nature of Businesses



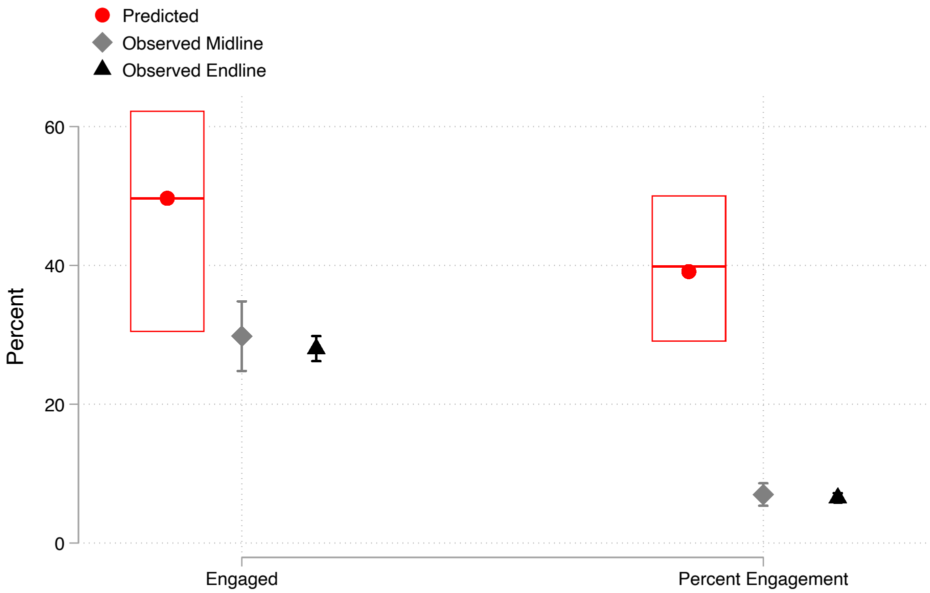
Notes: This figure shows the composition of the study samples across Midline and Endline in terms of nature of business of micro-entrepreneurs.

Figure 1.4: Engagement Survival Curve



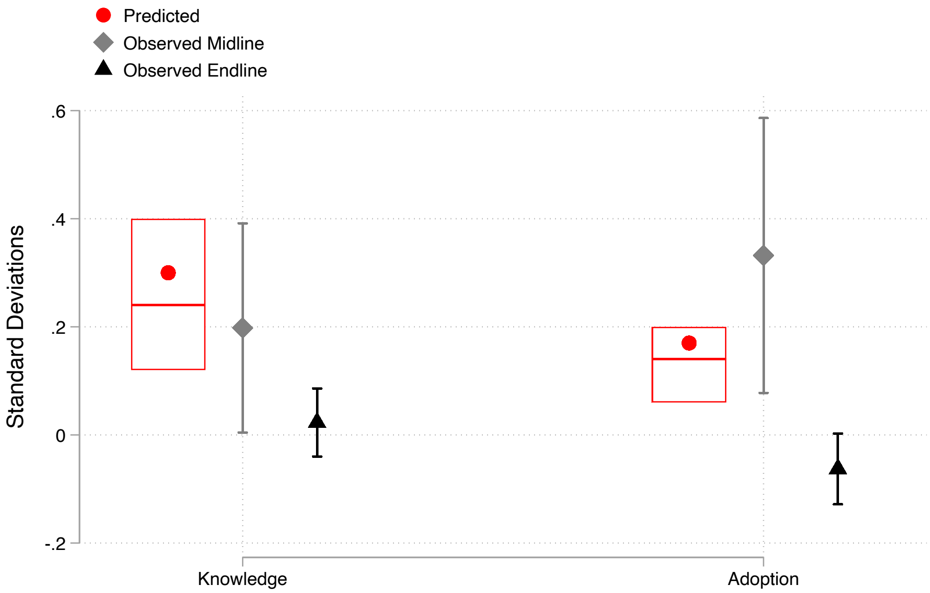
Notes: This figure illustrates how interactions with the SMS platform were distributed throughout the study period. The plot shows reverse cumulative engagement over time; for example, it shows that 80% of all the interactions with the chat-bot throughout the course of the study, had ended by 4/1/2022. The shaded areas represent the time-spans during which the Midline and Endline surveys were being conducted.

Figure 1.5: Predictions vs. Observations - Engagement



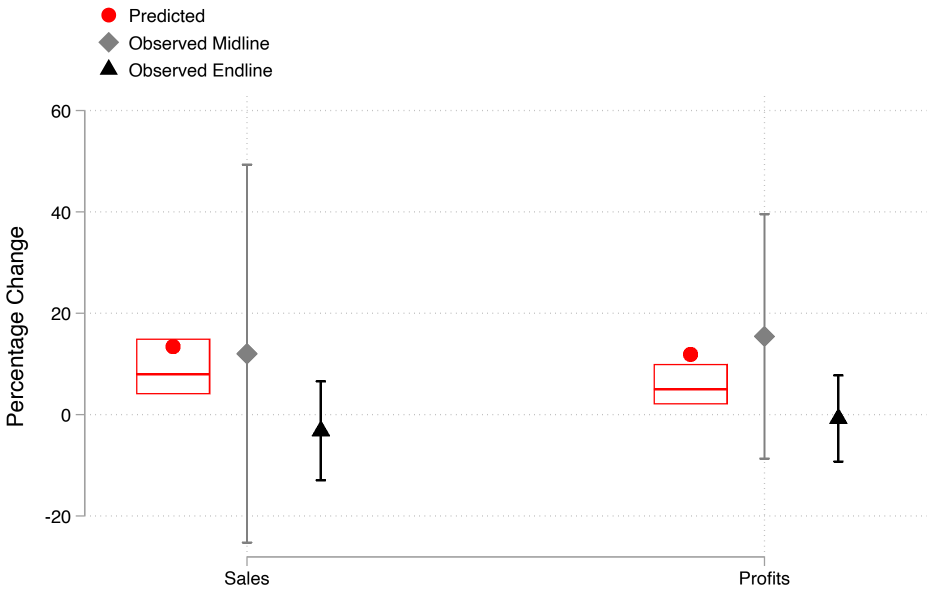
Notes: This figure shows how predicted treatment effects on extensive and intensive margin engagement for the Endline compare with observed Midline and Endline effects. Effects are in terms of percentage points. The distribution of the predicted treatment effect is illustrated in red, with the mean represented by the circle, and the rectangle representing the inter-quartile range, with the line inside the rectangle indicating the median. Observed Midline and Endline treatment effects are represented by a gray rhombus and a black triangle, respectively, with error bars representing 90% confidence intervals.

Figure 1.6: Predictions vs. Observations - Knowledge and Adoption



Notes: This figure shows how predicted treatment effects on knowledge and adoption of best practices for the Endline compare with observed Midline and Endline effects. Effects are in terms of control group standard deviations. The distribution of the predicted treatment effect is illustrated in red, with the mean represented by the circle, and the rectangle representing the inter-quartile range, with the line inside the rectangle indicating the median. Observed Midline and Endline treatment effects are represented by a gray rhombus and a black triangle, respectively, with error bars representing 90% confidence intervals.

Figure 1.7: Predictions vs. Observations - Sales and Profits



Notes: This figure shows how predicted treatment effects on business sales and profits in the last 30 days for the Endline compare with observed Midline and Endline effects. Effects are in terms of percentage changes. The distribution of the predicted treatment effect is illustrated in red, with the mean represented by the circle, and the rectangle representing the inter-quartile range, with the line inside the rectangle indicating the median. Observed Midline and Endline treatment effects are represented by a gray rhombus and a black triangle, respectively, with error bars representing 90% confidence intervals.

Chapter 2

Demand for SMS-Based Business Trainings Amongst Kenyan Micro-Entrepreneurs

2.1 Introduction

Most conventional in-person business training programs geared towards small businesses are unaffordable for small business owners in low-income countries. For example, the average Kenyan micro-enterprise generates about USD 440[1] in a month (*Micro, Small and Medium Enterprises Survey* (2016)), while the average per trainee cost of business trainings in Kenya[2] can range from USD 125 to USD 900 (Mehtha (2017)). Due to this, business training programs are typically provided to entrepreneurs free of cost, at substantial expense to government or donor organizations; some estimates suggest that over $1 billion is spent annually to train 4-5 million entrepreneurs in low-income countries globally (McKenzie (2020)). The high cost and model of delivery substantially limits the scale at which trainings can be provided to entrepreneurs and also gives rise to sustainability concerns in the long run.

Phone-based business trainings can be significantly lower-cost, and micro-entrepreneurs could theoretically afford to pay for them out of pocket. A market-based approach would be more sustainable and have stronger incentives for continuous improvement of service in the direction of local needs. However, there is little evidence on whether and how much small business owners in low-income settings would want to pay out of pocket for these trainings.

This paper seeks to address this gap by studying the demand for SMS-based business trainings amongst micro-entrepreneurs in Kenya. Leveraging key components of the field experiment studied in Chapter 1, I measure demand for trainings via two methods; First,

[1] Roughly equal to KES 48,000 at the time of the study.

[2] E.g., International Labour Organization (2024), which is a well known training implemented in Kenya and several other countries.

I examine buying decisions of 415 micro-entrepreneurs who were sent Take-It-Or-Leave-It (TIOLI) offers for an SMS business training via text messages, where the price was randomized across respondents over three levels; (i) free, (ii) half the marginal cost to service provider (KES 5), and (iii) double the marginal cost (KES 20). Second, I analyze data on willingness to pay for SMS business trainings collected in-person from 103 micro-entrepreneurs based in Nairobi, Kenya, via a modified version of the demand elicitation exercise pioneered by Becker, Degroot and Marschak (1964).

I find that micro-entrepreneurs are willing to pay a positive price for SMS business trainings, and the demand decreases with price. In the TIOLI sample, 70% of individuals chose to accept the additional training when it was offered for free, 68% accepted when the price was KES 5 (half the marginal cost to service provider), and about 50% accepted when the price was KES 20 (double the marginal cost). In the BDM sample, the average willingness to pay for SMS trainings was KES 50 (five times the marginal cost), and almost a quarter of the respondents were willing to buy the training for KES 100 (ten times the marginal cost). Data from both methods of measurement of demand thus reveal that micro-entrepreneurs were willing to pay a positive amount for SMS-based business management trainings, and a substantial proportion of them were willing to pay much more than the marginal cost to service providers. This suggests that a market-based approach for providing SMS business trainings might be feasible in this context.

Additionally, I find correlational evidence for the demand for trainings being higher amongst individuals with more children in the household, those that recently applied for a loan, those with more knowledge of best practices, and those with higher education levels. This is in line with the intuition that those who have more dependents, and those in need of funds recently are more likely to take on a potential opportunity to increase their business profits. Also, those with more knowledge about best practices and those who are more educated recognize the importance of information about best practices, and are more likely to want to learn more.

This study adds to the extremely limited work on demand for business trainings amongst entrepreneurs. Maffioli, McKenzie and Ubfal (2020) estimate the demand for a business training in Jamaica, however, the program they study is a conventional in-person training program rather than a remotely delivered phone-based training. Cole and Fernando (2020) estimate the willingness to pay for voice-based ICT advisory services in India, but those services are geared towards farmers. This study is the first to provide empirical evidence on demand for SMS-based business trainings amongst micro-entrepreneurs.

The remainder of the paper is organized as follows: Section 2.2 describes the context of the study, Section 2.3 outlines the demand elicitation methods employed, Section 2.4 describes the data sources, sample and timeline of the study activities, Section 2.5 presents the results, and Section 2.6 concludes.

2.2 Context

The study is based in Kenya - a country whose economy relies heavily on its small businesses,[3] and one in which adoption of best business management practices remains low amongst micro-entrepreneurs.

I study the demand for SMS-based business trainings through the same field experiment as that studied in Chapter 1 of this dissertation. The primary intervention in this experiment was an SMS-based business management training course developed by a Kenyan education technology company. Available in English as well as Swahili, the training conveyed information about best business management practices in an easy-to-internalize narrative format, describing decision-making of hypothetical micro-entrepreneurs in different scenarios. The content was organized in a fixed sequence of about 150 bite-sized chunks covering topics including marketing, advertising, pricing, record-keeping, and stock management. Those offered access could engage with the content at their own pace through an automated chat-bot using text messages for free. Treatment effects on knowledge and adoption of best practices, and business outcomes were measured three months (Midline) and twelve months (Endline) after the intervention.

Stratified by gender, the training was randomized across 4,701 micro-entrepreneurs, with 60% of the study sample being offered access to the SMS training, and the remaining 40% receiving placebo content designed to remind them about their business without providing any substantive information on best practices.

The following section describes how I leverage this setup to study the demand for SMS-based business trainings.

2.3 Research Design

I study willingness to pay for SMS-based trainings using two methods: (i) Randomized Take-It-Or-Leave-It (TIOLI) offers in the Treatment group, and (ii) In-person elicitation of maximum willingness to pay, adapting the method pioneered by Becker, Degroot and Marschak (1964) (BDM) for a subset of the overall sample. Using two different methods with different samples allows for corroboration of observed trends across the exercises, and thus lends more credibility to findings.

2.3.1 Take It Or Leave It Offers

The 2,820 micro-entrepreneurs in the Treatment group that were offered the SMS-based training were later offered access to a second SMS-based business training at a randomly selected price. The offer was sent over SMS once the users completed the first training, or if

[3] As noted in Chapter 1, MSMEs - of which micro-enterprises constitute 98% - engage over 90% of the active labor force in Kenya, and contribute over a third of the GDP.

they stopped engaging for at least two uninterrupted months. The elicitation was separately incentivised; individuals were promised additional airtime that was to be disbursed after they responded with their decision. If they chose to buy, the price of the training was deducted from the airtime value and the remaining airtime was disbursed. If they chose to not buy, the entire airtime was disbursed after they responded with their decision. The incentive was aimed at nudging people who did not want to buy to actually report their decision instead of just not responding to the invitation to choose, so as to not overestimate the demand.

Stratified by gender, the price in these TIOLI offers was randomized across individuals over three levels; (i) 1,419 individuals (50% of Treatment group) were offered the second training for free, (ii) 697 individuals (25% of Treatment group) were offered a price of KES 5, and (iii) 704 individuals (25% of Treatment group) were offered a price of KES 20. The marginal cost of provision of the entire training incurred by the implementing partner was KES 10 per person, thus the two positive price levels in the TIOLI design represented half and double the marginal cost of provision, respectively.

Observing buying decisions across these pricing arms allows me to confirm whether there is any positive willingness to pay for the trainings amongst entrepreneurs, and if it varies systematically with price.

2.3.2 Becker-DeGroot-Marschak Elicitation

Following the Endline Survey, I randomly select about 100 Nairobi-based business owners from my primary sample to conduct an in-person elicitation of willingness to pay for an additional SMS-based business training. Following Maffioli, McKenzie and Ubfal (2020), I used a modified version of the method proposed by Becker, Degroot and Marschak (1964), that uses a multiple-price list approach. The possible price level options were framed as the resulting prices from a lottery for the amount of discount offered to respondents.

Respondents were asked if they would buy the SMS training at a sequence of prices starting with zero and increasing in increments of KES 10, until the respondent switched their response from “Yes” to “No”. The respondents were then asked to quote their maximum willingness to pay between the price they rejected and the last price they accepted. After confirming if the respondent was sure about their response and that they would not be able to back out of their commitment to buy once the discount lottery was run, the enumerators ran the discount lottery and revealed the final price. The final incentive amount was disbursed via mobile money at the end of the interview, and, where applicable, an invitation to the additional training was sent to the respondents over SMS soon after.

Before the elicitation, respondents were provided a brief overview of the new SMS training content, and a detailed explanation about the elicitation method with hypothetical examples, highlighting that it was in the respondent’s interest to not commit to buying at a price that was lower or higher than their actual maximum willingness to pay for the SMS training.

To circumvent complications posed by the possibility of individuals reneging on their commitment to buy at any price drawn from the lottery which is less than or equal to the maximum willingness to pay they reported during the exercise,[4] respondents were informed up front that the payment for their potential purchase of the training would be taken out of the participation incentive amount committed to them before the start of the interview.

2.4 Data and Timeline

This section outlines the main data sources used for the study, describes the sample, and summarizes the timeline of activities.

2.4.1 Data Sources

There are three main sources of data for this study: (i) back-end data from the SMS platform, (ii) Endline survey, and (iii) the in-person BDM elicitation activity.

The implementing partner provided the back-end engagement data from the SMS training platform. This data contains information about engagement with the SMS trainings and buying decisions for the TIOLI demand elicitation.

Second, the Endline survey was conducted twelve months after the SMS training intervention, marking the end of the TIOLI elicitation window. The survey collected data on knowledge and adoption of best business management practices, and business outcomes, and was administered via phone calls. A total of 415 individuals responded to the TIOLI invitations, and 380 (91.6%) of these respondents were also covered in the Endline survey. Therefore, while analyses showing raw buying decisions will be based on data from all 415 respondents, all analyses linking buying decisions with other respondent characteristics will be based on data from the 380 respondents that overlap across the two samples.

Lastly, the in-person BDM-style demand elicitation was conducted at the end of the Endline survey, with 103 Nairobi-based business owners. Unlike the TIOLI offers, the sample for this exercise included treatment as well as control individuals. This will allow me to estimate if there is any effect of treatment assignment on willingness to pay for SMS trainings.

2.4.2 Sample

Table B.1.1 in Appendix B.1 presents summary statistics for the TIOLI and BDM samples. The samples appear more or less similar in terms of 'baseline covariates'[5]. The proportion of female entrepreneurs is similar across the samples at 44% and 53%, respectively. The average education level is the same at 12 years of schooling, while the average age is about 34 and 36 years, respectively. Number of adults (2.7 vs. 2.5) and children (2.1 vs 2.0) in the household are also similar. One key difference is that while almost half the TIOLI sample is based in rural areas, only 11% of the BDM sample is based in rural settings. This is expected

[4] As faced by Maffioli, McKenzie and Ubfal (2020) in their study.

[5] Term used in terms of the original field experiment in Chapter 1.

as the scope of the BDM elicitation was restricted to business owners in the city of Nairobi due to logistical constraints.

The samples are also largely comparable in terms of 'outcomes'.[6] Respondents on average fare similarly in terms of scores for knowledge (75% vs 79%) and adoption (67% vs 64%) of best business management practices. They were also similarly likely to have applied for a loan as well as to have missed a loan payment in the past 3 months. A key difference is that sales and profits, and time spent on business are much higher for the BDM sample. This is expected as Nairobi is the capital city of Kenya and a regional economic hub, so businesses based there are larger volume compared to the rest of the country.

2.4.3 Timeline

The timeline for the experiment implementation and data collection is summarized below:

Nov 2021 - Dec 2021	**Intervention:** Recruitment into Study and Intervention Deployment
Rolling	**TIOLI Offers Sent:** Invitation to participate in demand elicitation through TIOLI offers sent
Nov 2022 - Dec 2022	**Endline Survey:** Phone-based data collection on knowledge and adoption of best business practices, and business outcomes
Dec 2023 - Jan 2024	**BDM-styled Demand Elicitation:** In-person BDM-styled demand elicitation exercise conducted in Nairobi

2.5 Results

This section presents the main results from the study. Proceeding subsections cover construction of demand curves using the two demand elicitation methods, ruling out of alternative explanations for positive demand, and analyses exploring correlates of demand.

2.5.1 Demand Curves

2.5.1.1 TIOLI Offers

Of the 415 individuals who responded to the TIOLI invitations, 272 chose to buy the additional training, 111 chose not to buy, and 32 responses could not categorize by the system.

[6] Term used in terms of the original field experiment in Chapter 1.

I use a conservative acceptance rate by grouping the last category with those who explicitly chose to not buy.

Figure 2.1 shows the (inverse) demand curve constructed using buying decisions amongst the full TIOLI sample. The horizontal red line shows the marginal cost per person faced by the service provider. I observe that about 70% of individuals chose to accept the offer when the training was offered for free. When the price was half the marginal cost per user for the full training faced by the provider (KES 5), acceptance rate fell slightly to approximately 68%. From the price being half the marginal cost to double the marginal cost, there is a significant reduction in the acceptance rate, but even at KES 20, almost half of the individuals were willing to buy the training. Figure B.2.1 in Appendix B.1 shows that this demand curve looks very similar when constructed based only on the sample overlapping with the Endline survey.[7]

2.5.1.2 BDM Elicitation

The average maximum willingness to pay in the BDM sample was KES 50 (five times the marginal cost), with a standard deviation of KES 34.9. Figure B.2.2 in Appendix B.2 illustrates the distribution and shows bunching of responses at KES 0 (9%), KES 20 (10%), KES 50 (27%), and KES 100 (24%).[8]

I use the maximum willingness to pay for each respondent elicited using the BDM method to plot the proportion of the sample that would buy the training at each integer price level between 0 and 100 Kenyan Shillings. Figure 2.2 shows the resulting (inverse) demand curve. I find that the entire sample was willing to accept the SMS training when it was offered for free, while 23.3% of the sample was willing to pay KES 100, which is ten times the marginal cost.

Both methods of elicitation thus reveal a downward sloping demand curve, and show that micro-entrepreneurs are willing to pay a small amount to get access to SMS-based business trainings. However, this positive willingness to pay might not reflect actual demand - the following subsection discusses potential alternative explanations and offers arguments ruling them out in this context.

2.5.2 Ruling Out Alternative Explanations

Since about 70% of the treatment group did not start engaging with the content and no one in the control group had been exposed to it either, the willingness to pay in the TIOLI as well as BDM elicitation could be driven by those with less or no engagement with the content. If this was the case, I would see a clear negative relationship between exposure to training content and willingness to pay. I test for this by (i) comparing engagement levels in the TIOLI sample across those who bought the training and those who did not,

[7] Table B.1.2 shows that the randomized pricing arms are jointly balanced on observables using data from the Endline survey.

[8] This is not surprising in light of literature on round-number bias (Lynn, Flynn and Helion (2013)).

and (ii) estimating the causal effect of training on willingness to pay elicited through the BDM exercise. Table 2.1 shows that engagement in the TIOLI sample at both the extension margin[9] as well as intensive margin[10] is not lower amongst those who chose to buy the additional training, compared to those who chose not to. In fact, engagement rates were higher amongst those who chose to buy,[11] suggesting that those who are more exposed to the content are more likely to pay for an additional training. Furthermore, the causal effect of the SMS training on the maximum willingness to pay in the BDM sample is also non-negative - Table 2.3 shows an imprecisely measured positive effect of the SMS training on willingness to pay.[12] I therefore do not find evidence for a negative relationship between exposure to content and willingness to pay for SMS-based trainings.

Alternatively, the observed positive willingness to pay could also be due to reciprocity (Gouldner (1960)); respondents might have felt the need to give back for being part of the study by agreeing to buy the additional training. This might be a concern for the BDM sample since it involved enumerators meeting respondents in-person to conduct the elicitation. It is unlikely that reciprocity affected responses to TIOLI offers since that elicitation was done over SMS and was completely automated, and yet I still observe positive willingness to pay. I therefore conclude that reciprocity is unlikely to be driving the positive willingness to pay in this context.

The observed willingness to pay could also just be because a certain type of people would say yes to anything, especially if the price is so low. However, I argue that this is also unlikely to be driving the results since I find that willingness to pay systematically decreases with price level; the demand curves are downward sloping.

I therefore conclude that the decision to buy the SMS-trainings is a reflection of micro-entrepreneurs' intrinsic valuation of having access to the trainings.

2.5.3 Who Buys?

This section presents some correlational observations with regards to determinants of willingness to pay for SMS-based trainings.

Table 2.3 shows raw averages of individual level variables including demographic and enterprise characteristics in the TIOLI sample amongst those who accepted the TIOLI offer, those who rejected, and the difference between the two. I observe that number of children in household, knowledge about best business practices, and likelihood of having applied for a loan in the last three months is significantly higher amongst those that choose to buy compared to those that don't. Table 2.4 shows the results from OLS and Logit regressions of the decision to buy on the price levels in the TIOLI offers,[13] and the same individual

[9] The proportion of respondents offered the training who chose to start engaging with it.

[10] The proportion of the training content that the average respondent engaged with.

[11] Albeit, not statistically significantly so.

[12] Table B.1.3 shows that the treatment and control groups for the BDM sample are balanced on observables.

[13] Using the price of zero (free) as the base case.

characteristics. Column 1 shows the acceptance rate at the different price levels, in line with the figures from the demand curve in the preceding section. Column 2 shows the regression output when including demographic and enterprise characteristics. I observe that the significant differences in the raw comparisons in Table 2.3 survive controlling for all variables together - number of children in the household, knowledge about best practices, and having taken out a loan recently are all significantly positively related to the probability of accepting the TIOLI offer for another SMS training, controlling for other variables. The table also shows sales and profits to be statistically significantly related, but the magnitudes of the coefficients are negligible. Columns 3 and 4 show the same trend in terms of the log odds of accepting the TIOLI offer.

Table 2.5 shows the relationship between willingness to pay and individual characteristics in the BDM sample using bivariate OLS regressions. I observe willingness to pay is positively correlated with education statistically significantly. The correlation with Profits in last 30 days is also statistically significant, but the magnitude of the coefficient is negligible. Table 2.6 shows that the statistically significant positive association of willingness to pay with education is robust to controlling for other variables. Profits also remains statistically significant, but with a negligible coefficient.

Correlational analyses of determinants of demand for SMS-based business trainings in the TIOLI and BDM samples are thus in line with the intuition that those that have more dependents/mouths to feed at home, and those in need of funds recently are more likely to take on a potential opportunity to increase their business profits. Additionally, those with more knowledge about best business management practices and those who are more educated recognize the importance of information about best practices and are more likely to want to learn more.

2.6 Conclusion and Policy Implications

This paper examined the demand for SMS-based business trainings amongst Kenyan micro-entrepreneurs. Leveraging key components of a field experiment aimed at evaluating the effectiveness of SMS-based trainings, I studied demand using two methods. First, I analyzed buying decisions of 415 individuals who were sent TIOLI offers for an additional SMS business training via text messages, where the price was randomized across respondents over three levels; (i) free, (ii) half the marginal cost to service provider (KES 5), and (iii) double the marginal cost (KES 20). Second, I conducted an in-person BDM-style demand elicitation exercise aimed at measuring willingness to pay for SMS business trainings among 103 entrepreneurs based in Nairobi.

In the TIOLI sample, 70% of individuals chose to accept the additional training when it was offered for free, 68% accepted when the price was KES 5 (half the marginal cost), and about 50% accepted when the price was KES 20 (double the marginal cost). In the BDM sample, the average willingness to pay for SMS trainings was KES 50, and almost a quarter of the respondents were willing to buy the training for KES 100 (ten times the marginal cost to service provider). Both methods of demand elicitation thus show that micro-entrepreneurs were willing to pay a positive amount for SMS-based business management trainings, and a substantial proportion of them were willing to pay much more than the marginal cost to the service provider.

Additionally, the paper presented correlational evidence showing that demand for trainings was positively associated with number of children in the household, knowledge about best business practices, education level, and having recently applied for a loan.

Conventional in-person classroom-style business trainings are too expensive for micro-entrepreneurs in low-income settings and are usually provided free of cost at great expense to external parties. This limits the scale at which they can be deployed and creates sustainability concerns. SMS-based trainings offer an affordable alternative, and this study reveals that micro-entrepreneurs would likely be willing to pay for them out of pocket. A market-based approach to providing this service might thus be feasible, which would be more sustainable and have stronger incentives for continuous improvement of service in the direction of local needs.

Taken together with the results from Chapter 1, these findings also highlight that engagement with remotely provided content might not reflect true demand for SMS-based business trainings amongst micro-entrepreneurs, suggesting that behavioral drivers might be limiting utilization. Further work exploring these drivers and how to address them would be a promising research direction going forward.

2.7 Main Tables and Figures

2.7.1 Main Tables

Table 2.1: Engagement Levels Across TIOLI Purchase Decisions

Variable	Accept	Reject	Diff
Engaged	0.51	0.45	0.08
	(0.50)	(0.50)	(0.06)
% Engaged	0.14	0.11	0.03
	(0.25)	(0.22)	(0.03)
Observations	272	143	415

Notes: The table shows average extensive and intensive margin engagement levels amongst those who accepted the TIOLI offer in the treatment group, those that rejected, and the difference between them. Robust standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table 2.2: Effect of Training on Willing to Pay - BDM

	(1)	(2)	(3)
	OLS	IV	IV
Training	2.235		
	(7.438)		
Engaged		12.90	
		(42.57)	
Engaged $\geq$ 25%			50.00
			(167.2)
Female	-7.446	-8.162	-6.420
	(6.946)	(6.880)	(8.417)
P-value	.764	.762	.765
Control Mean	48.16	48.16	48.16
Observations	103	103	103

Notes: This table shows the intent-to-treat and local average treatment effect estimates of SMS trainings on maximum willingness to pay for SMS trainings elicited via the modified BDM method. Coefficients represent effects in terms of Kenyan Shillings. Column (1) shows output from an OLS regression, Column (2) shows output from a 2SLS regression where the endogenous variable is whether or not the individual engaged with training content, and Column (3) shows output from a 2SLS regression where the endogenous variable is whether or not the individual engaged with at least 25% of the training content. Robust standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table 2.3: TIOLI - Differences Between Buyers and Non-buyers

Variable	Accept	Reject	Diff
Female	0.45	0.41	0.04
	(0.50)	(0.49)	(0.05)
Rural	0.51	0.46	0.05
	(0.50)	(0.50)	(0.05)
Years of education	11.98	12.13	-0.15
	(2.48)	(2.46)	(0.27)
Age	33.76	33.41	0.35
	(8.38)	(9.01)	(0.93)
Num of adults in household	2.72	2.64	0.07
	(1.36)	(1.58)	(0.16)
Num of children in household	2.32	1.81	0.51***
	(1.44)	(1.41)	(0.16)
Knowledge	0.77	0.73	0.04**
	(0.14)	(0.14)	(0.02)
Adoption	0.67	0.67	0.00
	(0.17)	(0.16)	(0.02)
Sales in last 30 days	64386.18	50740.97	13645.21
	(119183.62)	(93513.92)	(12041.96)
Profits in last 30 days	19564.45	19674.42	-109.97
	(30842.64)	(31778.75)	(3375.95)
Applied for a loan	0.53	0.44	0.09*
	(0.50)	(0.50)	(0.05)
Loan payment missed/late	0.71	0.60	0.11
	(0.46)	(0.50)	(0.10)
Hours worked on business in last 30 days	221.45	202.29	19.16
	(126.64)	(109.27)	(13.11)
Hours worked on side jobs in last 30 days	21.77	27.08	-5.30
	(60.90)	(70.11)	(6.95)

Notes: This table shows the averages of listed variables amongst those who accepted the TIOLI offer for a second SMS training, those who rejected, and the difference between them. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2.4: TIOLI - Determinants of Decision to Buy

	OLS		**Logit**	
	(1)	(2)	(3)	(4)
Price = KSH 5	-0.0225 (0.0529)	-0.0180 (0.0561)	-0.105 (0.245)	-0.0678 (0.279)
Price = KSH 20	-0.229*** (0.0672)	-0.145* (0.0791)	-0.964*** (0.280)	-0.638* (0.349)
Female		0.0343 (0.0535)		0.141 (0.262)
Rural		-0.0091 (0.0525)		-0.0429 (0.251)
Years of education		0.00108 (0.0105)		0.000145 (0.0534)
Age		-0.00128 (0.00312)		-0.00727 (0.0147)
Num of adults in household		-0.0101 (0.0189)		-0.0549 (0.0902)
Num of children in household		0.0654*** (0.0186)		0.328*** (0.0982)
Knowledge		0.371** (0.181)		1.888** (0.861)
Adoption		0.0141 (0.153)		-0.0259 (0.739)
Sales in last 30 days		0.0000009*** (0.000000336)		0.00000636** (0.00000315)
Profits in last 30 days		-0.00000315** (0.00000128)		-0.0000205** (0.00000891)
Applied for a loan		0.090* (0.0506)		0.463* (0.249)
Loan payment missed/late		0.0633 (0.0628)		0.336 (0.333)
Hours worked on business in last 30 days		0.000148 (0.000230)		0.000646 (0.00117)
Hours worked on side jobs in last 30 days		-0.000116 (0.000492)		-0.000573 (0.00234)
Intercept	0.70*** (0.0305)	0.227 (0.247)	0.849*** (0.145)	-1.294 (1.228)
N	415	350	415	350

Notes: This table shows OLS and Logit regressions of the decision to accept the TIOLI offer for another SMS training, on the listed variables. Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2.5: BDM - Bivariate Analyses for Willingness to Pay

	Coefficient	Standard Error	Observations
Female	-7.595	(6.828)	103
Rural	7.233	(11.97)	103
Years of education	3.592*	(1.503)	103
Age	-0.226	(0.357)	103
Num of adults in household	-0.171	(2.896)	102
Num of children in household	0.260	(2.502)	102
Knowledge	14.77	(23.05)	103
Adoption	-9.340	(17.30)	103
Sales in last 30 days	0.0000325	(0.0000320)	103
Profits in last 30 days	0.000204*	(0.0000967)	103
Applied for a loan	-1.558	(7.005)	103
Loan payment missed/late	-11.95	(8.956)	103
Hours worked on business in last 30 days	0.0348	(0.0335)	103
Hours worked on side jobs in last 30 days	-0.0433	(0.0650)	103

Notes: This table shows results from bivariate OLS regressions of the maximum willingness to pay for an SMS training, on the listed variables. Each row represents a separate regression. Intercepts are included in the regression but not showed in the table for clarity. Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

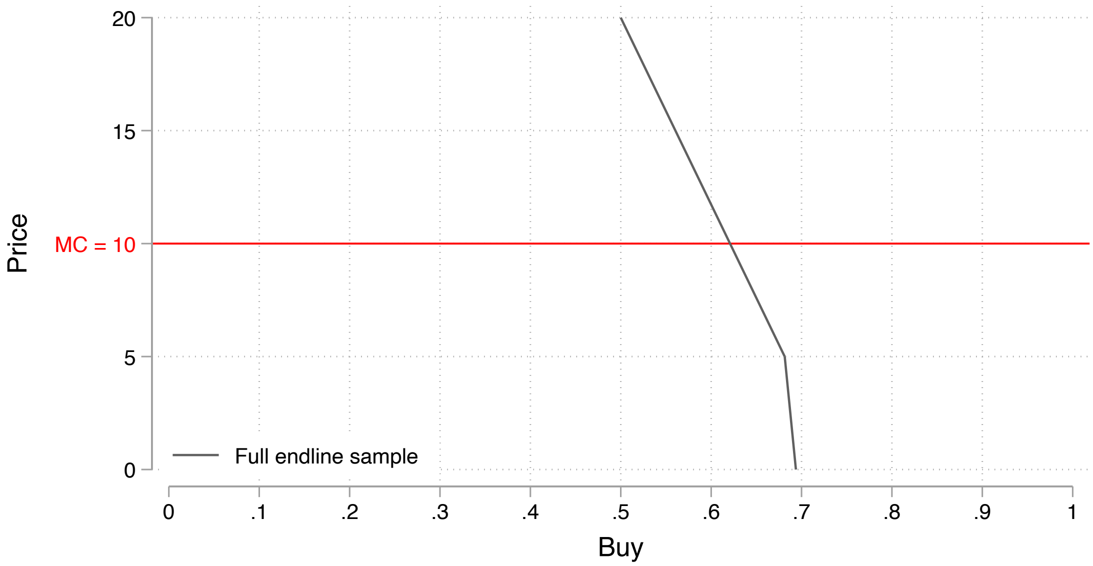
Table 2.6: BDM - Determinants of Willingness to Pay

	max_wtp
Female	-4.040
	(7.704)
Rural	9.901
	(15.01)
Years of education	3.673*
	(1.851)
Age	-.464
	(.372)
Num of adults in household	1.771
	(2.918)
Num of children in household	1.575
	(2.442)
Knowledge	13.33
	(26.85)
Adoption	-26.96
	(17.66)
Sales in last 30 days	-.0000252
	(.0000407)
Profits in last 30 days	.000219*
	(.000124)
Applied for a loan	.675
	(7.146)
Loan payment missed/late	-11.73
	(9.496)
Hours worked on business in last 30 days	.00416
	(.0429)
Hours worked on side jobs in last 30 days	-.0971
	(.0687)
Constant	21.97
	(30.38)
Observations	102

Notes: This table shows OLS regression of the maximum willingness to pay for an SMS training, on the listed variables. Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

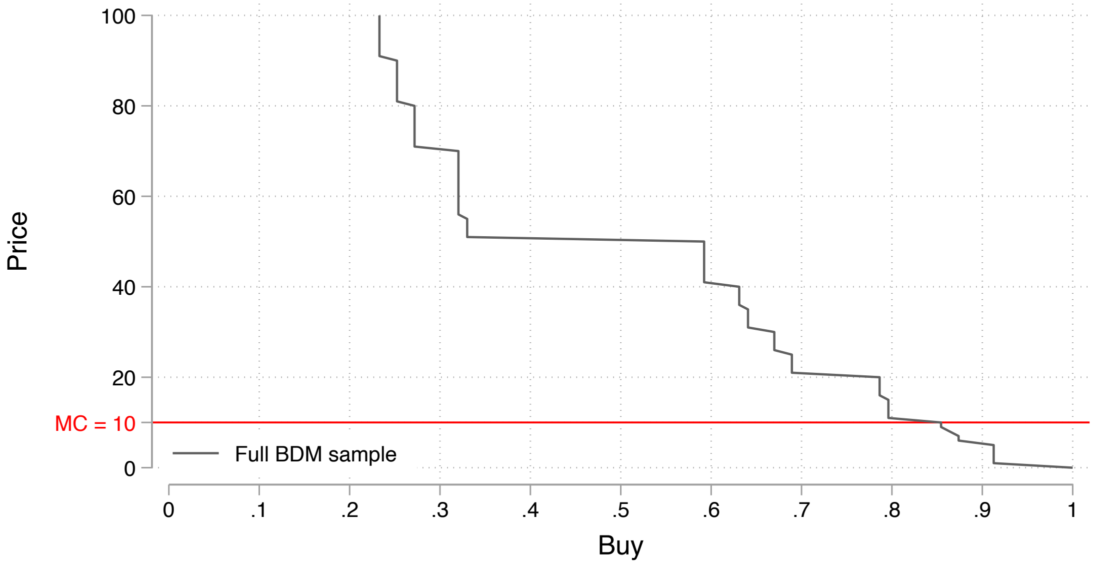
2.7.2 Main Figures

Figure 2.1: Willingness to Pay - TIOLI Offers



Notes: This figure shows the (inverse) demand curve based on buying decisions from randomized take-it-or-leave-it offers sent to treatment individuals. The horizontal red line represents the per person marginal cost faced by provider for delivering the entire training.

Figure 2.2: Willingness to Pay - BDM



Notes: This figure shows the (inverse) demand curve based on the maximum willingness to pay elicited using the in-person elicitation using the modified BDM method. The horizontal red line represents the per person marginal cost faced by provider for delivering the entire training.

Chapter 3

Command and Can't Control: Assessing Centralized Accountability in the Public Sector

3.1 Introduction

How the bureaucracy performs is fundamental to the provision of high-quality public services in the developing world (Besley et al. (2022)). Recent approaches to bolstering the functioning of public administration have focused on de jure improvements in formal contracting environments such as introducing pay-for-performance (Muralidharan and Sundararaman (2011); Dal Bó, Finan and Rossi (2013); Ashraf, Bandiera and Jack (2014); Deserranno (2019); Leaver et al. (2021)). However, the vast majority of reforms to government administration implemented at scale relate to shaping the de facto incentive environment in the bureaucracy instead of introducing changes in legal and fiscal environments. The Global Survey of Public Servants (Schuster et al. (2023)), run in 35 countries, reports that only 31% of public servants perceive their public service as actualizing de jure performance incentives, while 76% state that de facto reward systems are in operation.

A canonical de facto bureaucratic reform is *command-and-control* management, or hierarchical systems of control where officials are expected to follow centrally determined directions or face punishment. Finer (1997)'s magisterial overview of administrative arrangements of government throughout history emphasizes the continuous efforts of monarchies and autocracies towards the centralization of information and control around a sovereign. Modern military administrators across the world rely on command-and-control for effective governance across the hierarchy (Wilson (1989); Hoehn, Campbell and Bowen (2021)).

Faced with constraints on de jure changes in public sector incentives, civilian public sector bureaucracies have been attracted to adopt a command and control model. Following the purported success of British Prime Minister Tony Blair's 'delivery unit',[1] over 80 countries

[1] See The History of Government Blog (2022) for more details.

have set up centralized routines and offices (see Figure 3.1) that "combine functions such as target-setting, monitoring, accountability, and problem-solving with the aim of rapidly improving bureaucratic performance and service delivery" ((Education Commission, 2023, p. 7)). What distinguishes these reforms is the extraordinary political and executive backing they received around the world. However, evidence on the efficacy of applying command and control approaches to modern public administrations at scale remains scarce.

We study such a scheme implemented at scale in the education public administration of Punjab, Pakistan, where monthly education data from over 50 thousand public schools was channeled to the highest executive authority and used to set targets and establish accountability throughout the organization. This command-and-control scheme in Punjab is considered a showpiece of the centralized accountability delivery model: it was implemented to a very high standard for over six years, was advised by top experts in the world, and had the full backing and involvement of the most senior members of the executive (Barber (2013); Chaudhry and Tajwar (2021); Malik and Bari (2022)).[2]

Our analysis focuses on the efficacy of the scheme as a driver of improved educational outcomes. We collect the administrative data from all 52,000 schools in Punjab from December 2011 to May 2018 on which the scheme was built and digitize the monthly reports created for senior managers that flagged performing and underperforming school districts.[3] The monitoring reports present performance metrics drawn from this data, aggregated at the administrative unit, for a range of school outcomes including teacher presence, student attendance, functional facilities, and from September 2017, student test scores on standardized exams. Using this data, we examine how senior officials' high-frequency monitoring of public services and efforts to exert control impact subsequent school performance.

To more deeply assess the impacts of command and control approaches on public administration, we also collect data on key elements of the education administration related to financial and personnel resources, bureaucratic attention to individual schools, and the career progressions of affected officials. This data allows us to unpack the impact of the scheme across the hierarchical chain and explore a broad range of bureaucratic responses to the 'command-and-control' system.

In our core specifications, we use a stacked difference-in-differences design (Cengiz et al. (2019); Baker, Larcker and Wang (2022)) to assess the impact of a public official being flagged by the monitoring system on educational outcomes under their responsibility. An

[2] Education Commission (2023) write that "the chief minister... attended all 39 stocktake meetings to hold districts accountable, and took action to solve implementation bottlenecks in the quarterly high-stakes meetings" (p.16). A qualitative review of the scheme stated "At the core of the approach design was leveraging political interest and political capital to orient the bureaucratic structures involved in service delivery toward improvements at a fast pace" (Malik and Bari (2022)). The implementation in Punjab is highlighted as one of the success stories around the world. Reviewing the scheme in an interview in 2017, Michael Barber, one of the architects of the delivery approach around the world, stated, "Punjab is unique ... across the whole world for combining deliverology with really good and modern technology."

[3] The school-level data was collected by an agency within the education sector that is fully independent of the bureaucrats being monitored, and we validate its quality by using a distinct set of independent assessments.

official is only flagged if a sufficient percentage of schools within their jurisdiction have fallen below a threshold in the outcome of interest. The first difference in our research design compares schools in flagged and non-flagged administrative units. To make this comparison sharper, we additionally hone in on a sample of administrative units, labeled the 'threshold sample', that lies just above or below the threshold for flagging so that both administrative units see a comparable drop in the outcome of interest, but only one of them is flagged. The second difference compares the trajectory of treated and non-treated administrative units over time so that we can assess their response to shocks in the outcomes of interest.

We find precisely estimated evidence that the scheme had no substantive impact on targeted school outcomes: teacher and student attendance, functional school facilities, as well as English, Mathematics, and Urdu test scores. Though there is a small increase in the rate at which teachers return to schools after an absence, the limited magnitude on an outcome clearly within the authority of public managers - a 2 percentage point faster month-on-month improvement - in fact underlines the limitations of the scheme.

Despite these broadly null effects the program was maintained, and further developed, for 6 years. A potential reason for the persistence of the program is that a naive examination of before-after comparisons yields a strong positive effect of the program. Many outcomes in the policy domain exhibit reversion to the mean following idiosyncratic shocks, such as student test scores (Chay, McEwan and Urquiola (2005)). Our paper extends this finding to the overarching machinery of public administration. In our education setting, after a shock, schools in flagged areas follow a similar pattern of return to their equilibrium state of service delivery as their comparison schools in areas that were not flagged. Though senior managers observe the resolution of alert flags for particular administrative units, comparison to an appropriate counterfactual implies that this resolution does not seem to be due to their efforts.

It is possible that despite no overall impacts, the scheme produced significant changes in activity within the bureaucracy. We capitalize on our rich data on administrative activity to assess the impacts of flagging on key components of the public education production function. The scale of the data we have assembled allows us to estimate even small impacts with precision, painting an unusually rich picture of the impact of reforms on bureaucratic activity. We assess the financial and personnel decisions of bureaucratic managers responsible for the flagged areas. We do not observe more visits from relevant bureaucrats to affected schools, changes in their financial investments across schools, or bureaucratic transfers of teachers and head teachers. Thus overall, despite the enthusiasm for the reform of senior managers in Punjab, command and control management approaches did not motivate rank-and-file officers to change education outcomes in any substantively significant way.

We contribute to a growing literature on bureaucracy and development broadly (Finan, Olken and Pande (2015); Besley et al. (2022)), and on designing optimal incentive structures in the public sector more specifically (Banerjee et al. (2021); Ali et al. (2021)). Recent (frequently experimental) papers in this literature have made the important contribution of showcasing the efficacy of various incentive schemes such as financial rewards (Muralidharan and Sundararaman (2011); Dal Bó, Finan and Rossi (2013); Ashraf, Bandiera and Jack

(2014); Deserranno (2019); Leaver et al. (2021)), career incentives (Khan, Khwaja and Olken (2019); Bertrand et al. (2020); Deserranno, Leon and Kastrau (2022)), or other non-financial incentives (Ash and MacLeod (2015); Khan (2020); Honig (2021)). However, implementing many of these reforms at scale would require changes to the de jure environment which has been difficult to implement at scale.[4] Given the systemic nature of centralized accountability, command and control reforms are poorly suited to experimental evaluation. We present the first at-scale evidence in the economics literature on this classic pillar of Weberian bureaucracy: centralized control mechanisms.

Our findings are also relevant for the literature on the efficacy of management approaches in the public sector (Bloom and Van Reenen (2010*b*); Bloom et al. (2015); Rasul and Rogger (2018); Rasul, Rogger and Williams (2020); Banerjee et al. (2021); Ali et al. (2021); Carreri (2021)). Importantly, our study indicates that even strong centralized support for a management intervention can have passive impacts on public sector functioning. We are able to track key elements of the administrative production function precisely, supporting our null findings with evidence that the machinery of government was unmoved. Evidence on the impact of control mechanisms on public sector performance is mixed, with generally positive results for frontline settings (Olken (2007); Hussain (2015); Dhaliwal and Hanna (2017); Callen et al. (2020); Duflo, Hanna and Ryan (2012); Das et al. (2016)); and less supportive evidence from experiments about administrator's motivation and performance, or those dealing with organizational dynamics (Falk and Kosfeld (2006); Dickinson and Villeval (2008); Bandiera et al. (2021); Muralidharan and Singh (2020)). We extend this literature by providing evidence of the effects of centralized oversight on a broader administrative environment from an at-scale implementation in a large bureaucracy. By doing so, our study adds to the literature on the impacts of government-implemented schemes, which are argued to be a test of the external validity of pilot programs (Bold et al. (2018); Muralidharan and Niehaus (2017); Vivalt (2020)) and an assessment of the most widely used public sector reforms (de Ree et al. (2017)).[5]

We also add an early contribution to the nascent study of a key feature of bureaucracy: hierarchy. Though the theory of hierarchy in organizations continues to develop (Aghion and Tirole (1997); Dessein (2002); Chen (2017); Chen and Suen (2019)), there are few related

[4] By scale we mean both geographic coverage, but also temporal sustainability. Important exceptions are usually historical studies that examine major changes to civil service legislation (see for instance Xu (2018), Mehmood (2022), Aneja and Xu (2023), and Riaño (2021)). In fact, many papers examining these questions in modern bureaucracies refer to fixed de jure incentives under the Northcote-Trevelyan *system* that contain three features: competitive exam-based recruitment, rule-based promotions, and permanent civil service protected from political interference ((Besley et al., 2022, p. 400)). There are limited opportunities to examine how at scale changes in these impact the bureaucracy. See for instance Bertrand et al. (2020) how changes in the retirement age change career concerns in India.

[5] The paper provides a lens through which to understand the results of smaller pilots of centralized oversight, such as Callen et al. (2020), which show that flagging underperforming health facilities in Punjab positively affected health workers' attendance. However, when taken to scale, such pilots may not provide a sustainable means of managing the public administration (Banerjee, Duflo and Glennerster (2008); Banerjee et al. (2021)).

empirical tests in the literature. Recent evidence implies that understanding hierarchy in public organizations is critical to behavior there (Deserranno et al. (2022); Cilliers and Habyarimana (2023)). This paper shows that de-facto pressure directed through hierarchy may not engender substantial responses from public officials, however, salient senior management makes this form of incentive provision.

The paper proceeds as follows: Section 2 describes the setting of the public service we study and describes the centralized monitoring scheme. Section 3 introduces the data and presents our empirical approach. Section 4 presents the results of the evaluation of the scheme on school outcomes. Section 5 presents assessments of the scheme's impact on key elements of the education administration. Finally, Section 6 provides a discussion of our results in light of potential alternative uses of the data that was generated to run the scheme.

3.2 Public Education in Punjab

Punjab is Pakistan's most populous province, home to over 110 million people, half of the country's population. Twenty million are school-aged children, many attending approximately 52,000 public schools, with 400,000 teachers (*School Census* (2018)). The scale of managing education in the province is substantial.

The province is divided into 36 districts, which are subdivided into sub-units called tehsils, further subdivided into areas of responsibility called "maraakiz."[6] There are, on average, four tehsils per district and 48 maraakiz per tehsil. Thus, on average, any district education manager has 192 administrative units to track, and each markaz official must manage an average of 20 schools.

The School Education Department is responsible for organizing and overseeing the education sector's performance. The department has two arms: district education authorities, which coordinate the implementation of public education delivery, and the Program Monitoring and Implementation Unit (PMIU), which is responsible for independently collecting and disseminating data on school performance. Both are staffed and organized separately, and monitoring is generally seen as independent of implementation.

3.2.1 District education authorities

Each district in the province has one district education authority which reports directly to the School Education Department. Below them, the hierarchy consists of officers for each tehsil, and assistant education officers (AEO) for each markaz. Each layer of the hierarchy is expected to manage those officers under them. AEOs are the layer of hierarchy above school principals, thus completing a multi-link chain of command from senior executive to school level.[7]

[6] Plural of the term "markaz," the Urdu word for "center."

[7] Further, schools are categorized into one of three groups: elementary education female, male, and secondary education. Our study focuses on elementary level (male and female), comprising primary schools

Such a layered hierarchy is not unusual in administrative settings of this scale worldwide, as the physical constraint of traveling to schools, handling administrative tasks for each, and engaging with head teachers implies a limit on the scale of any individual official's ability for oversight. By contrast, a feature of large-scale measurement in management information systems is that it can alleviate the physical constraints and centralize the ability to supervise and censure at scale. By dramatically lowering the cost of monitoring individual schools and jurisdictions, digitization of public service delivery measures has opened up the possibility of centralized management throughout the hierarchy. Such a system of monitoring the administration requires an independent administration, which we turn to next.

3.2.2 The PMIU

While the district education authorities are responsible for outcomes in public schools, the Program Monitoring and Implementation Unit (PMIU) is tasked with monitoring the performance of district officers. To do so, it conducts an annual census of all public schools in the province and a monthly monitoring of schools to assess key aspects of the school environment. Undertaking these duties are monitoring assistants hired to collect data.

Across the analysis period, the monitoring assistants collected performance-related data from every school on an unannounced random date every month. The assignment of monthly school inspections to monitoring assistants was randomized to limit collusion with the school staff. As we discuss in the data section, our analysis of the consistency of different data sources on schools implies that this process produced valid assessments of school performance.

Data collected by the PMIU was used for monthly and quarterly performance reports, called 'data packs.' These data packs were first generated in December 2011 and then prepared monthly. We study the period until May 2018, just before the national elections and a change in administration. The data packs reported performance at the markaz level for each district along multiple dimensions: teacher presence, student attendance, visits by district staff, and status of school facilities (electricity, drinking water, toilets, and boundary wall).[8] From September 2017, the data packs also reported scores on standardized Math, English, and Urdu tests.

The reported performance on each dimension was color-coded in the data packs based on standardized performance thresholds set by the chief minister's team. A markaz could be coded red, orange, or green, with red being the primary flag for underperformance. Figure C.1.1 in the Appendix illustrates the color-coding. As such, an AEO (markaz-level officer) would be associated with any underperformance, although flagging was also done (in a less systematic way) at the tehsil and district level. The focus of the discussions was on markaz performance, and so that is the emphasis we follow in our empirical work, though we also provide consistent evidence for flagging at the district level.

(children aged 4 to 9) and middle schools (children aged 10 to 12). These makeup roughly 80 percent of all public schools in the province.

[8] Also included the number of schools surveyed, if they were found closed, statistics by male and female schools, and recommendations about which schools to focus on to improve outcomes.

3.2.3 Centralized Oversight Intervention

Using the PMIU-generated data on school performance, the chief minister of Punjab set up a centralized oversight regime for the education sector in 2011. He chaired an oversight committee and worked with the consultancy firm McKinsey International and a high-level advisor with expertise in centralized accountability.

Figure 3.2 describes the design of the monitoring scheme. Data on all schools in the province is collected in month t. Markaz-level average performance is presented to senior managers in month $t+1$. Maraakiz that do not reach specific (standardized) thresholds are flagged red or orange. The reports were used for senior management check-ins within the first ten days of every calendar month.

In addition, quarterly meetings were held where "the [chief minister] at that time was himself very very motivated and he would make it a point to not miss any one of the meetings."[9] The senior management of the province placed substantial weight on the system, and the chief minister "had full ownership of this reform and [sent] a signal to the bureaucracy that they were to take it seriously" ((Malik and Bari, 2022, p. 22)).

Senior managers did not change de jure power, such as making salaries conditional on performance. Some ad hoc financial bonuses were given to district officials but not to mid-level bureaucrats. We explore whether there is evidence of staff transfers or long-term impacts on career trajectories from poor performance. We do not find any such evidence. Instead, senior management was constrained by public service rules meant to avoid political influence. Thus, the system had to rely on de facto incentives to punish underperforming officials.

Interviews with district officials revealed that meetings mostly involved the officers flagged red getting censured in front of their peers. Quoting Malik and Bari (2022), "the red were reprimanded, and the greens were appreciated", where "The constant monitoring by the Chief Minister and the Chief Secretary played a very critical role." Officials stated that they did "not want to be punished in front of our colleagues." As the chief minister's staff officer recounts, "I wouldn't say it was fear necessarily but the point [is] that the quarterly rankings and the performance accountability caused a lot of concern."

The censoring generated incentives for district officials to motivate their subordinates, and this to those below them. The scheme intended that greater oversight by senior management would allow sanctions to serve as motivation through the chain of command. As such, the scheme relied on the interaction between measurable outcomes and personnel management. In public sector oversight models, the outputs can be reduced to observable quantities, but improvements in these still rely on multidimensional and non-contractible activities. So then, the question under evaluation is whether oversight and accountability regimes effectively motivate better personnel management throughout the hierarchy.

The political weight and international guidance ensured the scheme was effectively implemented. Reports were produced monthly from December 2011 to May 2018 as intended.

[9] Malik and Bari (2022) state that "All other practices of priority setting, target setting and use of data for monitoring were all feeding into the construction of this accountability mechanism that was arguably central to the design of the delivery approach that was instituted in Punjab."

To assess the data quality, we compared it with the Annual Census of Schools for the month the annual census was collected. Both data sources reported information about the number of teachers posted, enrolled students, and the functionality of school infrastructure. Figure C.1.2 in the Appendix compares both sources and shows that the overall error in reporting is low and there is a high overlap between both data sources. A comprehensive review of the data we use assesses it to be of generally high quality (World Bank (2020*a*)).

Despite slight modifications to the scheme's structure, these elements remained at its core. As a result, the design is a demonstration case of centralized, data-informed accountability regimes. The centrality of the scheme to the administration's management, the scale and quality of data collection, and the length of time that the scheme was in place all make the scheme a good test for the efficacy of such approaches in the public sector.

3.3 Evaluation methodology

3.3.1 Data

We used administrative data collected at the school level from December 2011 to May 2018.[10] The outcomes are monthly assessments of teacher presence, student attendance, and whether school facilities are functional. The first two are measured as the percentage of teachers/students present at the time of the visit by the monitoring assistants. The functional facilities measure the status of four types of school infrastructure: drinking water, electricity, toilets, and the boundary wall. We use an aggregate index of the share of functional facilities.

Additionally, starting in September 2017, PMIU began collecting data on student test scores in Math, English, and Urdu using standardized tests, administered by monitoring assistants to seven randomly selected 3rd-grade students in each school. Scores are measured as the percentage of correct answers. To understand the effect of bureaucratic behavior, we also use the data on district education staff visits to schools. We can identify each school's district, tehsil, and markaz, as well as the history of flagging across administrative tiers and units.

Over the entire period, 82% of maraakiz were flagged red at least once on some outcome, and 96% were flagged red or orange. Like any population of schools, there were some which were persistently high performers. 1.6% of schools never dropped below 90% on any of the outcomes. However, of the 82% of maraakiz flagged once, 79% got flagged again at some point. Thus, the oversight intervention was broad in its reach across maraakiz.

Flagging thresholds for color-coding in the datapacks were designed to be generally applicable to schools across the province, and based on the education authorities' pre-existing targets for performance measures. These targets were mostly the same across all districts and for all months of the year. In the case of student attendance, different targets were assigned

[10] The data excludes June, July, and August of each year, corresponding to summer vacations and public schools being closed.

across different districts and for different months of the year based on historical performance as it was felt in the case of that outcome a moving target was more appropriate.[11]

Table 3.1 report descriptive statistics. Panel A show that schools are relatively small, with an average of 4.6 teachers and 110 students. Roughly 3% of the schools have ever had more than 20 teachers. Those with more than 20 teachers are evenly distributed across the province. At the markaz level, Panel B shows a substantial variation in the number of schools within a markaz, broadly following differences in population size. However, the average number of schools an AEO must manage is 20, of which nearly 80% are elementary schools.

Panel A also shows descriptive statistics at the outcome-school-month level, separating between outcomes in flagged (on that outcome) and non-flagged maraakiz. Similarly, Panel B show descriptives at the outcome-markaz-month. By construction, the mean in a flagged markaz is lower than that in a non-flagged markaz. The month in which a markaz is flagged on a particular outcome, there is a drop in the mean level of that outcome. Comparison of the two sets of columns gives the order of magnitude of the differences. For example, flagged maraakiz have an average teacher presence of 80%, while in non-flagged maraakiz it is 93%.

In addition to the monthly flagging of AEOs/maraakiz, the districts were ranked each quarter. The ranking was based on an overall score of the performance in the previous months.[12] Panel C in Table 3.1 shows descriptive statistics for districts in the top/bottom positions. Bottom districts report a lower mean in the score. Panel C also shows the percentage of districts that entered the top/bottom five positions in each period. There is a relatively small number of cases where new districts fell into the top (7.7%) or bottom (8.3%) positions, suggesting a high degree of persistence in the ranking status.

Figure 3.3 presents this persistence graphically. For each quarterly meeting, we color-coded the quintile in which the district fell in the overall score distribution. The districts in the higher quintiles tend to maintain their high position in the ranking. In contrast, the districts in the lowest quintiles remained in last. The figure thus presents a descriptive sense that the flagging did not motivate poor performers sufficiently for their overall rankings to change.

A feature of the intervention environment is that almost all maraakiz were flagged at some point, and yet some districts and maraakiz remain systematically at the bottom of the distribution. Evidence from other settings indicates that education (and other environments) face structural constraints to improving outcomes (World Bank Group (2018)). However, they are also exposed to shocks (such as teachers getting sick) that substantially shift the absolute levels of service delivery. This would imply that Punjab's schools face shocks that sometimes push them under the flagging threshold irrespective of their baseline performance levels.

[11] Appendix C.1 provides further details about the thresholds for color-coding for each indicator of interest.

[12] Since this activity was based on a ranking, even if all districts were systematically improving, the ranking system kept rewarding districts with the highest relative scores and punishing those with the lowest scores.

The time series variation in outcomes among schools is consistent with this interpretation. Table 3.2 presents the standard deviations in school outcomes in each quintile of mean baseline performance. The top four quintiles of schools face comparable levels of variation. There is some significant probability of falling below the thresholds in each. This probability is almost a magnitude higher in the lowest quintile. The likelihood of flagging jumps toward the bottom of the distribution, implying a persistently challenging environment to manage.

3.3.2 Empirical strategy

To estimate the effect of the centralized accountability system on educational outcomes, we followed Cengiz et al. (2019) and Baker, Larcker and Wang (2022) to build a stacked dataset to avoid biases driven by the time-varying nature of the treatment (De Chaisemartin and d'Haultfoeuille (2020); Callaway and Sant'Anna (2021); Goodman-Bacon (2021)). The stacking consists of creating event-specific datasets for identifying control units that have not been treated during a specific period. The process is described in Figure C.1.3 in the Appendix. The result is a dataset with the treatment centered in relative time to eliminate its time-varying nature, conditional on indexing the estimations at the event-panel level. Following the stacked design of our data, we implemented a stacked difference-in-differences strategy.[13]

3.3.2.1 Markaz flagging

Our main specifications assess the impact of a markaz being flagged as red/underperforming on the flagged outcomes in schools within that markaz. We estimated the following equation:

$$\begin{aligned} Y_{smdte} = \gamma_1(T_{mde} \times Flag_{te}) + \gamma_2(T_{mde} \times Punish_{te}) + \\ \beta(T_{mde} \times AfterFlag_{te}) + \alpha_{mde} + \lambda_{te} + dt + \epsilon_{smdte} \end{aligned} \tag{3.1}$$

Subscripts s, m, d, t are for school, markaz, district, and time. All of the components are indexed at the event panel e. Y_{smdte} is the outcome for school s, within markaz m, in district d. T_{mde} equals 1 for schools in a flagged markaz m. $Flag_{te}$ equals 1 for the period data is collected and the flag is defined. $Punish_{te}$ equals 1 after the flagging, where the oversight committee meets and the accountability intervention occurs. $AfterFlag_{te}$ equals 1 after the

[13] The core empirical exercise we conduct in this paper uses specific features of the flagging system to estimate rigorous identification of its effects. However, in Appendix C.2.1 we also assess whether the introduction of the scheme itself created large changes in the trends of public education outcomes. We do so in three ways. First, assessing whether outcomes trended similarly before and after the introduction of the scheme in Punjab versus other territories in Pakistan. Second, whether the first flagging of any jurisdiction had a particular impact on its trajectory. Third, whether the first flagging of a jurisdiction in a district had any impact on the wider trajectory of schools there. We find no evidence on any of these margins: the introduction of the scheme did not affect the trajectory of school outcomes. This alleviates the concern that the relevant responses of bureaucrats to the scheme happened before (in expectation) or on impact. Such a coordinated and widespread response seems intuitively unlikely in a large and disparate environment.

punishment phase where we assess the intervention impact. α_{me} is for markaz fixed effects to control for constant characteristics of maraakiz, and λ_{te} is for time fixed effects to capture time-specific shocks. We include dt –a district binary and linear calendar index– to absorb district linear time trends. ϵ_{smdte} is the error term clustered at the markaz level (treatment level). In our main specifications, we stack for four pre-periods and seven post-periods.

Figure 3.4 presents the evolution of our outcomes in relative time, anchored on periods of flagging. Solid lines are schools in flagged maraakiz. Dotted lines are schools in non-flagged maraakiz. We present two dynamics: one that uses all schools (the blue lines) and one that uses only those that are "close" to the threshold for flagging (red lines). We highlight three periods corresponding to the month in which the data is collected and the flag is defined, the month in which these are reported to oversight committees and punishments occur, and the period after the flagging events, where we assess the impact of treatment.

We observe that treated and control units follow similar paths just before the flagging. In the month of flagging, the average school in a markaz that gets flagged suffers from a shock, contributing to the markaz being selected for treatment.[14] Thus, the treated units would not have followed the same transition as control units without the treatment, and the conditions for causality would be violated. To address the parallel trends violation, we follow Rambachan and Roth (2022) and redefine the base period as the one just before the negative transitory shock occurs (relative time -1).

To further account for the negative shock, we build a sample of comparable schools around the flagging thresholds for each outcome (plotted in red). We follow Calonico, Cattaneo and Farrell (2020) to identify the maraakiz within an optimal bandwidth on either side of the flagging threshold in time 0. We obtain optimal bandwidths separately for each event panel to build an stacked-threshold sample.[15] As can be observed, regardless of the sample, the transition of outcomes typically reverts to the pre-shock levels.

Then, γ_1 absorbs the effect of the negative transitory shock, and γ_2 captures the immediate recovery in the punishment period. β would estimate the effect of flagging on school performance after the shock. If flagging leads to higher outcomes on flagged units relative to non-flagged units, β should be positive. That is the core test of the specification. To illustrate the external validity of the results using the sample of schools around the flagging threshold, we also present results for the full set of schools.

3.3.2.2 District ranking

One concern is that markaz flagging might be less salient when the rest of the district performs well. We complement our core strategy with analysis at the district level. Above we noted that in quarterly oversight meetings, districts were ranked according to the aggregate performance in the prior quarter. Though we are far less powered to investigate the impact

[14] This situation is related to an Ashenfelter dip (Ashenfelter (1978); Ashenfelter and Card (1984); Heckman and Smith (1999)), which consists of self-selection into the treatment because of a negative shock.

[15] The threshold sample consists of 16% of observations of the full sample for teacher presence and student attendance, 9% for functional facilities and Urdu scores, 5% for Math scores, and 23% for English scores.

of this ranking, we apply a version of our main specification to being "flagged" as a top- or bottom-performing district on the subsequent performance of schools in that district, and additionally look at the interaction between district and markaz flagging.

We stack for four pre-periods and three post-periods as district meetings happen quarterly. We use as event time each month in which a meeting happened. Flagged units are defined as the schools in districts that were at the bottom/top of the ranking during the meeting in period 0. District rankings do not systematically receive a negative shock before the meeting, and thus do not require corrections for related self-selection and reversion to the mean. However, for consistency, we define -1 as the base period and build a threshold sample of the five districts closest to the treated five at the top/bottom to represent a threshold comparison. We estimate the effect of district ranking with the equation below:

$$\begin{gathered} Y_{smdte} = \gamma(Position_{de} \times Meeting_{te}) + \\ \beta(Position_{de} \times AfterMeeting_{te}) + \alpha_{de} + \lambda_{te} + \epsilon_{smdte} \end{gathered} \tag{3.2}$$

where $Position_{de}$ equals 1 for schools in bottom/top districts d. $Meeting_{te}$ equals 1 for the period when the quarterly meeting happens, so γ absorbs any immediate effect of the meeting. $AfterMeeting_{te}$ equals 1 for the months after the meeting, so β estimate the persistent effects of the flagging. α_{de} are district fixed effects and λ_{te} are time fixed effects. ϵ_{smdte} is the error term clustered at the district level. Interactions between this specification and the above markaz-level specification are natural extensions to these equations.

3.4 Results

3.4.1 Markaz flagging

Figure 3.5 reports the event studies for each outcome variable we study. The y-axis reports β coefficients in percentage point differences. The blue line is the full sample, while the red is the threshold sample. The event studies show that the pre-trends are not significant and are small in magnitude. Thus, the parallel trends assumption is plausible. As can be seen, most of the coefficients in both samples are statistically equivalent to zero at the 95% level in the *After flag* period, indicating null impact of the flagging. The full sample estimations exhibit a larger relative negative shock measured in period 0, but even this is almost recovered by the first *After flag* period.

We in fact see flagged schools taking longer than their equivalent non-flagged schools to return to their pre-existing levels. In particular, the coefficients related to student attendance (panel b) and English scores (panel e) take longer to reach the pre-shock level in flagged schools, though the magnitude of the effects are small. This is likely due to the fact that treatment schools have a marginally stronger shock in the outcome variable, and they may naturally have a more extended transition back to equilibrium.[16]

Table 3.3 presents the results of estimating equation 3.1. The first column for each variable reports the full sample, and the second shows the threshold sample. Panel A reports outcomes relating to school functioning. They are always negative and significant coefficients in the *Flag* and *Punish* periods for flagged relative to the non-flagged units. The coefficients for both periods represent the first negative shock and the subsequent immediate recovery, which we interpret as a reversion to the mean effect.

The coefficients for the *After flag* period (corresponding to β) are significant in both samples for teacher presence and student attendance. The coefficients are small (almost zero) compared to the mean of the dependent variable, but negative rather than positive. As observed graphically, the negative effect can be interpreted as a persistence of the negative shock. Panel B of Table 3.3 presents the results for the student test score variables. We note that the sample size is smaller here, given the reduced time frame for which we have these measures. We observe the same pattern of results as in Panel A. The results imply that the oversight scheme had no impact on school functioning nor student outcomes, but rather that flagged and non-flagged schools facing a similar shock returned to equilibria at roughly the same rate, and certainly did not improve disproportionately beyond their pre-existing levels.

To assess the robustness of these results, we present a series of additional specifications in Appendix C.2. We plot the estimate for each coefficient from a stacked data set including t additional periods to further test the stacked structure (Figure C.2.4). We assess the effects of using fixed effects that absorb the history of markaz flagging (Table C.2.1) and of changing our assumptions on the persistence in the impact of flagging (Table C.2.2). We present alternative difference-in-differences estimators (Figure C.2.6 and C.2.7). We estimate the results for the 'orange' flagging threshold (Figure C.2.8) and for flagging at the tehsil level, the layer of hierarchy above the markaz (Figure C.2.9). We assess the impact of centralized accountability separately for each month in which the scheme was implemented (Figure C.2.10). We also investigate the possibility of public officials anticipating the flagging (Figure C.2.11). In all cases, our results are qualitatively the same.

One possibility is that the system was not intended to improve student outcomes but rather to serve political ends. We therefore assess whether flagging had differential impacts across political environments. In Appendix C.2.5 we identify political alignment following Callen, Gulzar and Rezaee (2020) and use a difference-in-differences strategy to assess the effect of being in a politically aligned markaz. We compare aligned/non-aligned markaz, before/after the 2013 elections in places with high political competition (close elections). While we find no effect of political alignment on the probability of flagging, there are small effects of alignment on student attendance itself but no consistent effects elsewhere.

[16] As a robustness check, Figure C.2.5 reports the event studies for a stacked dataset with fewer post periods to test the sensibility of the results to an arbitrary number of periods. Results follow the same trends in both cases.

3.4.2 District ranking

To complement our main analysis, we assess the impact of being the top/bottom performing districts at quarterly oversight meetings. We restrict our analysis to measures of school functioning. Figure 3.6 presents the event studies for top/bottom performing districts. The figures illustrate that no pre-periods appear significant, suggesting the plausibility of the parallel trends assumption. The *After flag* period indicates that flagging once again has no significant effects on school functioning or outcomes.

Table 3.4 reports the treatment effects. Panel A for schools in the bottom districts shows that we detect a small but significant increase of 1.3 percentage points in student attendance for the threshold sample. Panel B for schools in top districts shows that being in it leads to a slight increase in teacher presence after the quarterly meeting. However, the coefficients are small in magnitude relative to the mean of the dependent variable before the meeting (91% in the full sample and 91.9% in the threshold sample). Hence, there is no evidence of significant increases in performance due to centralized monitoring of higher-level managers from the district-level rankings. The findings are consistent with the descriptive statistics in Panel C of Table 3.1 and Figure 3.3, showing that there is little movement into and out of the top quintiles of performance, with corresponding limits on the degree to which they might be motivating.

Despite finding zero overall impacts of flagging at the district level, we tested the impact of the interaction between district-level and markaz-level flagging. We hypothesize that the coincidence of flagging at both levels might create greater pressure throughout the hierarchy toward school improvement, leading to a differential increase in performance. We tested this hypothesis by estimating equation 3.2, including a triple interaction between schools in a bottom or top district in the quarterly meeting and those for which a markaz was also flagged in the month of the quarterly meeting. Appendix Table C.2.3 reports the results of the heterogeneity analysis. Panel A reports the results for the bottom districts, while Panel B reports the results for the top districts. The triple interactions for none of the panels, variables, and samples show positive and significant results, suggesting no major interaction between flagging district- and markaz-level performance.

3.4.3 Does punishment change the trend of recovery?

The recovery to pre-treatment means is a combination of mean reversion and the impact of the punishment period. A key advantage of the frequency of our data is that we can separately examine the impact of punishment beyond the regression to the mean trends in the outcomes. To do so, Figure 3.7 plots over time impacts on first-differenced outcomes that are reported above in Figure 3.5.

We can see that there exists a negative shock during the flagging month ($t = 0$). This negative shock is followed by a quick recovery in the month where punishment occurs ($t = 1$). If it were the case that punishment, where top down accountability occurs, was contributing to an improvement *beyond* the pre-existing path of recovery, we would expect the coefficient in period $t = 2$ to be larger than the coefficient in period $t = 1$ as the path to recovery would have accelerated.

We find evidence for the efficacy of punishment only in the case of teacher presence (panel a), where there is a small precisely estimated effect on the first differenced outcome (p-value of 0.06). This shows that the rate at which teachers return to schools is increased in the first month after flagging by 2 percentage points. From month 2 onwards, we see no difference between flagged and non-flagged schools. The results for flagging on other outcomes are all indistinguishable from zero, suggesting that punishment is not bringing any further improvement in the rate of recovery. Taken together, these results show that there is an impact of top-down accountability but it is small in magnitude and only occurs on the immediate next step on the causal chain.

3.5 Impacts on the machinery of government

Despite finding no impacts of the centralized accountability scheme on schooling outcomes, we can use the data we have collected to investigate if there were effects on other bureaucratic activities that we would expect to observe if the bureaucracy had been motivated to respond to the flagging. Specifically, we can analyze administrative action in terms of both personnel and financial resources, the two key inputs to effective government functioning. We look at bureaucratic effort through monitoring visits to affected schools, the movement of staff, and impacts on promotions. We also look at changes in school budgetary resources and the nature of expenditures at the school level.

3.5.1 Oversight visits

A natural immediate response by public officials flagged for poor performance would be to visit poorly performing schools to undertake diagnostic and remedial work on whatever area of school functioning had been flagged. School visits are a standard part of the AEOs work program and a mechanism to resolve issues that schools face in functioning effectively. We explore whether the flagging led to an increase in visits to (affected) schools. Table C.2.5 in the Appendix reports the results of each flagging on this measure of bureaucratic effort. The 'visited schools' measure equals 1 if the school received a visit from the relevant AEO. The coefficients of *Flag* and *Punish* periods account for changes in the probability of receiving a visit, given the negative shock. The flagging has no significant effects on bureaucratic visits to schools. The coefficients for the *After Flag* period are never significantly positive for both samples in none of the variables. Table C.2.6 shows the results in samples with specific characteristics to explore if bureaucrats gamed the system by strategically visiting bigger, worst performing, or most missing teachers schools. The results are small or non-significant.

3.5.2 School Budget Utilization

Another response by public officials is to channel budgetary resources to support struggling schools. We explore the relationship between flagging and the schools' resources by aggregating the panel at the year level and counting the number of times each school was in a flagged markaz. We used a panel regression with markaz and year fixed effects, and district-time trends to obtain estimates of the impact of the number of times flagged in a year on the amount of funds received and the expenditures undertaken by the schools in the next year. Further, to address potential endogeneity from resources assignment also affecting the flagging status, we use an instrumental variables approach and exploit the random position of the markaz around the arbitrary threshold. We instrument the number of times flagged by whether the schools were in a markaz flagged when staying within the threshold sample.

Panel A of Table C.2.7 in the appendix shows the results for each flag type on the amount of funding given by the government and the reported expenses at the school level for a year. For teacher presence, one more flag in the previous year is associated with an increase of 6% in non-government funds (those received from non-government sources such as parents), and one more flag on functional facilities increases non-government funds by 3%. For student attendance, one more flag leads to a 7% increase in government funds received by the school and a 3% rise in expenditures. The rest of the coefficients are small in magnitude and broadly insignificant. Panel B reports the results from an IV estimation strategy where we instrument the number of times flagged by the distance to the flagging threshold, which is akin to fuzzy RD setup. In this setup, we find no evidence of a response to any of the flagging in either the funds received by a school or its expenditures. Overall, there seems little systematic evidence that the flagging shifts budgetary resources or expenditures.

3.5.3 Transfers and Postings

Public officials can also intervene in the management of schools through the labor market by moving head teachers, or district officials across schools or districts in response to flagging. We study the rotation of officials at the school and district level, measuring rotation as a variable that equals 1 if the public official reported in period t is different from the one reported in $t-1$. First, we explored whether the markaz flagging induced a higher rotation of head teachers, as AEOs might use it to improve school performance within their administrative unit. We used equation 3.1 with rotation of head teachers as a dependent variable. We thus estimated the effect of being flagged on the probability of observing head teacher rotation. Overall, Appendix Table C.2.8 shows no significant changes in the probability of rotation of head teachers, except from math and english scores, for which we found a lower probability of rotation in the after-flag period.

Second, we used equation 3.2 at the district level to observe the rotation in district managers themselves. Because the district officer is a district attribute, we aggregated the data at the district level. Panel A of Appendix Table C.2.9 reports the results for bottom-performing district, and Panel B for a top-performing district. We bootstrapped the standard errors because of the low number of observations. No coefficient showed significant results, suggesting that the district flagging system based on rankings does not lead to a higher rotation of officers.

Finally, we explored for the district-level officers whether being in charge of a top/bottom district was related to whether they held a higher/lower-ranked position at the end of the scheme. In other words, whether the success of the districts in which they were in charge had any impact on their long-term progress through public service. We obtained data on the current employment of public officers in charge of a district between 2011 and 2015 and generated a ranking of the importance and status of each role. Appendix C.1.5 details how we constructed the rankings of district officer positions. We also calculated the months they were in charge of a top/bottom district. Then, we estimated a simple regression correlating the ranking of the current employment and the number of months they were in charge.[17] No coefficient is significant. However, we do have a relatively small number of observations and observe that the bottom (top) districts are negatively (positively) correlated with the rank of the current position of the public official.

Overall, there is no consistent evidence that the central accountability scheme induced any substantive impacts on how the government functioned in bureaucratic effort, budget, or public sector labor market, a result consistent with the null impacts that the scheme had on the targeted variables.

3.6 Discussion

Centralized command of the public administration, typically with few related changes in the de jure incentive structure, has been a dominant approach to the management of the public sector (Finer (1997); Education Commission (2023)). The rise of public service digital information systems has brought greater attention to the efficacy of this approach. As centralized analytical units have fed substantial volumes of data to senior managers, governments have been keen to showcase their responsiveness to this data through top-down methods of controlling service delivery. Despite the prevalence of this approach to managing government throughout history, as well as its continued implementation at scale worldwide, there have been limited evaluations to date on its efficacy.

We analyze the effectiveness of 'command and control' in government administration by evaluating a system from Punjab province in Pakistan that alerted senior government managers to poorly performing school districts. Despite flagging of poor performance leading to de facto accountability along the bureaucratic hierarchy, the scheme had no substantive impacts on schooling outcomes across any targeted outcome. By assessing the activities of public officials throughout the chain of service delivery, we find that this system had no impact on any aspect of government functioning beyond a slightly faster return of teachers to schools flagged as having low teacher attendance. Our data allow us to make these claims

[17] Regressions use bootstrapped standard errors to account for the low number of observations.

with a high degree of precision. Taken together, our results suggest that centralized command and control management approaches struggle to effectively manage unpredictable delivery environments. Such findings are consistent with emerging literature on large-scale incentive provision in the public service (see introduction).

An obvious caveat to our findings is that de jure incentives were not changed, and thus it could be argued that we would not expect to see responses by rational economic actors. However, widespread literature on the personnel economics of the state has documented the challenges to sustained changes in formal public sector contracts (Banerjee et al. (2021)) and the dominance of de facto public sector incentive schemes implemented in reality (Schuster et al. (2023)). As such, a frontier of that literature is to understand how de facto incentives (such as top-down accountability) may or may not improve service delivery outcomes.

Our identification strategy focuses on the impacts of the flagging of underperforming schools. There may have been larger benefits of the scheme, such as an immediate accountability effect or wider learning across the system upon its introduction. However, assessing the immediate impacts of the scheme using a range of approaches, we also fail to find evidence that its introduction substantially shifted outcomes. We also do not see any broad shift in the ranking of districts across the province, such that any learning did not improve the performance of the weakest performers. Rather, the relative rankings of school performance persisted. More broadly, a threshold-based approach to performance measurement is unlikely to be the most relevant method for a system to maximize learning, given its narrow lens. Alternative reporting based on the same data may have captured relative progress better.

What do these findings imply for large-scale data collection in the public sector? However detailed data-collection, management information systems struggle to document the full extent of many modern public service environments. As such, there have long been calls for autonomy for effective frontline service managers in related literatures (Simon (1983); Dixit (2002)). At the same time, large-scale datasets combined with modern analytics have been shown to be a powerful means for estimating important structural elements of the public sector production function (Fenizia (2022); Best, Hjort and Szakonyi (2017)). This would suggest that there is utility from taking an approach that builds on the comparative advantage of large-scale data analysis in estimating more permanent parameters of the education production function rather than variables that are potentially vulnerable to short-run stochastic shocks.

As an illustration of the power of large-scale data in the case of Punjab, we use the PMIU data to estimate the impacts of head teacher quality on the same outcomes that the centralized accountability system focused on. We follow Fenizia (2022) in using an AKM-model (Abowd, Kramarz and Margolis (1999)) of head-teacher productivity (Card, Heining and Kline (2013)). We identify three important insights. First, head teachers have different levels of added-value across distinct areas of school functioning, with some better at inducing teacher presence, and others better at improving test scores. Second, the rotation of head teachers across schools can have substantial impacts on school outcomes. Overall, a one standard deviation increase in head teacher quality accounts for approximately 3%

improvement in the corresponding outcome of the average school in our sample. Third, by using this information to optimally allocate head teachers to schools that are most in need of a particular set of skills, we find that PMIU could have raised levels of teacher presence by 19 percentage points in schools that were performing below the median on that margin.

Combining this illustrative analysis with our results indicates that centralized management of service-delivery through high-frequency monitoring and related control methods is an inefficient use of information management systems in the public sector. We find no evidence that this statement is mediated by features of the targeted outcome, with the 'command and control' scheme we study having no impacts along any point on the causal chain: from monitoring and budgetary allocation, facility construction and maintenance, to student and teacher presence at schools. However, insights using the data resulting from high-frequency monitoring can be powerfully used to identify structural parameters of the education production function that no official within the public administration could generate independently.

Moreover, complementing large-scale and high-frequency data collection with appropriate counterfactual analytics ensures that limited public resources are spent judiciously. We estimate, using only that data which PMIU would have had access to during the rollout of the scheme, that the limited impacts of the system could have been detected within months of it starting. Figure C.2.12 in the Appendix plots the after-flag β coefficients from equation 3.1 using only the data available up to month t.[18] As such, we mimic the analysis that the government could have undertaken during the scheme's operation.[19] The results are a long string of null or negative coefficients that would have been quickly perceptible to an analyst. Financial and personnel resources, and the attention paid to the scheme, could have been repurposed to other, potentially more effective, policies.

In conclusion, our paper provides a detailed evaluation of the concerns with centralized accountability systems debated in the literature (Kane and Staiger (2002); Besley and Coate (2003); Bardhan (2002); Bó et al. (2021)). Our results support the perspective that oversight and control approaches fail to induce changes throughout a public sector hierarchy. However, re-purposing the data that underlies an oversight scheme for analytical purposes related to structural determinants of public sector effectiveness has much greater promise (Lang (2010); Staiger and Rockoff (2010)).

[18] In the first month, we use data from the first month only. In the second, we use data from the first two months, and so on.

[19] We omit the results for school scores due to the short time series available for these variables.

3.7 Main Tables and Figures

3.7.1 Tables

Table 3.1: Descriptive statistics

Panel A: School-level variables								
	Mean	Median	Sd	N. Obs	Mean	Median	Sd	N. Obs
Number of teachers	4.6	3	3.8	2,305,208	.	.	.	.
Number of students	110	80	103	2,307,637	.	.	.	.
Outcomes (%)		**No flag**				**Flag**		
Teacher presence	93	100	15	2,095,004	83	100	22	209,599
Student attendance	90	93	12	1,899,734	81	85	17	403,409
Functional facilities	93	100	16	1,875,892	84	100	22	383,125
Math score	87	92	14	824,341	67	67	21	22,212
English score	80	83	17	659,293	65	67	20	187,236
Urdu score	85	89	15	810,220	67	67	20	36,291
Panel B: Markaz-level variables								
	Mean	Median	Sd	N. Obs	Mean	Median	Sd	N. Obs
Number of schools	21	15	19	130,364	.	.	.	.
Proportion elementary	80	100	40	130,364	.	.	.	.
Outcomes (%)		**No flag**				**Flag**		
Teacher presence	93	94	4.3	95,422	80	83	7.8	8,739
Student attendance	91	92	6	89,649	80	82	7.7	14,257
Functional facilities	95	98	11	90,029	81	84	11	13,846
Math score	87	88	6.4	60,069	65	66	4.9	2,100
English score	80	80	6.3	49,451	64	66	5.1	12,718
Urdu score	85	86	6.4	59,375	65	67	5.1	2,794
Panel C: District level variables								
		Top 5				**Bottom 5**		
Outcomes (%)	Mean	Median	Sd	N. Obs	Mean	Median	Sd	N. Obs
Overall score	94	95	3.8	70	78	78	10	70
New position	.077	0	.27	504	.083	0	.28	504

Notes: The unit for outcomes in Panel A is outcome-school-month; in Panel B it is outcome-markaz-month. Outcomes are measured in percentages. Student test scores are measured as the percentage of correct answers in standardized tests. A unit is flagged if it receives a flag in the data pack on that outcome in that month. Outcomes in Panel B correspond to the maraakiz that had elementary schools for which an AEO can be flagged. Panel C reports statistics at the district-quarter level. The "Overall score" is the weighted average of markaz outcomes for a district for the three months before the meeting for those ranked at the top/bottom in the respective meeting. The "New position" variable measures the percentage of districts that enter into the top/bottom in each quarterly meeting.

Table 3.2: Measures of Variation

School-level variation (sd) by quintiles of overall performance							
Outcomes (%)	Q1	Q2	Q3	Q4	Q5	All	N.Obs
Teacher presence	9.8	.96	.67	.72	1.5	7.4	51,534
Student attendance	13	1.3	.77	.69	1.5	9.6	51,507
Functional facilities	17	4.5	1.7	.46	.64	16	50,501
Math score	5.4	1.1	.81	.79	1.8	6.3	37,537
English score	5.8	1.4	1.1	1.2	3	8.2	37,536
Urdu score	5.5	1.3	.93	.92	2	7	37,536

Notes: The unit of observation for outcomes is presented at the school level. Teacher presence and student attendance are measured as the percentage of present teachers/students relative to the total teachers/students reported. The functional facilities variable is measured as the percentage of the school's functional infrastructure, including toilets, drinking water, boundary wall, and electricity. Math, English, and Urdu scores are measured as the percentage of correct answers in standardized tests. Each quintile is calculated separately based on the mean level of performance for each variable. The table shows the standard deviation for each school-level variable quintile.

Table 3.3: Monitoring effect on performance - markaz flagging

Panel A: School outcomes						
Dependent variable:	Teacher presence		Student attendance		Functional facilities	
T×Flag	-6.51***	-1.81***	-7.13***	-2.11***	-4.73***	-1.39***
	(0.15)	(0.19)	(0.18)	(0.16)	(0.34)	(0.28)
T×Punish	-2.71***	-1.63***	-3.06***	-2.24***	-1.19***	-0.66**
	(0.18)	(0.27)	(0.19)	(0.29)	(0.18)	(0.29)
T×After flag	-0.40***	-0.47***	-0.98***	-1.18***	-0.43***	-0.23
	(0.098)	(0.16)	(0.081)	(0.15)	(0.16)	(0.29)
N. of obs.	6,979,566	490,950	4,964,842	562,661	7,314,616	392,052
Mean Dep. Var. before	92.9	87.3	91.8	87.2	97.4	93.9
R^2	0.032	0.036	0.10	0.098	0.070	0.069
Panel B: Student scores						
Dependent variable:	Math		English		Urdu	
T×Flag	-13.7***	-2.74***	-10.3***	-2.34***	-10.7***	-2.63***
	(0.35)	(0.56)	(0.22)	(0.30)	(0.26)	(0.34)
T×Punish	-2.31***	-1.11	-3.34***	-2.47***	-1.48***	-0.46
	(0.46)	(0.82)	(0.28)	(0.46)	(0.36)	(0.55)
T×After flag	-0.14	-0.28	-1.53***	-1.55***	-0.087	-0.61*
	(0.28)	(0.52)	(0.20)	(0.30)	(0.22)	(0.35)
N. of obs.	2,182,972	53,066	804,855	146,692	1,936,332	119,016
Mean Dep. Var. before	86.9	71.7	78.1	70.4	84.5	71.7
R^2	0.100	0.15	0.065	0.069	0.10	0.13
Sample	Full	Threshold	Full	Threshold	Full	Threshold
Markaz FE	Yes	Yes	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes	Yes	Yes
District time trends	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Results from estimating equation 3.1. The school is the unit of observation for both panels. The first column for each outcome estimates for the full sample. The second column for each outcome estimates for the threshold sample, including schools in maraakiz that lie within the bandwidth obtained through regression discontinuity optimization methods. The flagging and threshold sample are based on the studied outcome. Teacher presence and student attendance are measured as the percentage of present teachers/students relative to the total teachers/students reported. The functional facilities variable is measured as the percentage of the school's functional infrastructure, including toilets, drinking water, boundary wall, and electricity. Math, English, and Urdu scores are measured as the percentage of correct answers in standardized tests. T equals 1 for schools in a flagged markaz. *Flag* equals 1 for the period in which the information is collected, and the markaz is flagged. *Punish* equals 1 for the period where the reports are distributed and the oversight meeting with the punishment occurs. *After flag* is equal to 1 for periods after the oversight meeting occurs. *Mean. Dep. Var before* shows the average outcome in the non-flagged maraakiz before the flagging occurs. Standard errors clustered by markaz, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

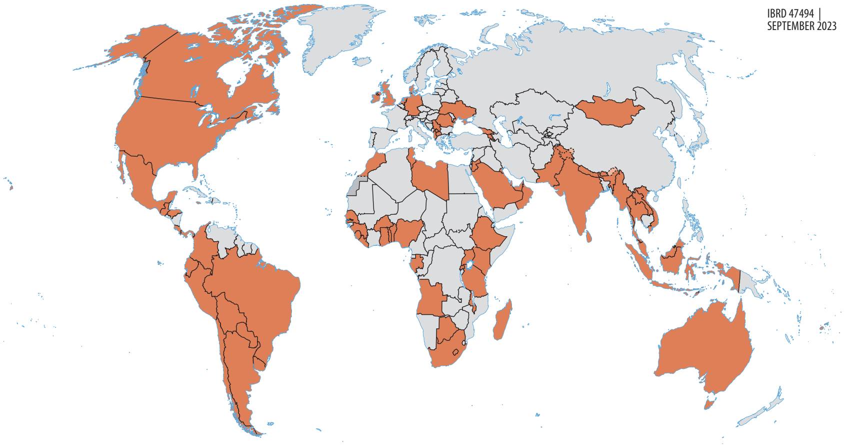
Table 3.4: Monitoring effect on performance - district ranking

Panel A: Bottom districts						
Dependent variable:	Teacher presence		Student attendance		Functional facilities	
Bottom×Meeting	-0.11	0.18	0.46	1.28*	-0.045	0.087
	(0.31)	(0.34)	(0.68)	(0.72)	(0.51)	(0.54)
Bottom×After meeting	0.38	0.27	0.72	1.31**	0.48	0.20
	(0.33)	(0.36)	(0.56)	(0.61)	(0.52)	(0.53)
N. of obs.	3,063,835	583,417	3,063,410	583,248	3,009,844	565,920
Mean Dep. Var. before	91.4	90.1	88.8	86.0	92.5	90.0
R^2	0.025	0.030	0.12	0.15	0.14	0.17
Panel B: Top districts						
Dependent variable:	Teacher presence		Student attendance		Functional facilities	
Top×Meeting	1.19***	0.71*	-0.76	-1.32*	0.43	0.29
	(0.30)	(0.38)	(0.51)	(0.73)	(0.30)	(0.28)
Top×After meeting	0.79***	0.82***	0.089	-0.50	0.073	0.66
	(0.25)	(0.23)	(0.31)	(0.68)	(0.42)	(0.46)
N. of obs.	3,111,642	682,461	3,111,048	682,369	3,036,557	672,780
Mean Dep. Var. before	91.0	91.9	87.4	90.0	91.6	92.7
R^2	0.027	0.026	0.12	0.12	0.14	0.15
Sample	Full	Threshold	Full	Threshold	Full	Threshold
District FE	Yes	Yes	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes	Yes	Yes

Note: Results from estimating equation 3.2. The school is the unit of observation for both panels. The first column for each outcome estimates for the full sample. The second column for each outcome estimates for the threshold sample, including the schools in the five districts closer to the five in the bottom/top. The bottom/top status and threshold sample are based on the aggregate district performance. Teacher presence and student attendance are measured as the percentage of present teachers/students relative to the total teachers/students reported. The functional facilities variable is measured as the percentage of the school's functional infrastructure, including toilets, drinking water, boundary wall, and electricity. *Bottom* equals 1 for schools in the bottom five districts and Top equals 1 for the schools in the top five districts on the date of the quarterly meeting. *Meeting* equals 1 in the period of the quarterly meeting. *Mean. Dep. Var before* shows the average outcome in the non-top/bottom districts before the meeting occurs. Standard errors clustered by district, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

3.7.2 Figures

Figure 3.1: Countries adopting the command-and-control delivery approach (shaded)



Source: Mansoor et al. (2023)

Figure 3.2: Monitoring scheme structure

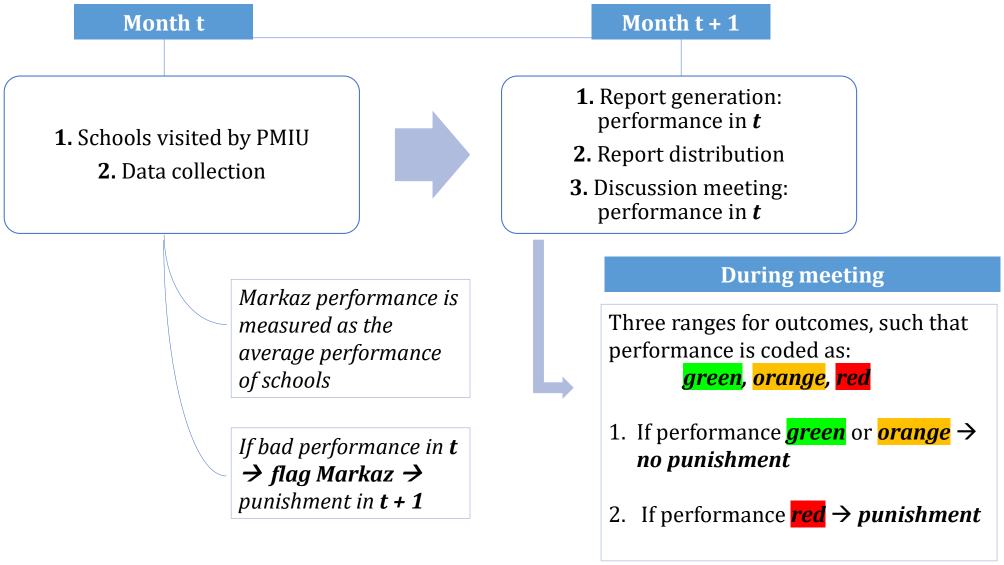
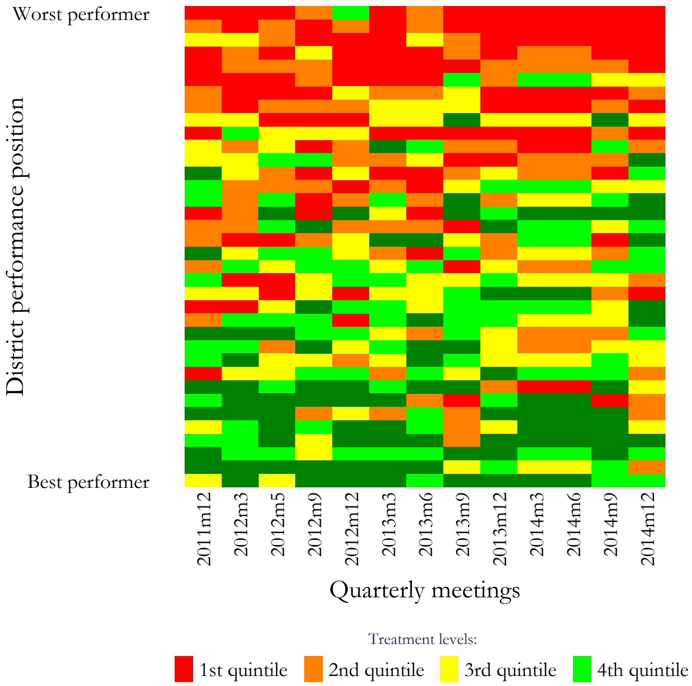
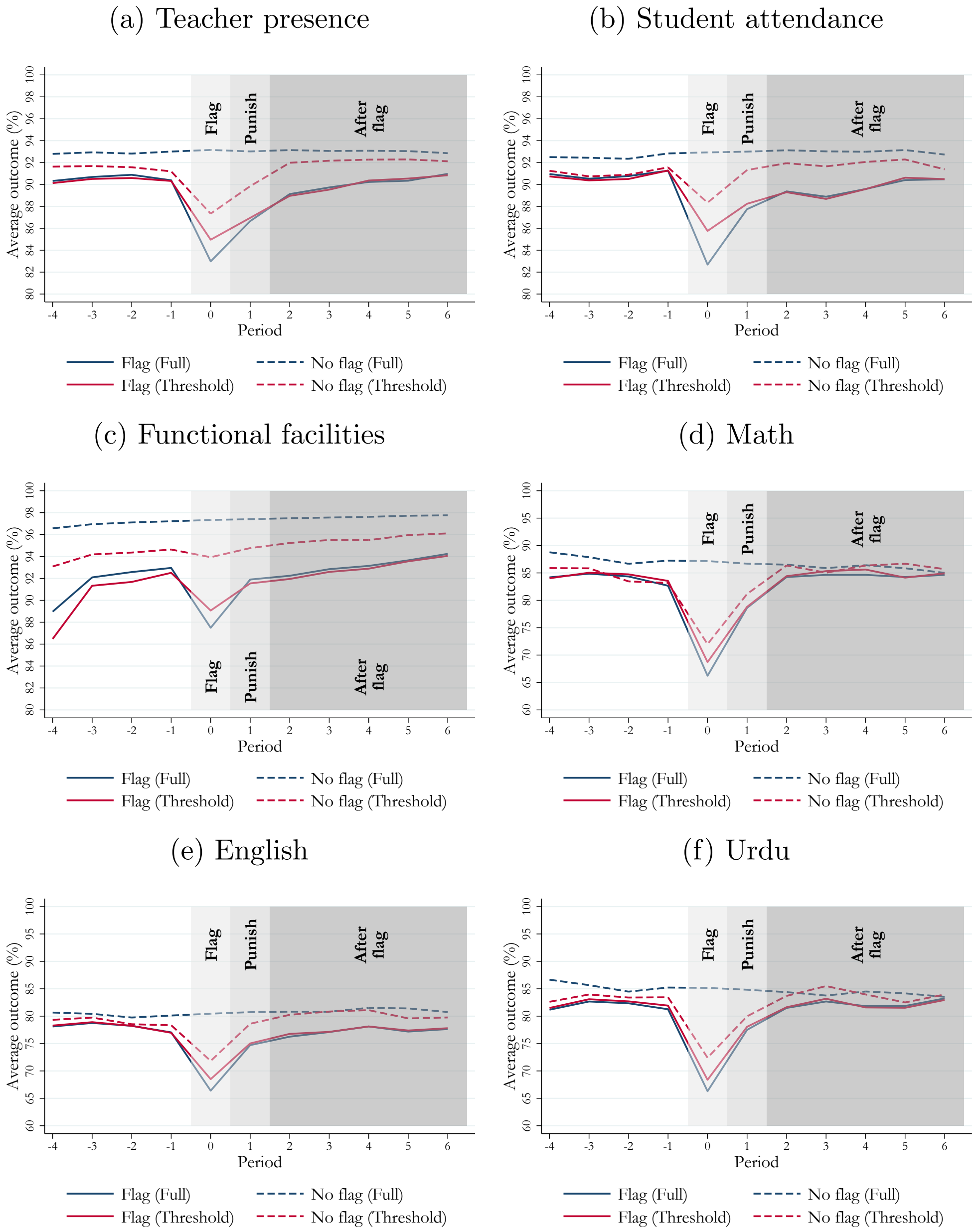

Month t
Month t + 1
1. Schools visited by PMIU
2. Data collection
1. Report generation: performance in ***t***
2. Report distribution
3. Discussion meeting: performance in ***t***
Markaz performance is measured as the average performance of schools
*If bad performance in **t** → **flag Markaz** → punishment in **t + 1***
During meeting
Three ranges for outcomes, such that performance is coded as:
green, ***orange***, ***red***
1. If performance ***green*** or ***orange*** → ***no punishment***
2. If performance ***red*** → ***punishment***

Figure 3.3: Distribution of quintiles of district performance



Note: This figure illustrates for each quarter the quintile of the overall district score distribution in which each district fell. District scores are measured based on the aggregate performance of teacher presence, student attendance, and functional facilities in each quarter. The figure ranks the districts based on their average performance of all the periods, such that the worst performing district at all times appears first.

Figure 3.4: Evolution of school outcomes in relative time - markaz flagging



(a) Teacher presence

(b) Student attendance

(c) Functional facilities

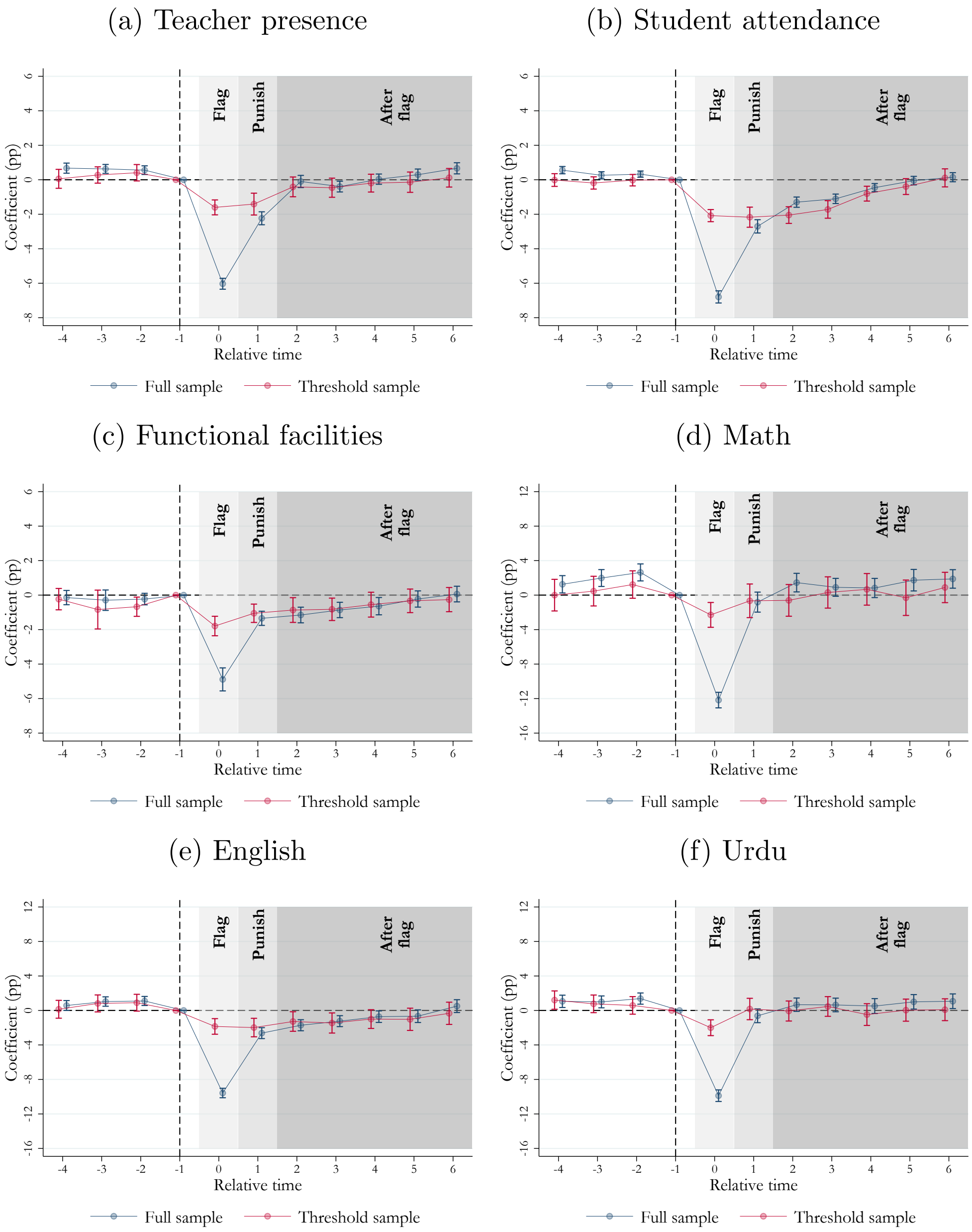
(d) Math

(e) English

(f) Urdu

Note: The figure presents the average evolution of schools in flagged (continuous line) and non-flagged (dashed line) maraakiz. Flagging is based on the outcome variable in focus. Blue lines represent the full sample. Red accounts for the threshold sample that is "close" to the flagging threshold. Relative time is divided into: *Flag*: period where information is collected and maraakiz are flagged; *Punish*: period where the reports are distributed and oversight meetings are held; *After flag*: periods after the meeting.

Figure 3.5: Event study - flagging effect on performance



(a) Teacher presence

(b) Student attendance

(c) Functional facilities

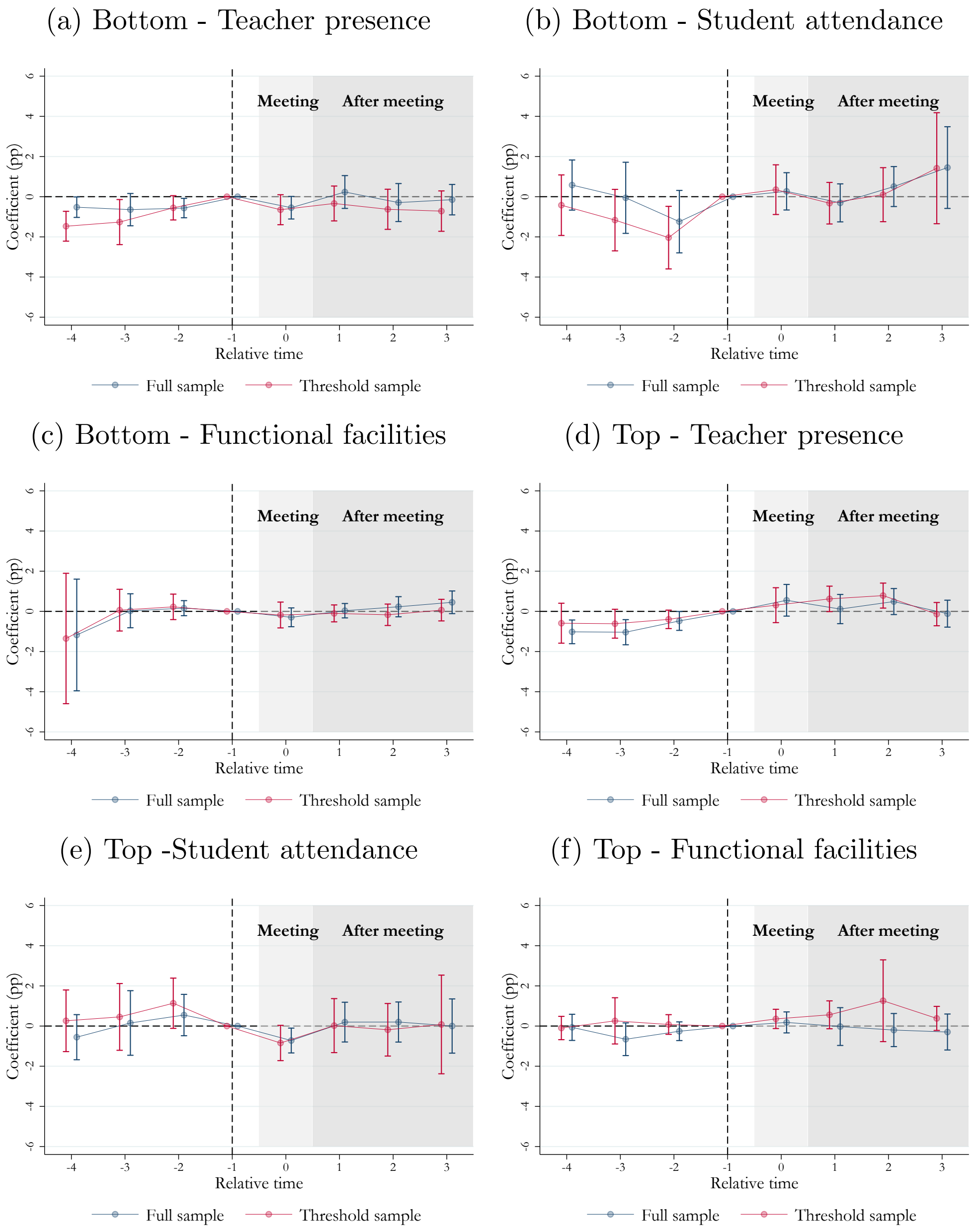
(d) Math

(e) English

(f) Urdu

Note: This figure presents results from estimating event studies based on equation 3.1 using -1 as the base period, comparing schools in flagged and non-flagged maraakiz. The blue line presents results for the full sample, while the red line presents results for the threshold sample, obtained through regression discontinuity optimization methods. The results are for flagging on the variable in the title of the panel. *Flag* is for the period in which the information is collected, and the markaz is flagged. *Punish* is for the period where the reports are distributed and the oversight meeting with the punishment occurs. *After flag* is for periods after the oversight meeting occurs. Error bars at the 95 percent level are presented for each coefficient.

Figure 3.6: Event study - district ranking effect on performance



(a) Bottom - Teacher presence

(b) Bottom - Student attendance

(c) Bottom - Functional facilities

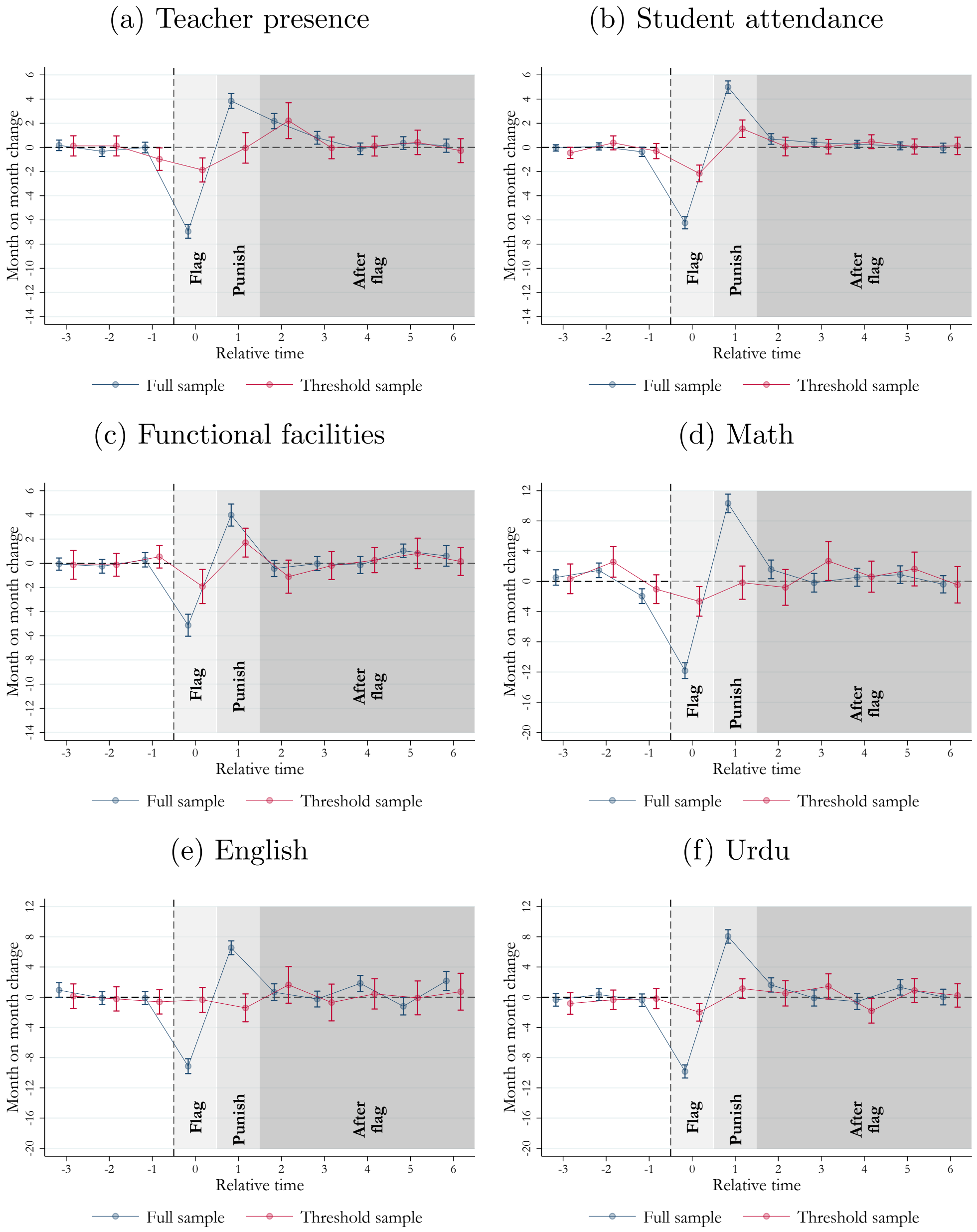
(d) Top - Teacher presence

(e) Top -Student attendance

(f) Top - Functional facilities

Note: This figure presents the results from estimating an event-study based on equation 3.2, using -1 as base period, comparing schools in top/bottom districts against schools out of the top/bottom districts. Bottom is for the schools in bottom five districts in the quarterly meeting. Top is for the schools in the Top five districts in the quarterly meeting. Blue line accounts for the result on the full sample, while the red accounts for the results using the threshold sample, including the schools in the five districts closer to the five in the bottom/top. *Meeting* is for the period of the quarterly meeting. Error bars at the 95 percent level are presented for each coefficient.

Figure 3.7: Punishment Period vs Reversion to Mean - Month on Month Changes



(a) Teacher presence

(b) Student attendance

(c) Functional facilities

(d) Math

(e) English

(f) Urdu

Note: This figure shows month-by-month coefficients from equation 3.1 for maraakiz not fully recovered from the negative shock during the punishment period. It compares schools in flagged and non-flagged maraakiz across consecutive months. The blue and red lines represent results for the full and threshold samples, respectively. Panel title shows flagging and dependent variable. *Flag* is for the period in which the information is collected, and the markaz is flagged. *Punish* is for the period where the reports are distributed and the oversight meeting with the punishment occurs. *After flag* is for periods after the oversight meeting occurs. Error bars represent 95% confidence intervals. P-values reported for a one-sided test for the coefficient of relative time 2 (after flag) being greater than that of relative time 1 (punishment) in the threshold sample: Panel (a) 0.06, Panel (b) 0.99, Panel (c) 0.99, Panel (d) 0.62, Panel (e) 0.08, Panel (f) 0.6.

Bibliography

Abowd, John M, Francis Kramarz and David N Margolis. 1999. “High wage workers and high wage firms.” *Econometrica* 67(2):251–333.

Aghion, Philippe and Jean Tirole. 1997. “Formal and Real Authority in Organizations.” *Journal of Political Economy* 105(1):1–29.
URL: *http://www.jstor.org/stable/2138869*

Ali, Aisha J, Javier Fuenzalida, Margarita Gómez and Martin J Williams. 2021. “Four lenses on people management in the public sector: an evidence review and synthesis.” *Oxford Review of Economic Policy* 37(2):335–366.
URL: *https://ideas.repec.org/a/oup/oxford/v37y2021i2p335-366..html*

Aneja, Abhay and Guo Xu. 2023. “Strengthening State Capacity: Civil Service Reform and Public Sector Performance during the Gilded Age.”.

Arifu: WhatsApp Chatbot Provides Tips for Micro-Retailers. N.d. https://strivecommunity.org/programs/arifu.

Arráiz, Irani, Syon Bhanot and Carla Calero. 2019. Less Is More: Experimental Evidence on Heuristics-Based Business Training in Ecuador. Technical report IDB Invest.

Ash, Elliott and W. Bentley MacLeod. 2015. “Intrinsic Motivation in Public Service: Theory and Evidence from State Supreme Courts.” *The Journal of Law and Economics* 58(4):863–913.
URL: *https://doi.org/10.1086/684293*

Ashenfelter, Orley. 1978. “Estimating the effect of training programs on earnings.” *The Review of Economics and Statistics* pp. 47–57.

Ashenfelter, Orley C and David Card. 1984. “Using the longitudinal structure of earnings to estimate the effect of training programs.”.

Ashraf, Nava, Oriana Bandiera and B Kelsey Jack. 2014. “No margin, no mission? A field experiment on incentives for public service delivery.” *Journal of public economics* 120:1–17.

Bai, Liang, Benjamin Handel, Edward Miguel and Gautam Rao. 2021. "Self-control and demand for preventive health: Evidence from hypertension in India." *Review of Economics and Statistics* 103(5):835–856.

Baker, Andrew C, David F Larcker and Charles CY Wang. 2022. "How much should we trust staggered difference-in-differences estimates?" *Journal of Financial Economics* 144(2):370–395.

Bandiera, Oriana, Michael Carlos Best, Adnan Qadir Khan and Andrea Prat. 2021. "The Allocation of Authority in Organizations: A Field Experiment with Bureaucrats*." *The Quarterly Journal of Economics* 136(4):2195–2242.
URL: *https://doi.org/10.1093/qje/qjab029*

Banerjee, Abhijit, Raghabendra Chattopadhyay, Esther Duflo, Daniel Keniston and Nina Singh. 2021. "Improving Police Performance in Rajasthan, India: Experimental Evidence on Incentives, Managerial Autonomy, and Training." *American Economic Journal: Economic Policy* 13(1):36–66.
URL: *https://www.aeaweb.org/articles?id=10.1257/pol.20190664*

Banerjee, Abhijit V., Esther Duflo and Rachel Glennerster. 2008. "Putting a Band-Aid on a Corpse: Incentives for Nurses in the Indian Public Health Care System." *Journal of the European Economic Association* 6(2-3):487–500.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1162/JEEA.2008.6.2-3.487*

Barber, Michael. 2013. The Good News from Pakistan. Technical report Reform, London.

Bardhan, Pranab. 2002. "Decentralization of Governance and Development." *Journal of Economic Perspectives* 16(4):185–205.
URL: *https://www.aeaweb.org/articles?id=10.1257/089533002320951037*

Becker, Gordon M., Morris H. Degroot and Jacob Marschak. 1964. "Measuring Utility by a Single-Response Sequential Method." *Behavioral Science* 9(3):226–232.

Bertrand, Marianne, Robin Burgess, Arunish Chawla and Guo Xu. 2020. "The glittering prizes: Career incentives and bureaucrat performance." *The Review of Economic Studies* 87(2):626–655.

Besley, Timothy, Robin Burgess, Adnan Khan and Guo Xu. 2022. "Bureaucracy and Development." *Annual Review of Economics* 14(1):397–424.
URL: *https://doi.org/10.1146/annurev-economics-080521-011950*

Besley, Timothy and Stephen Coate. 2003. "Centralized versus decentralized provision of local public goods: a political economy approach." *Journal of Public Economics* 87(12):2611–2637.
URL: *https://doi.org/10.1016/s0047-2727(02)00141-x*

Best, Michael Carlos, Jonas Hjort and David Szakonyi. 2017. Individuals and organizations as sources of state effectiveness. Technical report National Bureau of Economic Research.

Blattman, Christopher and Laura Ralston. 2015. "Generating Employment in Poor and Fragile States: Evidence from Labor Market and Entrepreneurship Programs.".

Bloom, Nicholas, Aprajit Mahajan, David McKenzie and John Roberts. 2010. "Why Do Firms in Developing Countries Have Low Productivity?" *American Economic Review* 100(2):619–623.

Bloom, Nicholas, Benn Eifert, Aprajit Mahajan, David McKenzie and John Roberts. 2013. "Does Management Matter? Evidence from India." *The Quarterly Journal of Economics* 128(1):1–51.

Bloom, Nicholas and John Van Reenen. 2010*a*. "Why Do Management Practices Differ across Firms and Countries?" *Journal of Economic Perspectives* 24(1):203–224.

Bloom, Nicholas and John Van Reenen. 2010*b*. "Why Do Management Practices Differ across Firms and Countries?" *Journal of Economic Perspectives* 24(1):203–24.
URL: *https://www.aeaweb.org/articles?id=10.1257/jep.24.1.203*

Bloom, Nicholas, Renata Lemos, Raffaella Sadun and John Van Reenen. 2015. "Does Management Matter in schools?" *The Economic Journal* 125(584):647–674.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/ecoj.12267*

Bó, Ernesto Dal, Frederico Finan, Nicholas Y. Li and Laura Schechter. 2021. "Information Technology and Government Decentralization: Experimental Evidence From Paraguay." *Econometrica* 89(2):677–701.
URL: *https://doi.org/10.3982/ecta17497*

Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Alice Ng'ang'a and Justin Sandefur. 2018. "Experimental evidence on scaling up education reforms in Kenya." *Journal of Public Economics* 168:1–20.
URL: *https://www.sciencedirect.com/science/article/pii/S0047272718301518*

Bruhn, Miriam, Dean Karlan and Antoinette Schoar. 2010. "What Capital Is Missing in Developing Countries?" *American Economic Review* 100(2):629–633.

Bruhn, Miriam, Dean Karlan and Antoinette Schoar. 2018. "The Impact of Consulting Services on Small and Medium Enterprises: Evidence from a Randomized Trial in Mexico." *Journal of Political Economy* 126(2):635–687.

Business Edge : Status and Disposition. 2006. Technical Report 4 World Bank Group Washington, DC: .

Callaway, Brantly and Pedro HC Sant'Anna. 2021. "Difference-in-differences with multiple time periods." *Journal of Econometrics* 225(2):200–230.

Callen, Michael, Saad Gulzar, Ali Hasanain, Muhammad Yasir Khan and Arman Rezaee. 2020. "Data and policy decisions: Experimental evidence from Pakistan." *Journal of Development Economics* 146:102523.

Callen, Michael, Saad Gulzar and Arman Rezaee. 2020. "Can political alignment be costly?" *The Journal of Politics* 82(2):612–626.

Calonico, Sebastian, Matias D Cattaneo and Max H Farrell. 2020. "Optimal bandwidth choice for robust bias-corrected inference in regression discontinuity designs." *The Econometrics Journal* 23(2):192–210.

Campos, Francisco, Michael Frese, Markus Goldstein, Leonardo Iacovone, Hillary C. Johnson, David McKenzie and Mona Mensmann. 2017. "Teaching Personal Initiative Beats Traditional Training in Boosting Small Business in West Africa." *Science* 357(6357):1287–1290.

Card, David, Jörg Heining and Patrick Kline. 2013. "Workplace heterogeneity and the rise of West German wage inequality." *The Quarterly journal of economics* 128(3):967–1015.

Carreri, Maria. 2021. "Can good politicians compensate for bad institutions? Evidence from an original survey of Italian mayors." *The Journal of Politics* 83(4):1229–1245.

Cengiz, Doruk, Arindrajit Dube, Attila Lindner and Ben Zipperer. 2019. "The effect of minimum wages on low-wage jobs." *The Quarterly Journal of Economics* 134(3):1405–1454.

Chaudhry, Rastee and Abdullah Waqar Tajwar. 2021. *The Punjab Schools Reform Roadmap: A Medium-Term Evaluation*. Cham: Springer International Publishing pp. 109–128.
URL: *https://doi.org/10.1007/978-3-030-57039-2*$_5$

Chay, Kenneth Y., Patrick J. McEwan and Miguel Urquiola. 2005. "The Central Role of Noise in Evaluating Interventions That Use Test Scores to Rank Schools." *American Economic Review* 95(4):1237–1258.
URL: *https://www.aeaweb.org/articles?id=10.1257/0002828054825529*

Chen, Cheng. 2017. "Management Quality and Firm Hierarchy in Industry Equilibrium." *American Economic Journal: Microeconomics* 9(4):203–44.
URL: *https://www.aeaweb.org/articles?id=10.1257/mic.20160305*

Chen, Cheng and Wing Suen. 2019. "The Comparative Statics of Optimal Hierarchies." *American Economic Journal: Microeconomics* 11(2):1–25.
URL: *https://www.aeaweb.org/articles?id=10.1257/mic.20170158*

Chioda, Laura, David Contreras-Loya, Paul Gertler and Dana Carney. 2021. “Making Entrepreneurs: Returns to Training Youth in Hard Versus Soft Business Skills.”.

Cho, Yoonyoung and Maddalena Honorati. 2014. “Entrepreneurship Programs in Developing Countries: A Meta Regression Analysis.” *Labour Economics* 28(C):110–130.

Cilliers, Jacobus and James Habyarimana. 2023. “Tackling Implementation Challenges with Information: Experimental Evidence from a School Governance Reform in Tanzania.”.

Cole, Shawn Allen and A. Fernando. 2020. ‘Mobile’Izing Agricultural Advice: Technology Adoption, Diffusion, and Sustainability. SSRN Scholarly Paper ID 2179008 Social Science Research Network Rochester, NY: .

Cole, Shawn Allen, Mukta Joshi and Antoinette Schoar. 2021. “Heuristics on Call: The Impact of Mobile Phone Based Business Management Advice.”.

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

Das, Jishnu, Abhijit Chowdhury, Reshmaan Hussam and Abhijit V Banerjee. 2016. “The impact of training informal health care providers in India: A randomized controlled trial.” *Science* 354(6308):aaf7384.

Davies, Elwyn, Peter Deffebach, Leonardo Iacovone and David Mckenzie. 2023. *Training Microentrepreneurs over Zoom: Experimental Evidence from Mexico.* Policy Research Working Papers The World Bank.

De Chaisemartin, Clément and Xavier d’Haultfoeuille. 2020. “Two-way fixed effects estimators with heterogeneous treatment effects.” *American Economic Review* 110(9):2964–96.

De Chaisemartin, Clément and Xavier D’Haultfoeuille. 2022. Difference-in-differences estimators of intertemporal treatment effects. Technical report National Bureau of Economic Research.

de Oliveira, Priscila. 2023. “Why Businesses Fail: Underadoption of Improved Practices by Brazilian Micro-Enterprises.”.

de Ree, Joppe, Karthik Muralidharan, Menno Pradhan and Halsey Rogers. 2017. “Double for Nothing? Experimental Evidence on an Unconditional Teacher Salary Increase in Indonesia*.” *The Quarterly Journal of Economics* 133(2):993–1039.
URL: *https://doi.org/10.1093/qje/qjx040*

Della Vigna, Stefano and Ulrike Malmendier. 2006. “Paying not to go to the gym.” *American Economic Review* 96(3):694–719.

DellaVigna, Stefano, Devin Pope and Eva Vivalt. 2019. “Predict Science to Improve Science.” *Science* 366(6464):428–429.

Deserranno, Erika. 2019. “Financial Incentives as Signals: Experimental Evidence from the Recruitment of Village Promoters in Uganda.” *American Economic Journal: Applied Economics* 11(1):277–317.
URL: *https://www.aeaweb.org/articles?id=10.1257/app.20170670*

Deserranno, Erika, Gianmarco Leon and Philipp Kastrau. 2022. Promotions and Productivity: The Role of Meritocracy and Pay Progression in the Public Sector. Technical report Working Paper.

Deserranno, Erika, Stefano Caria, Philipp Kastrau and Gianmarco León-Ciliotta. 2022. The Allocation of Incentives in Multi-Layered Organizations. Working paper Northwestern University.

Dessein, Wouter. 2002. “Authority and Communication in Organizations.” *The Review of Economic Studies* 69(4):811–838.
URL: *http://www.jstor.org/stable/1556723*

Dhaliwal, Iqbal and Rema Hanna. 2017. “The devil is in the details: The successes and limitations of bureaucratic reform in India.” *Journal of Development Economics* 124:1–21.

Dickinson, David and Marie-Claire Villeval. 2008. “Does monitoring decrease work effort?: The complementarity between agency and crowding-out theories.” *Games and Economic behavior* 63(1):56–76.

Dixit, Avinash. 2002. “Incentives and Organizations in the Public Sector: An Interpretative Review.” *The Journal of Human Resources* 37(4):696–727.
URL: *http://www.jstor.org/stable/3069614*

Drexler, Alejandro, Greg Fischer and Antoinette Schoar. 2014. “Keeping It Simple: Financial Literacy and Rules of Thumb.” *American Economic Journal: Applied Economics* 6(2):1–31.

Duflo, Esther, Rema Hanna and Stephen P Ryan. 2012. “Incentives work: Getting teachers to come to school.” *American Economic Review* 102(4):1241–78.

Education Commission. 2023. “Deliberate Disrupters: Can Delivery Approaches Deliver Better Education Outcomes?” *Technical Report* .

Estefan, Alejandro, Martina Improta, Romina Ordoñez and Paul Winters. 2023. “Digital Training for Micro-Entrepreneurs: Experimental Evidence from Guatemala.” *The World Bank Economic Review* p. lhad029.

Fabregas, Raissa, Michael Kremer, Matthew Lowes, Robert On and Giulia Zane. 2022. “Digital Information Provision and Behavior Change: Lessons from Six RCTs in East Africa.”.

Falk, Armin and Michael Kosfeld. 2006. “The hidden costs of control.” *American Economic Review* 96(5):1611–1630.

Fenizia, Alessandra. 2022. “Managers and productivity in the public sector.” *Econometrica* 90(3):1063–1084.

Finan, Frederico, Benjamin A Olken and Rohini Pande. 2015. “The personnel economics of the state.” *Handbook of Economic Field Experiments* .

Finer, S.E. 1997. *The History of Government from the Earliest Times: Volumes I-III.* Oxford University Press, USA.
URL: *https://books.google.com/books?id=-1kqswEACAAJ*

Goodman-Bacon, Andrew. 2021. “Difference-in-differences with variation in treatment timing.” *Journal of Econometrics* 225(2):254–277.

Gouldner, Alvin W. 1960. “The Norm of Reciprocity: A Preliminary Statement.” *American Sociological Review* 25(2):161–178.

Haddad, Josette. 2022. “Training Tanzanian Farmers Through Text Messaging.”.

Heckman, James J and Jeffrey A Smith. 1999. “The pre-programme earnings dip and the determinants of participation in a social programme. Implications for simple programme evaluation strategies.” *The Economic Journal* 109(457):313–348.

Hinrichsen, Simone and Samuel Ajadi. 2020. “Using Technology to Fight COVID-19: A Spotlight on SMS-based Education Start-up, Eneza Education.”.

Hoehn, John R, Caitlin Campbell and Andrew S Bowen. 2021. Defense primer: What is command and control. Congressional Research Service. https://crsreports. congress. gov/product . . . .

Honig, Dan. 2021. “Supportive management practice and intrinsic motivation go together in the public service.” *Proceedings of the National Academy of Sciences* 118(13):e2015124118.
URL: *https://www.pnas.org/doi/abs/10.1073/pnas.2015124118*

Hussain, Iftikhar. 2015. “Subjective performance evaluation in the public sector evidence from school inspections.” *Journal of Human Resources* 50(1):189–221.

ILO. 2019. Small Matters. Technical report International Labour Office Geneva: .

International Labour Organization. 2024. “Start and Improve Your Business (SIYB).”. Accessed: 2024-08-08.
URL: *https://www.ilo.org/start-and-improve-your-business-siyb*

Kane, Thomas J and Douglas O Staiger. 2002. “The Promise and Pitfalls of Using Imprecise School Accountability Measures.” *Journal of Economic Perspectives* 16(4):91–114.
URL: *https://doi.org/10.1257/089533002320950993*

Kenya Population and Housing Census. 2019. https://www.knbs.or.ke/?p=5621.

Khan, Adnan Q., Asim Ijaz Khwaja and Benjamin A. Olken. 2019. “Making Moves Matter: Experimental Evidence on Incentivizing Bureaucrats through Performance-Based Postings.” *American Economic Review* 109(1):237–70.
URL: *https://www.aeaweb.org/articles?id=10.1257/aer.20180277*

Khan, Muhammad Yasir. 2020. “Mission Motivation and Public Sector Performance: Experimental Evidence from Pakistan.” *Unpublished manuscript* .

Lang, Kevin. 2010. “Measurement Matters: Perspectives on Education Policy from an Economist and School Board Member.” *Journal of Economic Perspectives* 24(3):167–82.
URL: *https://www.aeaweb.org/articles?id=10.1257/jep.24.3.167*

Leaver, Clare, Owen Ozier, Pieter Serneels and Andrew Zeitlin. 2021. “Recruitment, Effort, and Retention Effects of Performance Contracts for Civil Servants: Experimental Evidence from Rwandan Primary Schools.” *American Economic Review* 111(7):2213–46.
URL: *https://www.aeaweb.org/articles?id=10.1257/aer.20191972*

Lynn, Michael, Sean Masaki Flynn and Chelsea Helion. 2013. “Do Consumers Prefer Round Prices? Evidence from Pay-What-You-Want Decisions and Self-Pumped Gasoline Purchases.” *Journal of Economic Psychology* 36:96–102.

Maffioli, Alessandro, David McKenzie and Diego Ubfal. 2020. *Estimating the Demand for Business Training : Evidence from Jamaica*. Policy Research Working Papers The World Bank.

Malik, Rabea and Faisal Bari. 2022. Improving service delivery via top-down data-driven accountability: Reform enactment of the Education Road Map in Pakistan. Technical report Working Paper.

Mansoor, Zahra, Dana Qarout, Kate Anderson, Celeste Carano, Liah Yecalo-Tecle, Veronika Dvorakova and Martin J. Williams. 2023. “A Global Mapping of Delivery Approaches.” *Technical Report* .

McKenzie, David. 2020. “Small Business Training to Improve Management Practices in Developing Countries: Reassessing the Evidence for “Training Doesn’t Work”.” *Policy Research Working Paper* p. 40.

McKenzie, David and Christopher Woodruff. 2014. “What Are We Learning from Business Training and Entrepreneurship Evaluations around the Developing World?” *World Bank Research Observer* 29(1):48–82.

McKenzie, David and Christopher Woodruff. 2017. “Business Practices in Small Firms in Developing Countries.” *Management Science* 63(9):2967–2981.

Mehmood, Muhammad Zia. 2023. “Predicting effects of an SMS-based business management training.” https://socialscienceprediction.org/.

Mehmood, Sultan. 2022. “The impact of Presidential appointment of judges: Montesquieu or the Federalists?” *American Economic Journal: Applied Economics* 14(4):411–445.

Mehtha, Susanne van Lieshout-Pranati. 2017. THE NEXT 15 MILLION Start and Improve Your Business Global Tracer Study 2011-15. Publication International Labour Organization.

Micro, Small and Medium Enterprises Survey. 2016. http://www.knbs.or.ke/?p=572.

M-Shule SMS Learning & Training, Kenya — UIL. 2022. https://uil.unesco.org/case-study/effective-practices-database-litbase-0/m-shule-sms-learning-training-kenya.

Muralidharan, Karthik and Abhijeet Singh. 2020. Improving Public Sector Management at Scale? Experimental Evidence on School Governance India. Working Paper 28129 National Bureau of Economic Research.
URL: *http://www.nber.org/papers/w28129*

Muralidharan, Karthik and Paul Niehaus. 2017. “Experimentation at Scale.” *Journal of Economic Perspectives* 31(4):103–24.
URL: *https://www.aeaweb.org/articles?id=10.1257/jep.31.4.103*

Muralidharan, Karthik and Venkatesh Sundararaman. 2011. “Teacher Performance Pay: Experimental Evidence from India.” *Journal of Political Economy* 119(1):39–77.

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

Otis, Nicholas, Rowan Clarke, Solène Delecourt, David Holtz and Rembrand Koning. 2024. “The Uneven Impact of Generative AI on Entrepreneurial Performance.”.

Rambachan, Ashesh and Jonathan Roth. 2022. A More Credible Approach to Parallel Trends. Technical report Working Paper.

Ramirez, Cristina. 2019. CEFE GLOBAL IMPACT STUDY 2019. Technical report CEFE International.

Rasul, Imran and Daniel Rogger. 2018. “Management of Bureaucrats and Public Service Delivery: Evidence from the Nigerian Civil Service.” *The Economic Journal* 128(608):413–446.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/ecoj.12418*

Rasul, Imran, Daniel Rogger and Martin J Williams. 2020. "Management, Organizational Performance, and Task Clarity: Evidence from Ghana's Civil Service." *Journal of Public Administration Research and Theory* 31(2):259–277.
URL: *https://doi.org/10.1093/jopart/muaa034*

Regan-Sachs, Rebecca. 2022. "No Smartphone? No Problem: 3 Keys to Training Unconnected Farmers %." https://www.technoserve.org/blog/3-keys-to-training-unconnected-farmers/.

Riaño, Juan Felipe. 2021. "Bureaucratic nepotism." *Available at SSRN 3995589* .

School Census. 2018. https://schoolportal.punjab.gov.pk/sed_census/.

Schuster, Christian, Kim Sass Mikkelsen, Daniel Rogger, Francis Fukuyama, Zahid Hasnain, Dinsha Mistree, Jan Meyer-Sahling, Katherine Bersch and Kerenssa Kay. 2023. "The Global Survey of Public Servants: Evidence from 1,300,000 Public Servants in 1,300 Government Institutions in 23 Countries." *Public Administration Review* 83(4):982–993.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/puar.13611*

Simon, William H. 1983. "Legality, Bureaucracy, and Class in the Welfare System." *The Yale Law Journal* 92(7):1198–1269.
URL: *http://www.jstor.org/stable/796270*

Spielman, David, Els Lecoutere, Simrin Makhija and Bjorn Van Campenhout. 2021. "Information and Communications Technology (ICT) and Agricultural Extension in Developing Countries." *Annual Review of Resource Economics* 13(Volume 13, 2021):177–201.
URL: *https://www.annualreviews.org/content/journals/10.1146/annurev-resource-101520-080657*

Staiger, Douglas O. and Jonah E. Rockoff. 2010. "Searching for Effective Teachers with Imperfect Information." *Journal of Economic Perspectives* 24(3):97–118.
URL: *https://www.aeaweb.org/articles?id=10.1257/jep.24.3.97*

Sun, Liyang and Sarah Abraham. 2021. "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects." *Journal of Econometrics* 225(2):175–199.

The History of Government Blog. 2022. "The Art of Delivery: The Prime Minister's Delivery Unit, 2001-2005." `https://history.blog.gov.uk/2022/08/26/the-art-of-delivery-the-prime-ministers-delivery-unit-2001-2005/`. Published on August 26, 2022.

Ulmann, Selina. 2023. "Improving Organic Farming Practices in Africa with SMS, IVR, App-Based Training." https://www.rural21.com/english/a-closer-look-at/detail/article/improving-organic-farming-practices-in-africa-with-sms-ivr-app-based-training.html.

United Nations Conference on Trade and Development. 2012. *Information Economy Report 2011: ICTs as an Enabler for Private Sector Development*. United Nations Conference on Trade and Development (UNCTAD) Information Economy Report (IER) UN.

van Vark, Caspar. 2012. "Empowering Farmers through SMS." *The Guardian* .

Vivalt, Eva. 2020. "How Much Can We Generalize From Impact Evaluations?" *Journal of the European Economic Association* 18(6):3045–3089.
URL: *https://doi.org/10.1093/jeea/jvaa019*

Wilson, J.Q. 1989. *Bureaucracy*. Basic Books.
URL: *https://books.google.com/books?id=jg-HAAAAMAAJ*

World Bank. 2020*a*. Technical Review of the PMIU Data Information System. Technical report World Bank Group.

World Bank. 2024. "SME Finance.". Accessed: 2024-08-08.
URL: *https://www.worldbank.org/en/topic/smefinance*

World Bank, Gender Data Portal. 2020*b*. "Entrepreneurship." https://genderdata.worldbank.org/topics/entrepreneurship.

World Bank Group. 2018. World Development Report 2019: LEARNING to Realize Education's Promise. Technical report World Bank Publications.

Xu, Guo. 2018. "The costs of patronage: Evidence from the british empire." *American Economic Review* 108(11):3170–3198.

Appendix A

Short Messages Fall Short for Micro-Entrepreneurs: Experimental Evidence from Kenya – Appendices

A.1 Summary Statistics and Balance Across Treatment and Control

Table A.1.1: Summary Statistics

	Midline			**Endline**		
Variable	Mean	SD	Obs.	Mean	SD	Obs.
Female	0.50	0.50	307	0.47	0.50	2,780
Rural	0.44	0.50	307	0.46	0.50	2,780
Years of education	11.81	2.52	307	11.88	2.70	2,779
Age	35.80	9.73	306	35.30	9.15	2,779
Num of adults in household	2.58	1.50	306	2.63	1.34	2,776
Num of children in household	2.16	1.70	306	2.16	1.49	2,776
Job before intervention	0.25	0.44	307	0.17	0.38	2,780
Business before intervention	0.89	0.31	307	0.85	0.36	2,780
Loan before intervention	0.41	0.49	307	0.38	0.49	2,779

Notes: This table shows the mean, standard deviation, and number of observations for pre-intervention covariates for Midline and Endline.

Table A.1.2: Balance Table

	Midline				**Endline**			
Variable	Full	Treatment	Control	Diff	Full	Treatment	Control	Diff
Female	0.50	0.53	0.44	0.08	0.47	0.46	0.48	-0.02
	(0.50)	(0.50)	(0.50)	(0.06)	(0.50)	(0.50)	(0.50)	(0.02)
Rural	0.44	0.47	0.38	0.09	0.46	0.46	0.46	-0.00
	(0.50)	(0.50)	(0.49)	(0.06)	(0.50)	(0.50)	(0.50)	(0.02)
Years of education	11.81	11.73	11.97	-0.25	11.88	11.84	11.95	-0.11
	(2.52)	(2.48)	(2.60)	(0.30)	(2.70)	(2.73)	(2.65)	(0.10)
Age	35.80	36.52	34.50	2.02*	35.30	35.00	35.75	-0.75**
	(9.73)	(10.19)	(8.73)	(1.16)	(9.16)	(9.03)	(9.33)	(0.35)
Num of adults in household	2.58	2.68	2.39	0.29	2.63	2.62	2.65	-0.03
	(1.50)	(1.66)	(1.11)	(0.18)	(1.34)	(1.34)	(1.34)	(0.05)
Num of children in household	2.16	2.16	2.17	-0.01	2.16	2.14	2.20	-0.07
	(1.70)	(1.75)	(1.61)	(0.20)	(1.49)	(1.50)	(1.49)	(0.06)
Job before intervention	0.25	0.25	0.27	-0.02	0.17	0.18	0.17	0.00
	(0.44)	(0.43)	(0.44)	(0.05)	(0.38)	(0.38)	(0.38)	(0.01)
Business before intervention	0.89	0.89	0.88	0.01	0.85	0.84	0.86	-0.02*
	(0.31)	(0.31)	(0.33)	(0.04)	(0.36)	(0.37)	(0.35)	(0.01)
Loan before intervention	0.41	0.42	0.40	0.02	0.38	0.39	0.37	0.01
	(0.49)	(0.49)	(0.49)	(0.06)	(0.49)	(0.49)	(0.48)	(0.02)
F-test p-value ($\beta_{diff} \neq 0$)				0.24				0.18
Observations	307	198	109	307	2,779	1,668	1,111	2,779

Notes: This table shows the balance of pre-intervention covariates across treatment and control groups for Midline and Endline samples. Standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

A.2 Midline

Table A.2.1: Midline: Engagement by Age

	Full		**Age $\geq$ 34**		**Age < 34**		**Diff**	
	Engaged	% Engaged	Engaged	% Engaged	Engaged	% Engaged	Engaged	% Engaged
Training	0.298***	0.0698***	0.235***	0.0555***	0.375***	0.0896***		
	(0.0306)	(0.00988)	(0.0356)	(0.0116)	(0.0542)	(0.0181)		
Train x Age $\geq$ 34							-0.141**	-0.0341
							(0.0649)	(0.0215)
P-value	0	0	0	0	0	0	0.0310	0.114
Control Mean	0	0	0	0	0	0	0	0
Observations	307	307	160	160	146	146	306	306

Notes: This table shows the effect of treatment assignment on extensive and intensive margin engagement at Midline for the full sample (Columns 1 and 2, respectively), the sample with median and above age (Columns 3 and 4, respectively), the sample with below median age (Columns 5 and 6, respectively), and the difference in treatment effects across median and above, and below median age samples (Columns 7 and 8, respectively). Standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table A.2.2: Midline: Knowledge and Adoption of Advertising

	Knowledge			**Adoption**		
	OLS	IV	IV	OLS	IV	IV
Training	.0913			.0839**		
	(.0633)			(.0344)		
Engaged		.310			.282**	
		(.216)			(.120)	
Covered advertising			.875			.812**
			(.614)			(.360)
Female	.0579	.0401	.0193	-.123***	-.140***	-.155***
	(.0696)	(.0726)	(.0786)	(.0430)	(.0493)	(.0562)
P-value	.150	.151	.154	.0154	.0183	.0241
Control Mean	.385	.385	.385	.0380	.0380	.0380
Observations	307	307	307	297	297	297

Notes: This table shows the intent-to-treat and local average treatment effect estimates of SMS trainings on knowledge and adoption of advertising at Midline. The dependent variable in the first three columns is a binary variable that indicates whether the individual responded correctly to the question testing knolwedge of advertising, while in the last three columns it is a binary variable that indicates whether the individual advertised any of their products in the last three months. Columns (1) and (4) show output from OLS regressions, Columns (2) and (5) show output from a 2SLS regressions where the endogenous variable is whether or not the individual engaged with training content, and Columns (3) and (6) show output from 2SLS regressions where the endogenous variable is whether or not the individual engaged with the part of the training content that covered advertising. Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.2.3: Midline: Balance Across Treatment and Control by Age

	Age $\geq$ 34				**Age $<$ 34**			
Variable	Full	Treatment	Control	Diff	Full	Treatment	Control	Diff
Female	0.59	0.64	0.50	0.14*	0.39	0.39	0.39	0.01
	(0.49)	(0.48)	(0.50)	(0.08)	(0.49)	(0.49)	(0.49)	(0.08)
Rural	0.46	0.50	0.38	0.12	0.41	0.44	0.37	0.07
	(0.50)	(0.50)	(0.49)	(0.08)	(0.49)	(0.50)	(0.49)	(0.08)
Years of education	10.85	10.72	11.12	-0.39	12.86	12.92	12.75	0.17
	(2.62)	(2.57)	(2.73)	(0.44)	(1.93)	(1.73)	(2.23)	(0.33)
Num of adults in household	2.78	2.92	2.50	0.42	2.36	2.39	2.30	0.10
	(1.59)	(1.77)	(1.08)	(0.27)	(1.36)	(1.48)	(1.15)	(0.23)
Num of children in household	2.50	2.45	2.60	-0.14	1.79	1.80	1.77	0.03
	(1.67)	(1.65)	(1.71)	(0.28)	(1.66)	(1.81)	(1.41)	(0.28)
Job before intervention	0.19	0.19	0.21	-0.03	0.32	0.31	0.32	-0.00
	(0.40)	(0.39)	(0.41)	(0.07)	(0.47)	(0.47)	(0.47)	(0.08)
Business before intervention	0.92	0.91	0.94	-0.03	0.86	0.88	0.82	0.05
	(0.27)	(0.29)	(0.24)	(0.05)	(0.35)	(0.33)	(0.38)	(0.06)
Loan before intervention	0.46	0.49	0.40	0.09	0.36	0.33	0.40	-0.08
	(0.50)	(0.50)	(0.50)	(0.08)	(0.48)	(0.47)	(0.49)	(0.08)
Observations	160	108	52	160	146	89	57	146

Notes: This table shows the balance of pre-intervention covariates across treatment and control groups at Midline for the sample with median and above age, and the sample with below median age. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.2.4: Midline: Primary Business Performance by Age

	Full			**Age $\geq$ 34**			**Age < 34**			**Diff**		
	Sales	Profits	Survival	Sales	Profits	Survival	Sales	Profits	Survival	Sales	Profits	Survival
Training	5721.0	1680.6	0.0414	-23403.5	-450.1	-0.0289	35607.9**	3993.1	0.116**			
	(10814.3)	(1601.9)	(0.0326)	(16185.3)	(2031.0)	(0.0388)	(14825.0)	(2517.4)	(0.0482)			
Train x Age $\geq$ 34										-59011.5***	-4443.2	-0.145**
										(21949.1)	(3234.3)	(0.0619)
P-value	0.597	0.295	0.204	0.150	0.825	0.458	0.0180	0.115	0.0180	0.00800	0.171	0.0200
Control Mean	47581.2	10886.9	0.908	64450	11349.4	0.962	32586.7	10482.1	0.860	47581.2	10886.9	0.908
Observations	290	294	307	152	151	160	138	143	146	290	294	306

Notes: This table shows the effect of treatment assignment on primary business sales in the last 30 days, primary business profits in the last 30 days, and business survival at Midline for the full sample (Columns 1, 2 and 3, respectively), the sample with median and above age (Columns 4, 5, and 6, respectively), the sample with below median age (Columns 7, 8, and 9, respectively), and the difference in treatment effects across median and above, and below median age samples (Columns 10, 11, and 12, respectively). Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.2.5: Midline: Hours Spent Across All Businesses

	Hrs. worked	
	OLS	IV
Training	31.46*	
	(17.74)	
Engaged		116.9*
		(66.91)
Female	-10.73	-17.15
	(17.72)	(18.82)
P-value	.0774	.0807
Control Mean	198.5	198.5
Observations	267	267

Notes: This table shows the intent-to-treat and local average treatment effect estimates of SMS trainings on time spent across all businesses in the last 30 days at Midline. Coefficients represent effects in terms of hours worked. Column (1) shows output from an OLS regression, and Column (2) shows output from a 2SLS regression where the endogenous variable is whether or not the individual engaged with training content. Robust standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table A.2.6: Midline: Side Jobs

	Job		**Job Hours**	
	OLS	IV	OLS	IV
Training	-.0295		-5.951	
	(.0533)		(10.27)	
Engaged		-.102		-20.72
		(.184)		(35.75)
Female	-.0786	-.0722	-22.81**	-21.45*
	(.0557)	(.0582)	(10.15)	(11.52)
P-value	.580	.580	.563	.562
Control Mean	.236	.236	35.06	35.06
Observations	296	296	291	291

Notes: This table shows the intent-to-treat and local average treatment effect estimates of SMS trainings on employment in and time spent on side jobs in the last 30 days at Midline. Coefficients in Columns (1) and (2) represent effects in terms of probability of having a side job, while those in Columns (3) and (4) represent effects in terms of hours worked. Columns (1) and (3) show output from OLS regressions, while Columns (2) and (4) show output from 2SLS regressions where the endogenous variable is whether or not the individual engaged with training content. Robust standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table A.2.7: Midline: Time Spent on Business by Age

	Full		**Age $\geq$ 34**		**Age < 34**		**Diff**	
	Primary	All	Primary	All	Primary	All	Primary	All
Training	28.88*	31.46*	-14.94	-14.55	67.23***	70.90***		
	(16.52)	(17.74)	(23.74)	(24.50)	(22.06)	(24.69)		
Train x Age $\geq$ 34							-82.17**	-85.46**
							(32.41)	(34.78)
P-value	0.0817	0.0774	0.530	0.553	0.00300	0.00500	0.0120	0.0150
Control Mean	178.6	198.5	214.8	238.6	148.0	164	178.6	198.5
Observations	269	267	139	139	129	127	268	266

Notes: This table shows the effect of treatment assignment on hours spent working on primary business and across all businesses in the last 30 days at Midline for the full sample (Columns (1), and (2), respectively), the sample with median and above age (Columns (3), and (4), respectively), the sample with below median age (Columns (5), and (6), respectively), and the difference in treatment effects across median and above, and below median age samples (Columns (7), and (8), respectively). Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.2.8: Midline: Labor Hours Employed in Last 30 days

	HH Labor Hrs. - Primary		**Outside Labor Hrs. - Primary**		**HH Labor Hrs. - All**		**Outside Labor Hrs. - All**	
	OLS	IV	OLS	IV	OLS	IV	OLS	IV
Training	2.496		-3.282		-8.693		-7.263	
	(12.03)		(27.48)		(14.53)		(33.37)	
Engaged		8.430		-11.33		-29.50		-25.06
		(40.42)		(94.36)		(49.21)		(114.6)
Female	-37.96***	-38.40***	-40.06*	-39.36*	-50.76***	-49.06***	-50.49	-48.95
	(11.80)	(12.54)	(23.91)	(23.09)	(14.11)	(15.09)	(31.73)	(30.13)
P-value	.836	.835	.905	.904	.550	.549	.828	.827
Control Mean	42.50	42.50	83.03	83.03	64.19	64.19	107.3	107.3
Observations	302	302	297	297	307	307	297	297

Notes: This table shows the intent-to-treat and local average treatment effect estimates of SMS trainings on household and outside labor employed in primary business and across all businesses in the last 30 days at Midline. Coefficients across all Columns represent effects in terms of labor hours employed. Columns (1), (3), (5), and (7) show output from OLS regressions, while Columns (2), (4), (6), and (8) show output from 2SLS regressions where the endogenous variable is whether or not the individual engaged with training content. Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.2.9: Midline: Loan Amount Applied for and Received by Age

	Full		**Age ≥ 34**		**Age < 34**		**Diff**	
	Applied	Received	Applied	Received	Applied	Received	Applied	Received
Training	-365.6	987.1	-15646.6	-9935.4	15236.9*	12081.4*		
	(7629.7)	(5570.2)	(13477.2)	(9249.4)	(8471.6)	(7082.0)		
Train x Age ≥ 34							-30883.5*	-22016.8*
							(15921.5)	(11650.4)
P-value	0.962	0.859	0.247	0.284	0.0740	0.0900	0.0530	0.0600
Control Mean	13818.3	10104.6	23442.3	16980.8	5038.6	3831.6	13818.3	10104.6
Observations	307	307	160	160	146	146	306	306

Notes: This table shows the effect of treatment assignment on loan amount applied for and received in Kenyan Shillings in the last 3 months at Midline for the full sample (Columns (1), and (2), respectively), the sample with median and above age (Columns (3), and (4), respectively), the sample with below median age (Columns (5), and (6), respectively), and the difference in treatment effects across median and above, and below median age samples (Columns (7), and (8), respectively). Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

A.3 Endline Vs Midline Samples

Table A.3.1: Endline VS Midline Samples: covariates

Variable	Matched with Midline	Unmatched with Midline	Diff
Female	0.48	0.47	0.02
	(0.50)	(0.50)	(0.03)
Years of education	11.91	11.88	0.03
	(2.40)	(2.73)	(0.19)
Age	36.26	35.22	1.04
	(9.00)	(9.17)	(0.63)
Rural	0.46	0.46	-0.00
	(0.50)	(0.50)	(0.03)
Num of adults in household	2.55	2.63	-0.08
	(1.25)	(1.35)	(0.09)
Num of children in household	2.11	2.17	-0.06
	(1.42)	(1.50)	(0.10)
Job before intervention	0.15	0.18	-0.03
	(0.36)	(0.38)	(0.03)
Business before intervention	0.85	0.85	-0.00
	(0.36)	(0.36)	(0.02)
Loan before intervention	0.39	0.38	0.01
	(0.49)	(0.49)	(0.03)
F-test p-value ($\beta_{diff} \neq 0$)			0.96
Observations	227	2,553	2,780

Notes: This table shows comparison of pre-intervention covariates across the Endline sample matched with the Midline, and the Endline sample not matched with the Midline. Standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table A.3.2: Endline VS Midline Samples: Control Outcomes

Variable	Matched with Midline	Unmatched with Midline	Diff
Basic Knowledge	0.74	0.73	0.01
	(0.16)	(0.18)	(0.02)
Advanced Knowledge	0.77	0.79	-0.02
	(0.18)	(0.18)	(0.02)
Overall Knowledge	0.75	0.76	-0.01
	(0.14)	(0.14)	(0.02)
Basic Adoption	0.64	0.70	-0.06**
	(0.22)	(0.21)	(0.02)
Advanced Adoption	0.67	0.67	-0.00
	(0.26)	(0.22)	(0.03)
Overall Adoption	0.66	0.69	-0.03
	(0.17)	(0.18)	(0.02)
Owns Business	0.92	0.92	0.00
	(0.28)	(0.28)	(0.03)
Num of Businesses Owned	0.98	1.01	-0.03
	(0.38)	(0.44)	(0.05)
Business Registered	0.51	0.46	0.05
	(0.50)	(0.50)	(0.06)
Num of Businesses Registered	0.55	0.50	0.05
	(0.57)	(0.55)	(0.07)
7-day Sales from Primary Business	15454.17	15780.95	-326.79
	(21356.77)	(23449.46)	(2644.42)
30-day Sales from Primary Business	62773.81	59075.36	3698.45
	(93706.85)	(90622.07)	(10312.33)
7-day Sales from All Businesses	15993.45	16975.15	-981.70
	(21397.30)	(25210.13)	(2831.15)
30-day Sales from All Businesses	64934.52	64413.61	520.91
	(93808.84)	(102112.05)	(11521.62)
7-day Profits from Primary Business	4725.00	4793.07	-68.07
	(5138.23)	(6190.39)	(694.35)
30-day Profits from Primary Business	18878.57	19500.64	-622.07
	(19334.51)	(25710.48)	(2870.13)
7-day Profits from All Businesses	4908.93	5221.90	-312.97
	(5180.22)	(6997.48)	(780.62)
30-day Profits from All Businesses	19560.71	21215.50	-1654.78
	(19391.96)	(29073.36)	(3230.30)
Applied for a Loan	0.52	0.47	0.06
	(0.50)	(0.50)	(0.06)
Loan Amount Applied	20970.24	20345.08	625.16
	(39903.22)	(61354.25)	(6810.69)
Loan Amount Received	17396.90	16765.28	631.63
	(32061.82)	(55158.15)	(6102.47)
Loan Application Success Rate	0.91	0.88	0.03
	(0.29)	(0.33)	(0.05)
Loan Payment Missed/Late	0.53	0.56	-0.03
	(0.51)	(0.50)	(0.12)
F-test p-value ($\beta_{diff} \neq 0$)			0.998
Observations	84	1,027	1,111

Notes: This table shows comparison of control group outcomes across the Endline sample matched with the Midline, and the Endline sample not matched with the Midline. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

A.4 Endline

Table A.4.1: Endline: Engagement by Age

	Full		**Age $\geq$ 34**		**Age < 34**		**Diff**	
	Engaged	% Engaged	Engaged	% Engaged	Engaged	% Engaged	Engaged	% Engaged
Training	0.280^{***}	0.0651^{***}	0.281^{***}	0.0605^{***}	0.279^{***}	0.0700^{***}		
	(0.0110)	(0.00411)	(0.0157)	(0.00549)	(0.0155)	(0.00616)		
Train x Age $\geq$ 34							0.00194	-0.00945
							(0.0220)	(0.00825)
P-value	0	0	0	0	0	0	0.930	0.252
Control Mean	0	0	0	0	0	0	0	0
Observations	2780	2780	1426	1426	1353	1353	2779	2779

Notes: This table shows the effect of treatment assignment on extensive and intensive margin engagement at Endline for the full sample (Columns 1 and 2, respectively), the sample with median and above age (Columns 3 and 4, respectively), the sample with below median age (Columns 5 and 6, respectively), and the difference in treatment effects across median and above, and below median age samples (Columns 7 and 8, respectively). Standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table A.4.2: Endline: Knowledge and Adoption of Best Practices Using Midline Sample

	Basic Knowledge		**Basic Adoption**		**Advanced Knowledge**		**Advanced Adoption**	
	OLS	IV	OLS	IV	OLS	IV	OLS	IV
Training	-.0393		.159		-.132		-.0128	
	(.143)		(.139)		(.171)		(.137)	
Engaged		-.103		.404		-.347		-.0333
		(.376)		(.356)		(.453)		(.353)
Female	.0523	.0607	-.114	-.150	-.153	-.125	.161	.164
	(.154)	(.155)	(.139)	(.147)	(.213)	(.215)	(.129)	(.141)
P-value	.785	.783	.256	.257	.442	.444	.926	.925
Control Mean	0	0	0	0	0	0	0	0
Observations	227	227	217	217	227	227	216	216

Notes: This table shows the intent-to-treat and local average treatment effect estimates of SMS trainings on means effect indices of basic and advanced knowledge and adoption of best practices at Endline, using the sample matched with Midline only. Basic knowledge and basic adoption indices are similar to the knowledge and adoption indices analysed for the Midline, while the advanced knowledge and adoption indices are based on best practices are a bit more advanced and not necessarily directly mentioned in the SMS-trainings. Coefficients represent effects in terms of control group standard deviations. Columns (1), (3), (5), and (7) show output from OLS regressions, and columns (2), (4), (6), and (8) show output from 2SLS regressions where the endogenous variable is whether or not the individual engaged with training content. Robust standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table A.4.3: Endline: Primary Business Sales, Profits and Survival Using Midline Sample

	Sales		**Profits**		**Survival**	
	OLS	IV	OLS	IV	OLS	IV
Training	-17305.1		-182.8		.00879	
	(11060.5)		(3005.2)		(.0407)	
Engaged		-45882.8		-484.7		.0232
		(29928.4)		(7915.2)		(.107)
Female	-36873.9***	-32984.0***	-10376.6***	-10335.5***	-.0380	-.0399
	(8353.9)	(9411.3)	(3298.1)	(3571.9)	(.0455)	(.0445)
P-value	.119	.125	.952	.951	.829	.828
Control Mean	62773.8	62773.8	18878.6	18878.6	.917	.917
Observations	226	226	226	226	227	227

Notes: This table shows the intent-to-treat and local average treatment effect estimates of SMS trainings on primary business sales and profits from last 30 days, and business survival at Endline, using the sample matched with Midline only. Coefficients in columns (1) through (4) represent effects in terms of Kenyan Shillings, while those in columns (5) and (6) represent probability of individual having an active business. Columns (1), (3) and (5) show output from OLS regressions, and columns (2), (4) and (6) show output from 2SLS regressions where the endogenous variable is whether or not the individual engaged with training content. Robust standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table A.4.4: Endline: Sales and Profits Across All Businesses

	Sales		**Profits**	
	OLS	IV	OLS	IV
Training	668.7		482.7	
	(4075.1)		(1153.7)	
Engaged		2391.6		1727.9
		(14564.1)		(4127.7)
Female	-42183.0***	-42219.9***	-11698.1***	-11723.4***
	(3951.9)	(3966.5)	(1128.2)	(1130.6)
P-value	.870	.870	.676	.676
Control Mean	64453.1	64453.1	21089.9	21089.9
Observations	2772	2772	2770	2770

Notes: This table shows the intent-to-treat and local average treatment effect estimates of SMS trainings on sales and profits across all businesses from last 30 days at Endline. Coefficients across all columns represent effects in terms of Kenyan Shillings. Columns (1), and (3) show output from OLS regressions, and Columns (2), and (4) show output from 2SLS regressions where the endogenous variable is whether or not the individual engaged with training content. Robust standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table A.4.5: Endline: Sales and Profits Across All Businesses Using Midline Sample

	Sales		**Profits**	
	OLS	IV	OLS	IV
Training	-12427.6		986.0	
	(11442.1)		(3111.7)	
Engaged		-32950.4		2614.3
		(30620.7)		(8201.6)
Female	-37827.5***	-35034.0***	-11608.2***	-11829.9***
	(9534.4)	(10050.8)	(3475.7)	(3764.6)
P-value	.279	.282	.752	.750
Control Mean	64934.5	64934.5	19560.7	19560.7
Observations	226	226	226	226

Notes: This table shows the intent-to-treat and local average treatment effect estimates of SMS trainings on sales and profits across all businesses from last 30 days at Endline, using the sample matched with Midline only. Coefficients across all columns represent effects in terms of Kenyan Shillings. Columns (1), and (3) show output from OLS regressions, and Columns (2), and (4) show output from 2SLS regressions where the endogenous variable is whether or not the individual engaged with training content. Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.4.6: Endline: Primary Business Sales, Profits and Survival By Age

	Full			**Age ≥ 34**			**Age < 34**			**Diff**		
	Sales	Profits	Survival	Sales	Profits	Survival	Sales	Profits	Survival	Sales	Profits	Survival
Training	-2206.5	-220.6	-0.0159	-2217.1	2.944	-0.0178	-688.6	-259.5	-0.00680			
	(3534.1)	(1009.0)	(0.0112)	(5244.1)	(1429.6)	(0.0131)	(4680.3)	(1426.9)	(0.0184)			
Train x Age $\geq$ 34										-1528.5	262.4	-0.0110
										(7029.0)	(2019.8)	(0.0226)
P-value	0.532	0.827	0.154	0.673	0.998	0.174	0.883	0.856	0.712	0.828	0.897	0.627
Control Mean	59356.0	19453.4	0.915	62630.4	19478.0	0.944	55466.9	19424.2	0.882	59356.0	19453.4	0.915
Observations	2772	2770	2779	1419	1417	1425	1352	1352	1353	2771	2769	2778

Notes: This table shows the effect of treatment assignment on primary business sales in the last 30 days, primary business profits in the last 30 days, and business survival at Endline for the full sample (Columns (1), (2) and (3), respectively), the sample with median and above age (Columns (4), (5), and (6), respectively), the sample with below median age (Columns (7), (8), and (9), respectively), and the difference in treatment effects across median and above, and below median age samples (Columns (10), (11), and (12), respectively). Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.4.7: Endline: Primary Business Sales, Profits and Survival By Age Using Midline Sample

	Full			**Age ≥ 34**			**Age < 34**			**Diff**		
	Sales	Profits	Survival	Sales	Profits	Survival	Sales	Profits	Survival	Sales	Profits	Survival
Training	-17305.1	-182.8	0.00879	-30631.6**	-1845.4	-0.00491	604.9	1903.7	0.0318			
	(11060.5)	(3005.2)	(0.0407)	(13859.9)	(3784.4)	(0.0498)	(17903.8)	(4890.2)	(0.0707)			
Train x Age $\geq$ 34										-31236.5	-3749.1	-0.0368
										(22628.4)	(6179.8)	(0.0865)
P-value	0.119	0.952	0.829	0.0290	0.627	0.922	0.973	0.698	0.654	0.169	0.545	0.671
Control Mean	62773.8	18878.6	0.917	67967.3	18761.2	0.939	55502.9	19042.9	0.886	62773.8	18878.6	0.917
Observations	226	226	227	126	126	127	100	100	100	226	226	227

Notes: This table shows the effect of treatment assignment on primary business sales in the last 30 days, primary business profits in the last 30 days, and business survival at Endline, using the sample matched with Midline only, for the full sample (Columns (1), (2) and (3), respectively), the sample with median and above age (Columns (4), (5), and (6), respectively), the sample with below median age (Columns (7), (8), and (9), respectively), and the difference in treatment effects across median and above, and below median age samples (Columns (10), (11), and (12), respectively). Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.4.8: Endline: Sales and Profits Across All Businesses By Age

	Full		**Age $\geq$ 34**		**Age < 34**		**Diff**	
	Sales	Profits	Sales	Profits	Sales	Profits	Sales	Profits
Training	668.7	482.7	3484.0	1572.3	-484.9	-408.1		
	(4075.1)	(1153.7)	(6118.3)	(1647.0)	(5362.3)	(1633.0)		
Train x Age $\geq$ 34							3968.8	1980.4
							(8135.6)	(2319.3)
P-value	0.870	0.676	0.569	0.340	0.928	0.803	0.626	0.393
Control Mean	64453.1	21089.9	67194.8	20851.7	61196.7	21372.8	64453.1	21089.9
Observations	2772	2770	1419	1417	1352	1352	2771	2769

Notes: This table shows the effect of treatment assignment on sales and profits across all businesses in the last 30 days at Endline, for the full sample (Columns (1), and (2), respectively), the sample with median and above age (Columns (3), and (4), respectively), the sample with below median age (Columns (5), and (6), respectively), and the difference in treatment effects across median and above, and below median age samples (Columns (7), and (8), respectively). Robust standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table A.4.9: Endline: Sales and Profits Across All Businesses By Age Using Midline Sample

	Full		**Age $\geq$ 34**		**Age < 34**		**Diff**	
	Sales	Profits	Sales	Profits	Sales	Profits	Sales	Profits
Training	-12427.6	986.0	-26692.7*	-410.7	8250.5	3014.2		
	(11442.1)	(3111.7)	(14273.5)	(4055.5)	(18940.3)	(4952.3)		
Train x Age $\geq$ 34							-34943.2	-3424.9
							(23701.7)	(6397.8)
P-value	0.279	0.752	0.0640	0.919	0.664	0.544	0.142	0.593
Control Mean	64934.5	19560.7	71191.8	19634.7	56174.3	19457.1	64934.5	19560.7
Observations	226	226	126	126	100	100	226	226

Notes: This table shows the effect of treatment assignment on sales and profits across all businesses in the last 30 days at Endline, using the sample matched with Midline only, for the full sample (Columns (1), and (2), respectively), the sample with median and above age (Columns (3), and (4), respectively), the sample with below median age (Columns (5), and (6), respectively), and the difference in treatment effects across median and above, and below median age samples (Columns (7), and (8), respectively). Robust standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table A.4.10: Endline: Time Spent on Primary Business in last 30 days Using Midline Sample

	Hrs. Worked	
	OLS	IV
Training	9.672	
	(17.29)	
Engaged		25.65
		(45.96)
Female	-4.735	-6.909
	(18.98)	(20.02)
P-value	.576	.577
Control Mean	208.7	208.7
Observations	226	226

Notes: This table shows the intent-to-treat and local average treatment effect estimates of SMS trainings on time spent on primary business in the last 30 days at Endline, using the sample matched with Midline only. Coefficients represent effects in terms of hours worked. Column (1) shows output from an OLS regression, and Column (2) shows output from a 2SLS regression where the endogenous variable is whether or not the individual engaged with training content. Robust standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table A.4.11: Endline: Time Spent on All Businesses

	Hrs. in 7 days		**Hrs. in 30 days**	
	OLS	IV	OLS	IV
Training	-2.366**		-9.152**	
	(1.149)		(4.577)	
Engaged		-8.442**		-32.65**
		(4.113)		(16.38)
Female	-1.903*	-1.775	-8.249*	-7.752*
	(1.121)	(1.126)	(4.472)	(4.491)
P-value	.0396	.0401	.0457	.0462
Control Mean	55.66	55.66	223.4	223.4
Observations	2779	2779	2779	2779

Notes: This table shows the intent-to-treat and local average treatment effect estimates of SMS trainings on time spent across all businesses in the last 7 and 30 days at Endline. Coefficients represent effects in terms of hours worked. Columns (1) and (3) show output from an OLS regressions, and Columns (2) and (4) show output from a 2SLS regressions where the endogenous variable is whether or not the individual engaged with training content. Robust standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table A.4.12: Endline: Time Spent Across All Businesses Using Midline Sample

	Hrs. in 7 days		**Hrs. in 30 days**	
	OLS	IV	OLS	IV
Training	3.909		13.57	
	(4.487)		(17.75)	
Engaged		10.30		35.77
		(11.94)		(47.14)
Female	-3.413	-4.253	-13.21	-16.12
	(4.765)	(5.082)	(19.03)	(20.18)
P-value	.385	.388	.445	.448
Control Mean	53.26	53.26	215.0	215.0
Observations	227	227	227	227

Notes: This table shows the intent-to-treat and local average treatment effect estimates of SMS trainings on time spent across all businesses in the last 7 and 30 days at Endline, using the sample matched with Midline only. Coefficients represent effects in terms of hours worked. Columns (1) and (3) show output from an OLS regressions, and Columns (2) and (4) show output from a 2SLS regressions where the endogenous variable is whether or not the individual engaged with training content. Robust standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table A.4.13: Endline: Side Jobs in Last 30 days

	Job		**Job Hours**	
	OLS	IV	OLS	IV
Training	-.0141		-1.856	
	(.0143)		(2.411)	
Engaged		-.0503		-6.625
		(.0511)		(8.606)
Female	-.0910***	-.0902***	-17.72***	-17.62***
	(.0138)	(.0138)	(2.292)	(2.283)
P-value	.325	.325	.442	.441
Control Mean	.171	.171	23.36	23.36
Observations	2780	2780	2779	2779

Notes: This table shows the intent-to-treat and local average treatment effect estimates of SMS trainings on employment in and time spent on side jobs in the last 30 days at Endline. Coefficients in Columns (1) and (2) represent effects in terms of probability of having a side job, while those in Columns (3) and (4) represent effects in terms of hours worked. Columns (1) and (3) show output from OLS regressions, while Columns (2) and (4) show output from 2SLS regressions where the endogenous variable is whether or not the individual engaged with training content. Robust standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table A.4.14: Endline: Side Jobs in Last 30 days Using Midline Sample

	Job		**Job Hours**	
	OLS	IV	OLS	IV
Training	.0273		-1.446	
	(.0555)		(8.879)	
Engaged		.0720		-3.810
		(.145)		(23.28)
Female	-.0870	-.0928	-21.58**	-21.27**
	(.0631)	(.0641)	(8.634)	(8.986)
P-value	.623	.619	.871	.870
Control Mean	.155	.155	24.36	24.36
Observations	227	227	227	227

Notes: This table shows the intent-to-treat and local average treatment effect estimates of SMS trainings on employment in and time spent on side jobs in the last 30 days at Endline, using the sample matched with Midline only. Coefficients in Columns (1) and (2) represent effects in terms of probability of having a side job, while those in Columns (3) and (4) represent effects in terms of hours worked. Columns (1) and (3) show output from OLS regressions, while Columns (2) and (4) show output from 2SLS regressions where the endogenous variable is whether or not the individual engaged with training content. Robust standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table A.4.15: Endline: Time Spent on Business in Last 30 days by Age

	Full		**Age ≥ 34**		**Age < 34**		**Diff**	
	Primary	All	Primary	All	Primary	All	Primary	All
Training	-9.702**	-9.152**	-11.54*	-10.60*	-5.769	-5.859		
	(4.459)	(4.577)	(5.986)	(6.133)	(6.635)	(6.838)		
Train x Age ≥ 34							-5.769	-4.737
							(8.936)	(9.185)
P-value	0.0296	0.0457	0.0540	0.0840	0.385	0.392	0.519	0.606
Control Mean	215.6	223.4	223.7	230.3	205.9	215.2	215.6	223.4
Observations	2777	2779	1423	1425	1353	1353	2776	2778

Notes: This table shows the effect of treatment assignment on hours spent working on primary business and across all businesses in the last 30 days at Endline for the full sample (Columns (1), and (2), respectively), the sample with median and above age (Columns (3), and (4), respectively), the sample with below median age (Columns (5), and (6), respectively), and the difference in treatment effects across median and above, and below median age samples (Columns (7), and (8), respectively). Robust standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table A.4.16: Endline: Time Spent on Business in Last 30 days by Age Using Midline Sample

	Full		**Age ≥ 34**		**Age < 34**		**Diff**	
	Primary	All	Primary	All	Primary	All	Primary	All
Training	9.672	13.57	-11.21	-4.813	40.41	41.62		
	(17.29)	(17.75)	(23.31)	(24.33)	(25.98)	(26.09)		
Train x Age ≥ 34							-51.62	-46.43
							(34.89)	(35.66)
P-value	0.576	0.445	0.631	0.843	0.123	0.114	0.140	0.194
Control Mean	208.7	215.0	228.5	234.9	181.0	187.1	208.7	215.0
Observations	226	227	126	127	100	100	226	227

Notes: This table shows the effect of treatment assignment on hours spent working on primary business and across all businesses in the last 30 days at Endline, using the sample matched with Midline only, for the full sample (Columns (1), and (2), respectively), the sample with median and above age (Columns (3), and (4), respectively), the sample with below median age (Columns (5), and (6), respectively), and the difference in treatment effects across median and above, and below median age samples (Columns (7), and (8), respectively). Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.4.17: Endline: Time Spent on Business in Last 7 days by Age

	Full		**Age ≥ 34**		**Age < 34**		**Diff**	
	Primary	All	Primary	All	Primary	All	Primary	All
Training	-2.534**	-2.366**	-3.414**	-3.125**	-1.065	-1.079		
	(1.113)	(1.149)	(1.482)	(1.529)	(1.667)	(1.729)		
Train x Age ≥ 34							-2.349	-2.046
							(2.231)	(2.308)
P-value	0.0229	0.0396	0.0210	0.0410	0.523	0.533	0.292	0.375
Control Mean	53.68	55.66	55.96	57.64	50.97	53.31	53.68	55.66
Observations	2777	2779	1423	1425	1353	1353	2776	2778

Notes: This table shows the effect of treatment assignment on hours spent working on primary business and across all businesses in the last 7 days at Endline for the full sample (Columns (1), and (2), respectively), the sample with median and above age (Columns (3), and (4), respectively), the sample with below median age (Columns (5), and (6), respectively), and the difference in treatment effects across median and above, and below median age samples (Columns (7), and (8), respectively). Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.4.18: Endline: Time Spent on Business in Last 7 days by Age Using Midline Sample

	Full		**Age $\geq$ 34**		**Age < 34**		**Diff**	
	Primary	All	Primary	All	Primary	All	Primary	All
Training	2.781	3.909	-2.080	-0.221	9.963	10.32		
	(4.319)	(4.487)	(5.821)	(6.192)	(6.475)	(6.502)		
Train x Age $\geq$ 34							-12.04	-10.54
							(8.704)	(8.976)
P-value	0.520	0.385	0.721	0.972	0.127	0.116	0.168	0.242
Control Mean	51.60	53.26	56.47	58.22	44.77	46.31	51.59	53.26
Observations	226	227	126	127	100	100	226	227

Notes: This table shows the effect of treatment assignment on hours spent working on primary business and across all businesses in the last 7 days at Endline, using the sample matched with Midline only, for the full sample (Columns (1), and (2), respectively), the sample with median and above age (Columns (3), and (4), respectively), the sample with below median age (Columns (5), and (6), respectively), and the difference in treatment effects across median and above, and below median age samples (Columns (7), and (8), respectively). Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.4.19: Endline: Labor Hours Employed in Last 30 Days

	HH Labor Hrs. - Primary		**Outside Labor Hrs. - Primary**		**HH Labor Hrs. - All**		**Outside Labor Hrs. - All**	
	OLS	IV	OLS	IV	OLS	IV	OLS	IV
Training	-4.496		1.021		-5.465		2.763	
	(3.145)		(5.555)		(3.486)		(6.584)	
Engaged		-16.04		3.643		-19.50		9.859
		(11.24)		(19.80)		(12.47)		(23.47)
Female	-25.58***	-25.33***	-40.51***	-40.57***	-29.51***	-29.21***	-58.40***	-58.55***
	(2.963)	(2.994)	(5.404)	(5.432)	(3.282)	(3.318)	(6.417)	(6.446)
P-value	.153	.154	.854	.854	.117	.118	.675	.674
Control Mean	34.43	34.43	55.39	55.39	39.74	39.74	68.64	68.64
Observations	2779	2779	2778	2778	2779	2779	2779	2779

Notes: This table shows the intent-to-treat and local average treatment effect estimates of SMS trainings on household and outside labor employed in primary business and across all businesses in the last 30 days at Endline. Coefficients across all Columns represent effects in terms of labor hours employed. Columns (1), (3), (5), and (7) show output from OLS regressions, while Columns (2), (4), (6), and (8) show output from 2SLS regressions where the endogenous variable is whether or not the individual engaged with training content. Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.4.20: Endline: Labor Hours Employed in Last 30 Days Using Midline Sample

	HH Labor Hrs. - Primary		**Outside Labor Hrs. - Primary**		**HH Labor Hrs. - All**		**Outside Labor Hrs. - All**	
	OLS	IV	OLS	IV	OLS	IV	OLS	IV
Training	-4.280		-12.80		-4.110		-4.364	
	(13.09)		(19.93)		(14.75)		(23.45)	
Engaged		-11.28		-33.73		-10.83		-11.50
		(34.27)		(53.12)		(38.60)		(61.66)
Female	-25.18*	-24.26	-56.25***	-53.50**	-34.10**	-33.22*	-63.74***	-62.80**
	(14.48)	(14.82)	(19.73)	(22.33)	(16.34)	(17.13)	(24.45)	(26.97)
P-value	.744	.742	.521	.525	.781	.779	.853	.852
Control Mean	43.06	43.06	55.05	55.05	49.39	49.39	61.71	61.71
Observations	227	227	227	227	227	227	227	227

Notes: This table shows the intent-to-treat and local average treatment effect estimates of SMS trainings on household and outside labor employed in primary business and across all businesses in the last 30 days at Endline, using the sample matched with Midline only. Coefficients across all Columns represent effects in terms of labor hours employed. Columns (1), (3), (5), and (7) show output from OLS regressions, while Columns (2), (4), (6), and (8) show output from 2SLS regressions where the endogenous variable is whether or not the individual engaged with training content. Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.4.21: Endline: Labor Hours Employed in Last 7 Days

	HH Labor Hrs. - Primary		**Outside Labor Hrs. - Primary**		**HH Labor Hrs. - All**		**Outside Labor Hrs. - All**	
	OLS	IV	OLS	IV	OLS	IV	OLS	IV
Training	-.965		.221		-1.206		.617	
	(.796)		(1.352)		(.870)		(1.601)	
Engaged		-3.442		.790		-4.304		2.200
		(2.845)		(4.819)		(3.110)		(5.708)
Female	-6.367***	-6.314***	-9.912***	-9.924***	-7.247***	-7.181***	-14.28***	-14.32***
	(.750)	(.758)	(1.314)	(1.322)	(.819)	(.828)	(1.562)	(1.571)
P-value	.226	.226	.870	.870	.166	.166	.700	.700
Control Mean	8.584	8.584	13.62	13.62	9.848	9.848	16.92	16.92
Observations	2779	2779	2778	2778	2779	2779	2779	2779

Notes: This table shows the intent-to-treat and local average treatment effect estimates of SMS trainings on household and outside labor employed in primary business and across all businesses in the last 7 days at Endline. Coefficients across all Columns represent effects in terms of labor hours employed. Columns (1), (3), (5), and (7) show output from OLS regressions, while Columns (2), (4), (6), and (8) show output from 2SLS regressions where the endogenous variable is whether or not the individual engaged with training content. Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.4.22: Endline: Labor Hours Employed in Last 7 Days

	HH Labor Hrs. - Primary		**Outside Labor Hrs. - Primary**		**HH Labor Hrs. - All**		**Outside Labor Hrs. - All**	
	OLS	IV	OLS	IV	OLS	IV	OLS	IV
Training	-.785		-3.756		-.860		-1.477	
	(3.292)		(4.830)		(3.647)		(5.772)	
Engaged		-2.068		-9.898		-2.266		-3.893
		(8.611)		(12.93)		(9.536)		(15.20)
Female	-6.012	-5.843	-13.89***	-13.08**	-7.987**	-7.802*	-16.06***	-15.74**
	(3.651)	(3.738)	(4.720)	(5.389)	(4.029)	(4.209)	(6.042)	(6.696)
P-value	.812	.810	.438	.444	.814	.812	.798	.798
Control Mean	10.62	10.62	13.89	13.89	12.15	12.15	15.56	15.56
Observations	227	227	227	227	227	227	227	227

Notes: This table shows the intent-to-treat and local average treatment effect estimates of SMS trainings on household and outside labor employed in primary business and across all businesses in the last 7 days at Endline, using the sample matched with Midline only. Coefficients across all Columns represent effects in terms of labor hours employed. Columns (1), (3), (5), and (7) show output from OLS regressions, while Columns (2), (4), (6), and (8) show output from 2SLS regressions where the endogenous variable is whether or not the individual engaged with training content. Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.4.23: Endline: Loan Amount Applied for and Received in Last 3 months by Age

	Full		**Age $\geq$ 34**		**Age $<$ 34**		**Diff**	
	Applied	Received	Applied	Received	Applied	Received	Applied	Received
Training	-667.9	-397.2	-2400.2	-973.9	2173.1	1156.1		
	(2321.2)	(2071.2)	(3633.6)	(3319.1)	(2660.8)	(2276.2)		
Train x Age $\geq$ 34							-4573.3	-2130.0
							(4503.7)	(4024.6)
P-value	0.774	0.848	0.509	0.769	0.414	0.612	0.310	0.597
Control Mean	20392.3	16813.0	25000	21000	15000	12000	20000	17000
Observations	2780	2780	1426	1426	1353	1353	2779	2779

Notes: This table shows the effect of treatment assignment on loan amount applied for and received in Kenyan Shillings in the last 3 months at Endline for the full sample (Columns (1), and (2), respectively), the sample with median and above age (Columns (3), and (4), respectively), the sample with below median age (Columns (5), and (6), respectively), and the difference in treatment effects across median and above, and below median age samples (Columns (7), and (8), respectively). Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

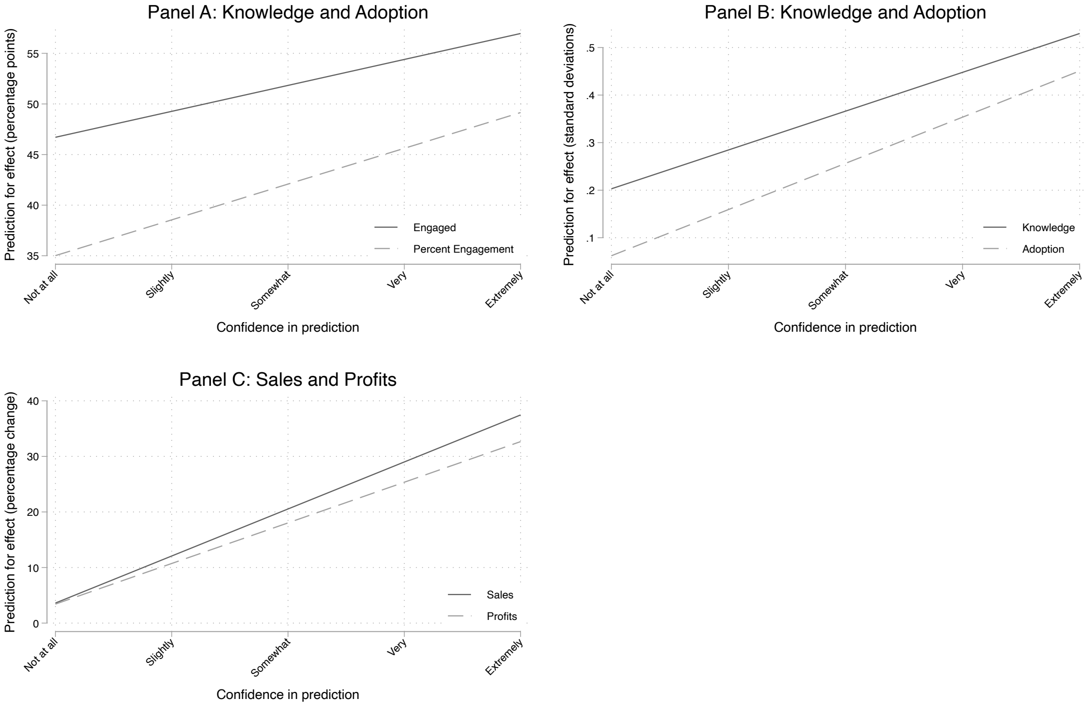
Table A.4.24: Endline: Loan Amount Applied for and Received in Last 3 months by Age Using Midline Sample

	Full		**Age ≥ 34**		**Age < 34**		**Diff**	
	Applied	Received	Applied	Received	Applied	Received	Applied	Received
Training	-4311.5 (5851.4)	-2953.9 (5222.6)	-7823.7 (8891.0)	-5418.3 (7796.6)	1186.8 (6960.9)	661.6 (6795.3)		
Train x Age ≥ 34							-9010.5 (11294.0)	-6079.9 (10342.6)
P-value	0.462	0.572	0.381	0.488	0.865	0.923	0.426	0.557
Control Mean	20970.2	17396.9	24000	19000	17000	15000	21000	17000
Observations	227	227	127	127	100	100	227	227

Notes: This table shows the effect of treatment assignment on loan amount applied for and received in Kenyan Shillings in the last 3 months at Endline, using the sample matched with Midline only, for the full sample (Columns (1), and (2), respectively), the sample with median and above age (Columns (3), and (4), respectively), the sample with below median age (Columns (5), and (6), respectively), and the difference in treatment effects across median and above, and below median age samples (Columns (7), and (8), respectively). Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

A.5 Predictions

Figure A.5.1: Predictions and Confidence behind Predictions



Notes: This figure shows the best-fit line representing the correlation between predictions of treatment effects and how confident the respondents reported to be in their predictions, ranging from not at all confident, to extremely confident.

Appendix B

Demand for SMS-Based Business Trainings Amongst Kenyan Micro-Entrepreneurs – Appendices

B.1 Summary Statistics and Balance

Table B.1.1: Summary Statistics

	TIOLI			BDM		
	Mean	SD	Obs.	Mean	SD	Obs.
Panel A: Baseline Covariates						
Female	0.44	0.50	380	0.53	0.50	103
Rural	0.49	0.50	380	0.11	0.31	103
Years of education	12.0	2.47	380	12.0	2.30	103
Age	33.6	8.59	380	36.3	9.34	103
Num of adults	2.69	1.44	379	2.45	1.24	102
Num of children	2.14	1.45	379	1.98	1.52	102
Panel B: Outcomes						
Knowledge	0.75	0.14	380	0.79	0.13	103
Adoption	0.67	0.17	351	0.64	0.21	103
Sales in last 30 days	59,754	111,199	380	69,238	113,295	103
Profits in last 30 days	19,602	31,122	380	23,852	32,155	103
Applied for loan	0.50	0.50	380	0.57	0.50	103
Missed loan payment	0.20	0.40	380	0.18	0.39	103
Hours worked on business	214.9	121.2	380	261.4	99.7	103
Hours worked on side jobs	23.6	64.1	380	17.5	47.6	103

Notes: This table shows the the mean, standard deviation and number of observations for pre-intervention covariates, and outcomes studied in the evaluation of the SMS-based training intervention. The summary statistics are presented separately for the Take-It-Or-Leave-It and Becker-DeGroot-Marschak demand elicitation samples.

Table B.1.2: Balance Across TIOLI Pricing Arms

Variable	P_1	P_2	P_3	Diff (P_1, P_2)	Diff (P_2, P_3)	Diff (P_1, P_3)
Panel A: Baseline Covariates						
Female	0.43	0.50	0.36	-0.06	0.13*	0.07
	(0.50)	(0.50)	(0.48)	(0.06)	(0.08)	(0.07)
Rural	0.51	0.48	0.47	0.03	0.01	0.04
	(0.50)	(0.50)	(0.50)	(0.06)	(0.08)	(0.07)
Years of education	12.00	12.15	11.93	-0.16	0.22	0.06
	(2.59)	(2.27)	(2.42)	(0.29)	(0.38)	(0.38)
Age	33.12	34.76	33.34	-1.64	1.42	-0.23
	(8.33)	(9.28)	(8.05)	(1.01)	(1.43)	(1.23)
Num of adults in household	2.76	2.63	2.57	0.12	0.06	0.19
	(1.43)	(1.52)	(1.33)	(0.17)	(0.24)	(0.21)
Num of children in household	2.21	2.12	1.93	0.09	0.19	0.28
	(1.45)	(1.43)	(1.51)	(0.17)	(0.24)	(0.22)
Panel B: Outcomes						
Knowledge	0.74	0.79	0.75	-0.05***	0.04*	-0.01
	(0.14)	(0.13)	(0.14)	(0.02)	(0.02)	(0.02)
Adoption	0.67	0.67	0.68	-0.01	-0.00	-0.01
	(0.16)	(0.18)	(0.19)	(0.02)	(0.03)	(0.03)
Sales in last 30 days	61449	60733	51741	715.76	8992	9707
	(109417)	(118607)	(103972)	(13162)	(18395)	(16068)
Profits in last 30 days	19731	19199	19922	532	-723	-192
	(31201)	(29921)	(33585)	(3592)	(5040)	(4709)
Applied for a loan	0.52	0.48	0.50	0.04	-0.02	0.02
	(0.50)	(0.50)	(0.50)	(0.06)	(0.08)	(0.07)
Loan payment missed/late	0.19	0.24	0.14	-0.05	0.10	0.05
	(0.39)	(0.43)	(0.35)	(0.05)	(0.07)	(0.06)
Hours worked on business in last 30 days	214.26	209.64	227.74	4.62	-18.10	-13.48
	(126.40)	(112.38)	(119.86)	(14.21)	(18.57)	(18.55)
Hours worked on side jobs in last 30 days	26.27	24.04	12.95	2.24	11.09	13.32
	(67.19)	(66.90)	(43.99)	(7.83)	(9.72)	(9.34)
Wald Chi-Sq. Test P-value = 0.26						

Notes: This table shows the balance of pre-intervention covariates across TIOLI pricing arms for the TIOLI sample overlapping with the Endline data. The reported p-value at the bottom is from a test of joint orthogonality across all randomization groups. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

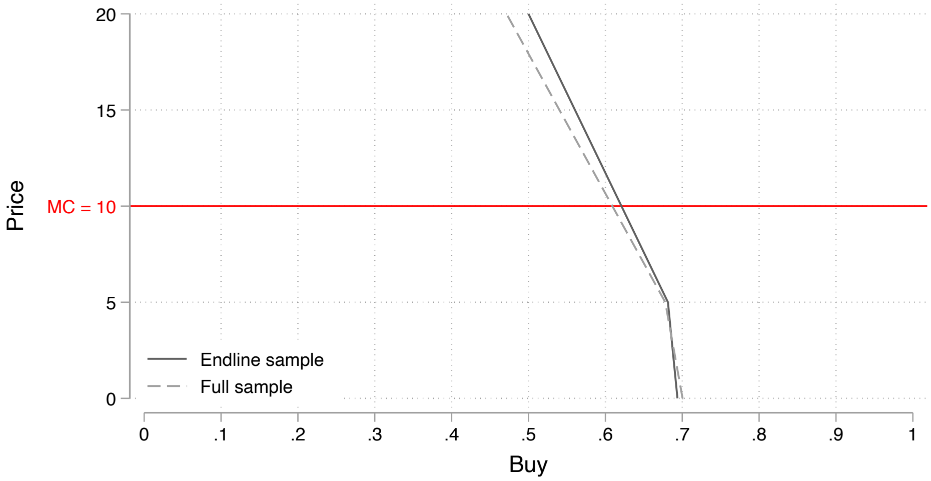
Table B.1.3: Balance Across Treatment and Control in BDM Sample

Variable	Treatment	Control	Difference
Female	0.51	0.58	-0.07
	(0.50)	(0.50)	(0.10)
Rural	0.14	0.05	0.09
	(0.35)	(0.23)	(0.06)
Years of education	11.86	12.21	-0.35
	(2.24)	(2.41)	(0.47)
Age	35.82	37.21	-1.40
	(9.74)	(8.66)	(1.91)
Num of adults in household	2.48	2.39	0.09
	(1.33)	(1.08)	(0.26)
Num of children in household	1.94	2.05	-0.12
	(1.45)	(1.66)	(0.31)
Knowledge	0.76	0.82	-0.06**
	(0.14)	(0.11)	(0.03)
Adoption	0.65	0.62	0.03
	(0.21)	(0.20)	(0.04)
Sales in last 30 days	69729.23	68397.37	1331.86
	(117252.40)	(107718.30)	(23249.43)
Profits in last 30 days	24344.31	23010.53	1333.78
	(31404.55)	(33812.18)	(6597.39)
Applied for a Loan	0.65	0.45	0.20**
	(0.48)	(0.50)	(0.10)
Loan payment missed/late	0.17	0.21	-0.04
	(0.38)	(0.41)	(0.08)
Hours worked on business in last 30 days	256.72	269.47	-12.75
	(104.83)	(91.06)	(20.42)
Hours worked on side jobs in last 30 days	19.88	13.47	6.40
	(50.94)	(41.56)	(9.74)
F-test p-value ($\beta_{diff} \neq 0$)			0.13

Notes: This table shows the balance of pre-intervention covariates across TIOLI pricing arms for the TIOLI sample overlapping with the Endline data. The reported p-value at the bottom is from a test of joint orthogonality across all randomization groups. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

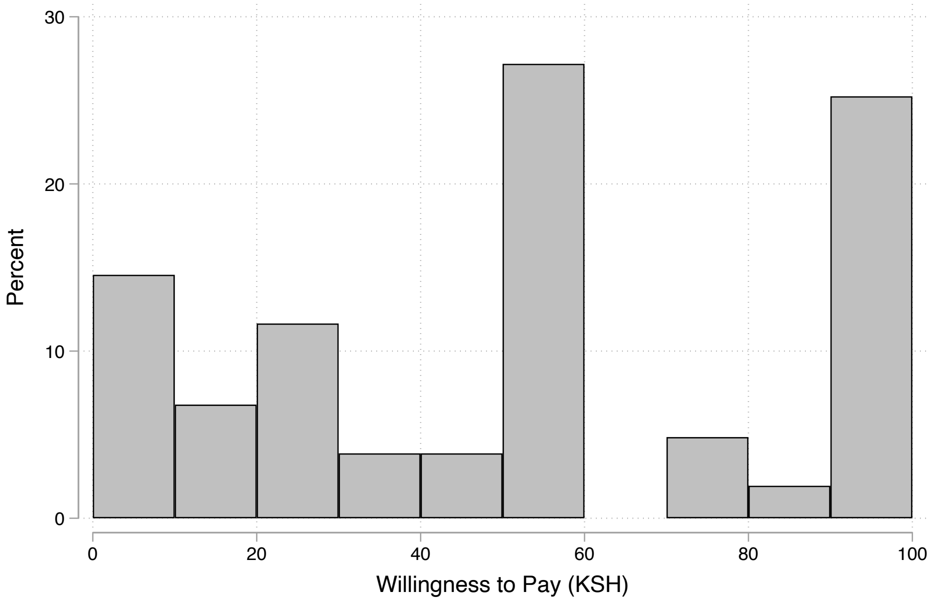
B.2 Demand for SMS Trainings

Figure B.2.1: TIOLI Demand Curve - Full Sample vs. Endline Sample



Notes: This figure shows the (inverse) demand curves based on buying decisions from randomized TIOLI offers sent to treatment individuals in the full TIOLI sample and that overlapping with the Endline sample. The horizontal red line represents the per person marginal cost faced by the service provider for delivering the entire training.

Figure B.2.2: BDM Willingness to Pay Distribution



Notes: This figure shows the distribution of maximum willingness to pay measured through the in-person BDM elicitation exercise.

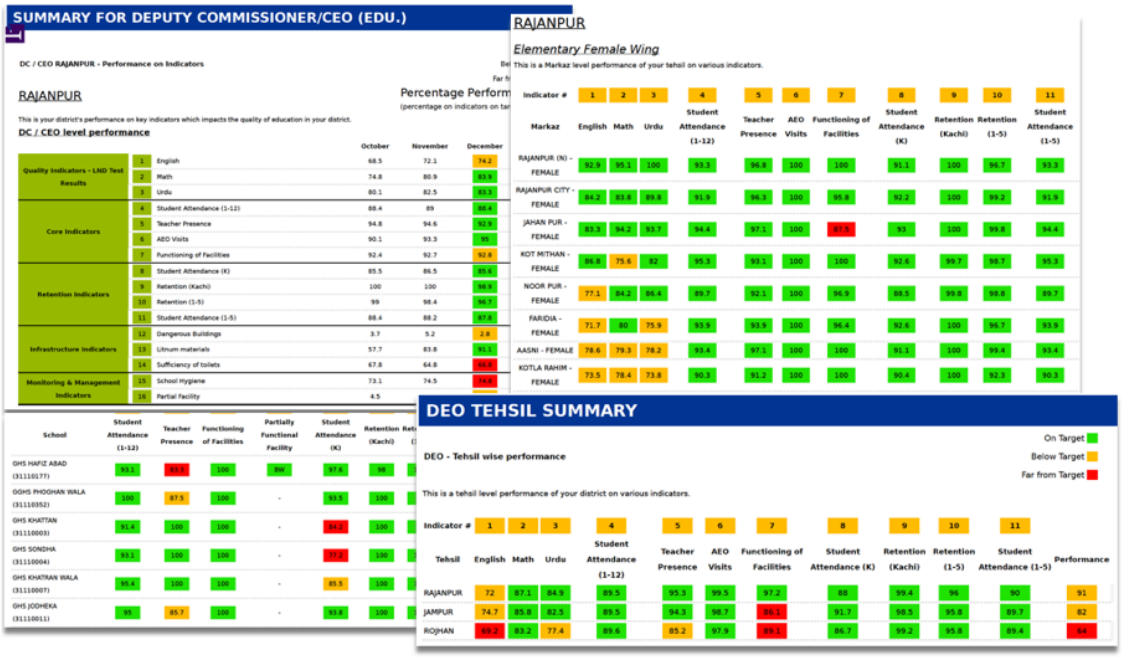
Appendix C

Command and Can’t Control: Assessing Centralized Accountability in the Public Sector – Appendices

C.1 Data and Design Details

C.1.1 Data pack

Figure C.1.1: Data pack screenshot



SUMMARY FOR DEPUTY COMMISSIONER/CEO (EDU.)

DC / CEO RAJANPUR - Performance on Indicators

RAJANPUR

Percentage Perform
(percentage on indicators on tar

This is your district's performance on key indicators which impacts the quality of education in your district.

DC / CEO level performance

			October	November	December
Quality Indicators - LND Test Results	1	English	68.5	72.1	74.2
	2	Math	74.8	80.9	83.9
	3	Urdu	80.1	82.5	83.3
Core Indicators	4	Student Attendance (1-12)	88.4	89	88.4
	5	Teacher Presence	94.8	94.6	92.9
	6	AEO Visits	90.1	93.3	95
	7	Functioning of Facilities	92.4	92.7	92.8
Retention Indicators	8	Student Attendance (K)	85.5	86.5	85.6
	9	Retention (Kachi)	100	100	98.9
	10	Retention (1-5)	99	98.4	96.7
	11	Student Attendance (1-5)	88.4	88.2	87.8
Infrastructure Indicators	12	Dangerous Buildings	3.7	5.2	2.8
	13	Litnum materials	57.7	83.8	91.1
	14	Sufficiency of toilets	67.8	64.8	66.8
Monitoring & Management Indicators	15	School Hygiene	73.1	74.5	74.8
	16	Partial Facility	4.5		

School	Student Attendance (1-12)	Teacher Presence	Functioning of Facilities	Partially Functional Facility	Student Attendance (K)	Retention (Kachi)
GHS HAFIZ ABAD (31110177)	93.1	83.3	100	BW	97.6	98
GGHS PHOGHAN WALA (31110352)	100	87.5	100	-	93.5	100
GHS KHATTAN (31110003)	91.4	100	100	-	84.2	100
GHS SONDHA (31110004)	93.1	100	100	-	77.2	100
GHS KHATRAN WALA (31110007)	95.4	100	100	-	85.5	100
GHS JODHEKA (31110011)	95	85.7	100	-	93.8	100

RAJANPUR

Elementary Female Wing

This is a Markaz level performance of your tehsil on various indicators.

Indicator #	1	2	3	4	5	6	7	8	9	10	11
Markaz	English	Math	Urdu	Student Attendance (1-12)	Teacher Presence	AEO Visits	Functioning of Facilities	Student Attendance (K)	Retention (Kachi)	Retention (1-5)	Student Attendance (1-5)
RAJANPUR (N) - FEMALE	92.9	95.1	100	93.3	96.8	100	100	91.1	100	96.7	93.3
RAJANPUR CITY - FEMALE	84.2	83.8	89.8	91.9	96.3	100	95.8	92.2	100	99.2	91.9
JAHAN PUR - FEMALE	83.3	94.2	93.7	94.4	97.1	100	87.5	93	100	99.8	94.4
KOT MITHAN - FEMALE	86.8	75.6	82	95.3	93.1	100	100	92.6	99.7	98.7	95.3
NOOR PUR - FEMALE	77.1	84.2	86.4	89.7	92.1	100	96.9	88.5	99.8	98.8	89.7
FARIDIA - FEMALE	71.7	80	75.9	93.9	93.9	100	96.4	92.6	100	96.7	93.9
AASNI - FEMALE	78.6	79.3	78.2	93.4	97.1	100	100	91.1	100	99.4	93.4
KOTLA RAHIM - FEMALE	73.5	78.4	73.8	90.3	91.2	100	100	90.4	100	92.3	90.3

DEO TEHSIL SUMMARY

DEO - Tehsil wise performance

On Target 🟩
Below Target 🟨
Far from Target 🟥

This is a tehsil level performance of your district on various indicators.

Indicator #	1	2	3	4	5	6	7	8	9	10	11	
Tehsil	English	Math	Urdu	Student Attendance (1-12)	Teacher Presence	AEO Visits	Functioning of Facilities	Student Attendance (K)	Retention (Kachi)	Retention (1-5)	Student Attendance (1-5)	Performance
RAJANPUR	72	87.1	84.9	89.5	95.3	99.5	97.2	88	99.4	96	90	91
JAMPUR	74.7	85.8	82.5	89.5	94.3	98.7	86.1	91.7	98.5	95.8	89.7	82
ROJHAN	69.2	83.2	77.4	89.6	85.2	97.9	89.1	86.7	99.2	95.8	89.4	64

C.1.2 Color-coded performance thresholds

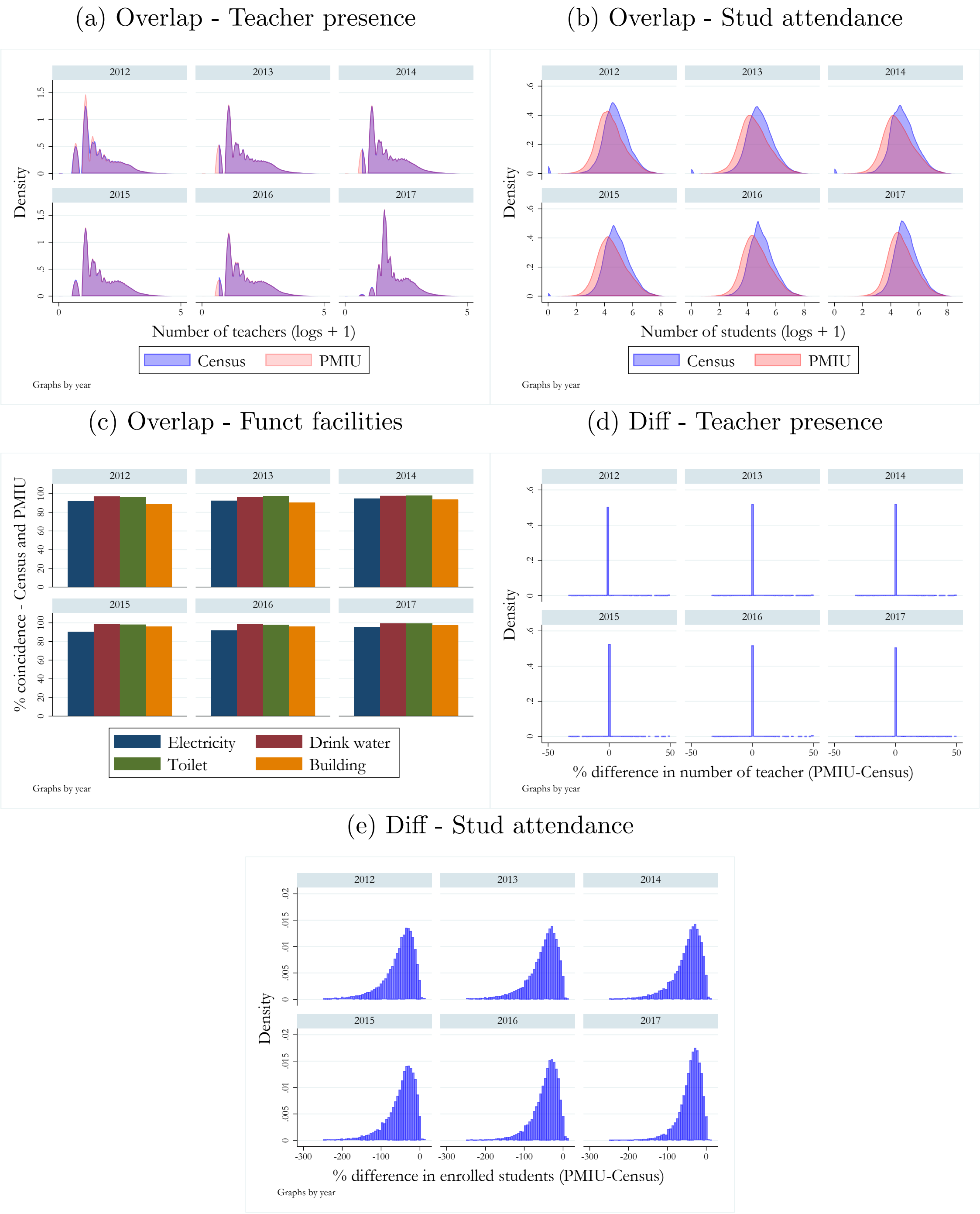
Teacher presence at every aggregation was coded red when it fell below 86%, orange when it was 86% and above but below 90%, and green when it was 90% or higher. These thresholds were the same for all districts and for all months of the year. Functioning facilities thresholds were 90% and 95%, and were the same across all districts and months of the year.

Thresholds for student attendance varied across districts and months. Districts were divided into three categories, A, B and C, where category A consisted of historically highest performing districts, category C consisted of historically lowest performing districts, and B consisted of the rest. Further, the months in the year were divided into two groups - December-March were considered high attendance months and April-November were considered low attendance months. This division accounted for differential attendance expected due to exams during the school year. Different thresholds were set for each category of districts, for each group of months; for category A districts during December-March, student attendance was coded red if it was below 89%, orange if it was 89% and above but below 92%, and green if it was 92% and above. During April-November, the thresholds were 87% and 90%. For category B districts the thresholds were 87% and 90% during December-March and 85% and 88% during April-November. For category C districts the thresholds were 84% and 87% during December-March and 82% and 85% during April-November.

C.1.3 Compliance

We have data-pack reports for 60 months from December 2011 to May 2018, which account for 100% of the reporting (without June, July, and August). To asses the quality of the data we compare the data-pack reports with the annual school census in the month that the relevant census was undertaken (October). Figure C.1.2 compares the distribution of the variables both sources report. Panels (a) and (b) show for teacher presence and student attendance that both sources overlap, suggesting that the population was mapped consistently. Panel (c) plots the percentage of schools where the functionality coincides, which is near 100%. Panel (d) and (e) plots the distribution of the differences in the reporting. For teacher presence, we observe almost no difference. For enrolled students, it shows a high mass around zero, though a tail of negative values.

Figure C.1.2: Data validation - monthly PMIU vs. Census



(a) Overlap - Teacher presence

(b) Overlap - Stud attendance

(c) Overlap - Funct facilities

(d) Diff - Teacher presence

(e) Diff - Stud attendance

Note: This figure compares October PMIU data and corresponding school-level quantities from the Annual School Census. Panel (a) and (b) plot the distribution of (log+1) teachers and students. Panel (c) plots the coincidence in the reporting of functional facilities (= 1 if functional). Panel (d) and (e) plots the distribution of school differences as percentage change (PMIU - Census)/PMIU dropping the data below percentile 1 and above percentile 99.

C.1.4 Stacking process

Figure C.1.3: Stacking process

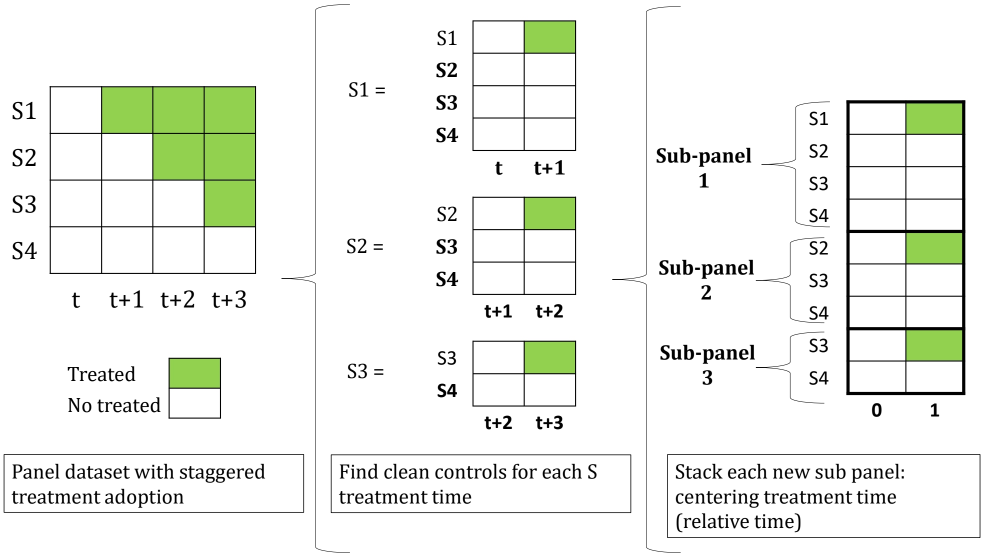

S1
S2
S3
S4
t
t+1
t+2
t+3
Treated
No treated
Panel dataset with staggered treatment adoption
S1 =
S2 =
S3 =
Find clean controls for each S treatment time
Sub-panel 1
Sub-panel 2
Sub-panel 3
0
1
Stack each new sub panel: centering treatment time (relative time)

Figure C.1.3 describe the stacking procedure. Each row/column corresponds to a subject/period treatment status. Green indicates treatment. We (arbitrarily) choose one period before and the period of treatment adoption. For S_1, in t_{+1}, units $S_{2,3,4}$ are controls. For S_2 in t_{+2} units $S_{3,4}$ are not treated. For S_3 in t_{+3}, unit S_4 is not treated. For each treated unit we build a two-period panel with its own controls, assign a unique identifier for each, and stack them together by normalizing in relative time so no bias from treatment timing adoption appears from using two-way fixed effects. As the same unit can appear at different events, the fixed effects must be interacted with panel identifiers to account for repeated units and differences in relative time origins.

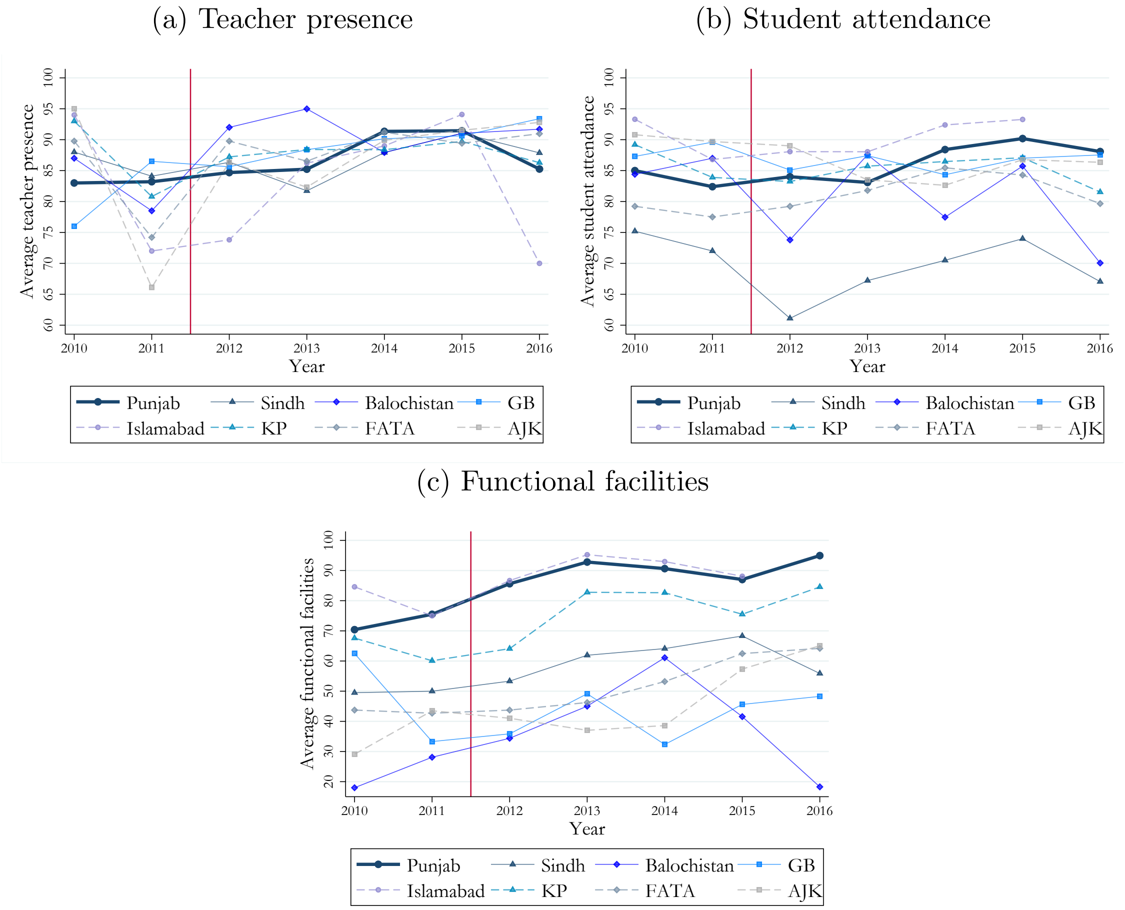
C.1.5 Ranking of district officer positions

District officers can be rewarded/punished in terms of transfers to more/less preferred postings based on performance. To estimate the effect of the oversight scheme on the career trajectory, we collected information on the postings for each senior officer before and after they were posted as district officers. We ranked all designations by seniority to ascertain whether a officer was rewarded/punished determining if a change in position was a promotion/demotion. The ranking of designations was generated through extensive research about seniority levels within the Pakistani bureaucracy and was vetted by two senior bureaucrats.

C.2 Additional Results and Robustness

C.2.1 Immediate impact of monitoring system implementation

Figure C.2.1: Pakistan provinces average outcomes trends

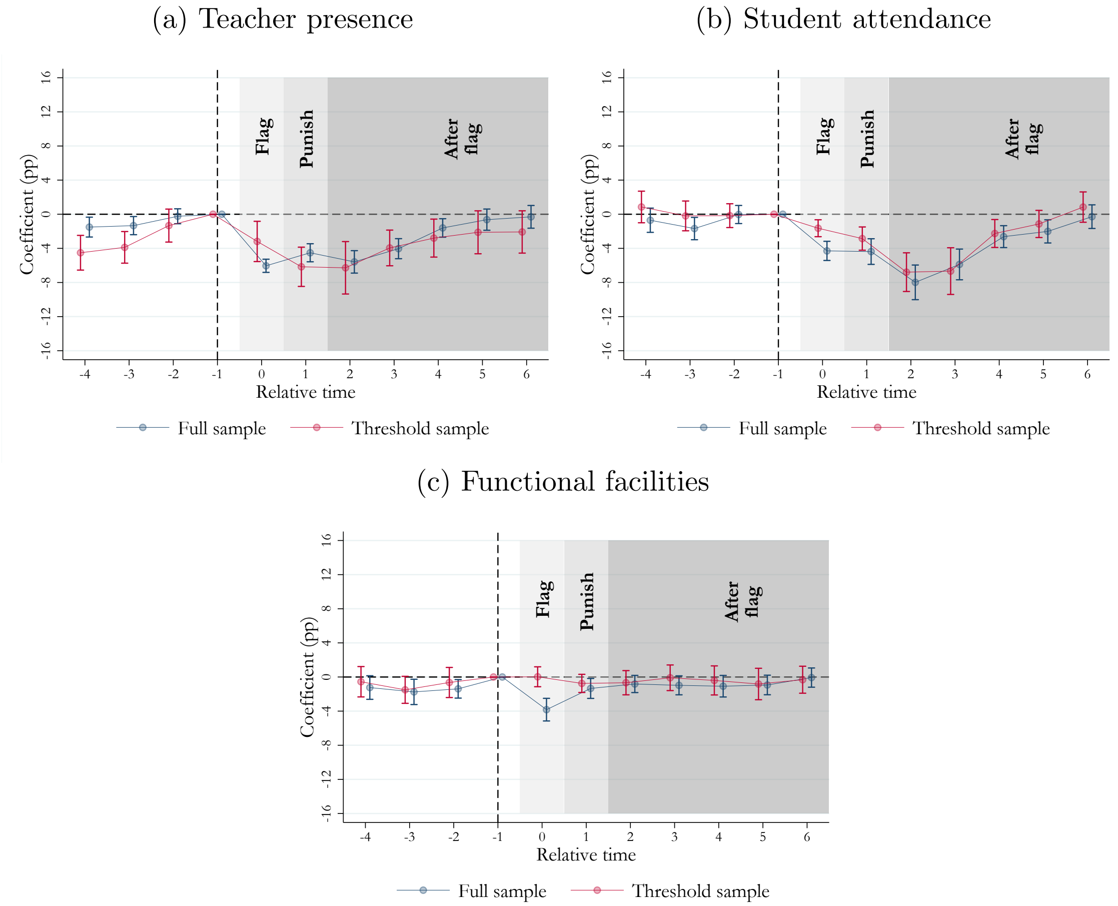


Note: The figure show the Data from ASER Pakistan (`aserpakistan.org`)

We assess whether the lack of an effect from flagging might be explained by a general impact of the policy across Punjab. Figure C.2.1 shows the average trends of education outcomes in all Pakistan provinces.[1] Note that most provinces are either improving or in a similar trend to Punjab (darker blue line). So despite some underperforming provinces, most of the country faces similar evolving trends.

[1] We recover province-level data for the period 2010-2016 from the Annual Status of Education Report - ASER - Pakistan (`aserpakistan.org`), which have been independently and consistently conducting household and school surveys to assess the education advancements in the country.

Figure C.2.2: Event study - first time flagging effect on performance - flagged units


(a) Teacher presence

(b) Student attendance

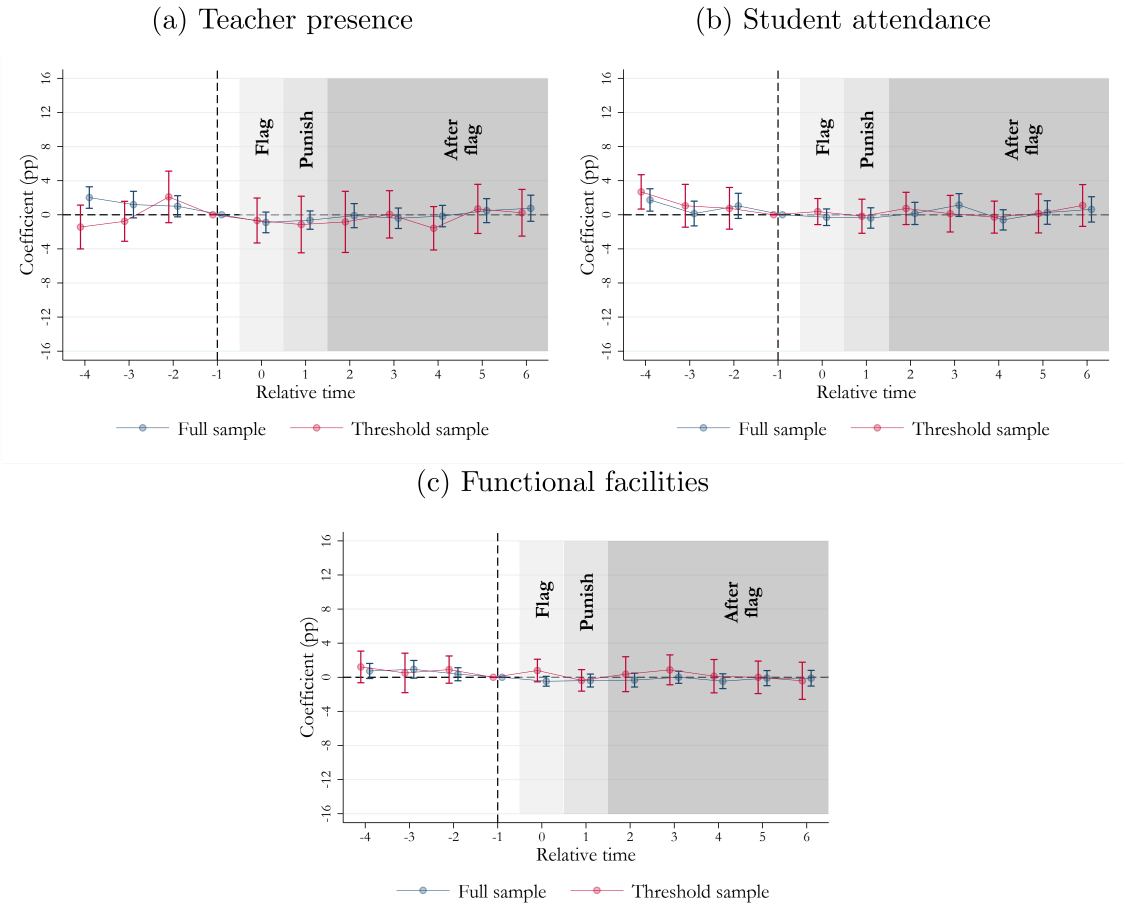
(c) Functional facilities

Note: This figure displays the γ_i coefficients from an event study based on equation C.1, only for the first month of the oversight scheme implementation, using -1 as the base period, and comparing schools in flagged and non-flagged maraakiz. The blue line presents results for the full sample, while the red line presents results for the threshold sample, obtained through regression discontinuity optimization methods. The results are for flagging on the variable in the title of the panel. *Flag* is for the period in which the information is collected, and the markaz is flagged. *Punish* is for the period where the reports are distributed and the oversight meeting with the punishment occurs. *After flag* is for periods after the oversight meeting occurs. Error bars at the 95 percent level are presented for each coefficient. Error bars at the 95 percent level are presented for each coefficient.

We also test the oversight scheme's immediate impacts by studying the effect of being flagged in the first month of implementation and being a neighbor of a flagged unit. We estimate for the first month of implementation a modified equation 3.1 including an additional treatment for maraakiz with a school neighboring a flagged markaz. In the equation below $N_{mde} = 1$ represents neighbors, γ_i coefficients for the first time flagged, and β_i for the effect of flagging on neighbors of flagged units.

$$
\begin{aligned}
Y_{smdte} = \gamma_1(T_{mde} \times Flag_{te}) + \gamma_2(T_{mde} \times Punish_{te}) + \gamma_3(T_{mde} \times AfterFlag_{te}) + \\
\beta_1(N_{mde} \times Flag_{te}) + \beta_2(N_{mde} \times Punish_{te}) + \beta_3(N_{mde} \times AfterFlag_{te}) + \\
\alpha_{mde} + \lambda_{te} + dt + \epsilon_{smdte}
\end{aligned}
\tag{C.1}
$$

Figure C.2.3: Event study - first time flagging effect on performance - neighbor units



(a) Teacher presence

(b) Student attendance

(c) Functional facilities

Note: This figure displays the β_i coefficients from an event study based on equation C.1, only for the first month of the oversight scheme implementation, using -1 as the base period, and comparing schools in flagged and non-flagged maraakiz. The blue line presents results for the full sample, while the red line presents results for the threshold sample, obtained through regression discontinuity optimization methods. The results are for flagging on the variable in the title of the panel. *Flag* is for the period in which the information is collected, and the markaz is flagged. *Punish* is for the period where the reports are distributed and the oversight meeting with the punishment occurs. *After flag* is for periods after the oversight meeting occurs. Error bars at the 95 percent level are presented for each coefficient. Error bars at the 95 percent level are presented for each coefficient.

Figure C.2.2 reports the results for the first time flagged units in the full (blue) and threshold (red) samples. We observe no positive effects. Instead, both samples of teacher presence and student attendance suggest that flagged units took more time to recover from the negative shock that leads them to be flagged for the first time. Figure C.2.3 reports the results for the neighbors of flagged maraakiz. We observe no positive effects.

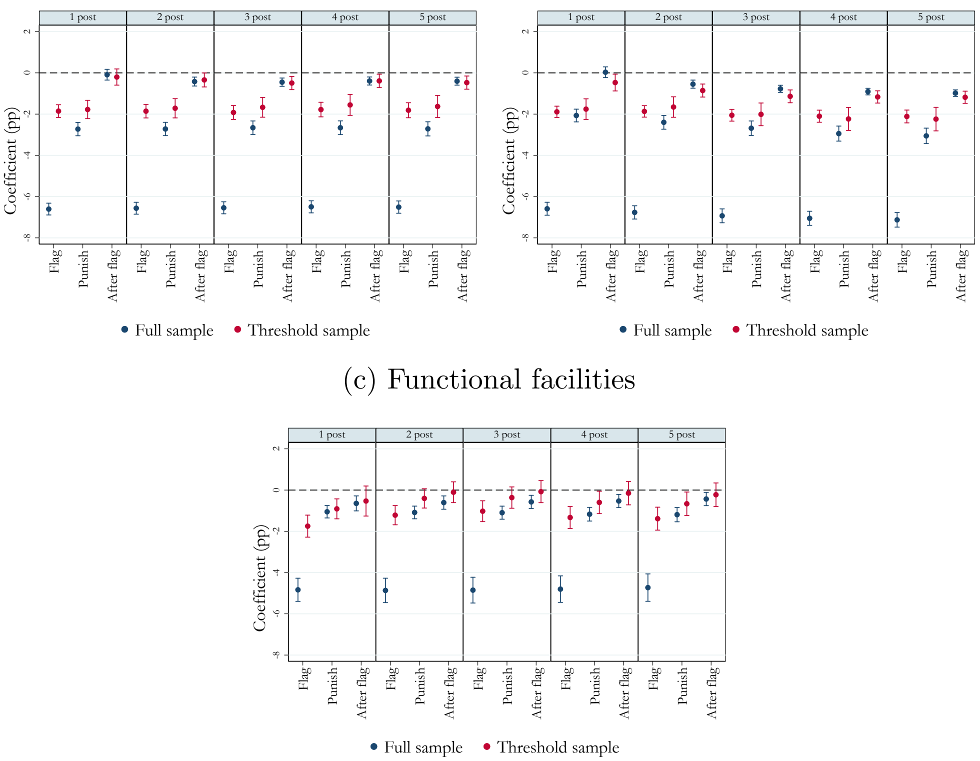
C.2.2 Robustness of the stacked design

Since the flagging might turn on/off, the election of post-periods leads us to assume that the unit remains treated, and we might be losing information. We present results to different post-period window stacking, and additional estimators. Figure C.2.4 plots the estimates from estimating equation 3.1 with a stacking including t periods. $Flag$ show the temporary nature of the negative shock. The coefficients of $Punish$ remain qualitatively similar, showing the immediate recovery. The $AfterFlag$ coefficients remain close to zero. Figure C.2.5 reports the event study for a shorter stacking, showing that the trends before and after suggest a similar evolving path as Figure 3.5. Figure C.2.6 estimates the event study following Sun and Abraham (2021) on the non-stacked dataset, under the assumption of staggered treatment timing, so flagged maraakiz remain treated after the first occurrence.[2] Figure C.2.7 estimates the event study using the DID_l estimator following De Chaisemartin and D'Haultfoeuille (2022), which allows to consider the effect of those switching on/off the treatment, which is also robust to differences in treatment timing.[3]

[2] The authors show that in TWFE dynamic specification with staggered adoption, leads/lag coefficients are contaminated by the effect on other relative periods. It is an special case of Callaway and Sant'Anna (2021) with no covariates (Baker, Larcker and Wang (2022)).

[3] We use a lower number of post periods as the dynamic estimator is obtained as a weighted average of difference in differences comparing the t and $t-l-1$ outcome evolution, between switchers in $t-l$ and non-switchers cohorts (De Chaisemartin and D'Haultfoeuille (2022)).

Figure C.2.4: Average effects by additional after flag t flagging effect on performance



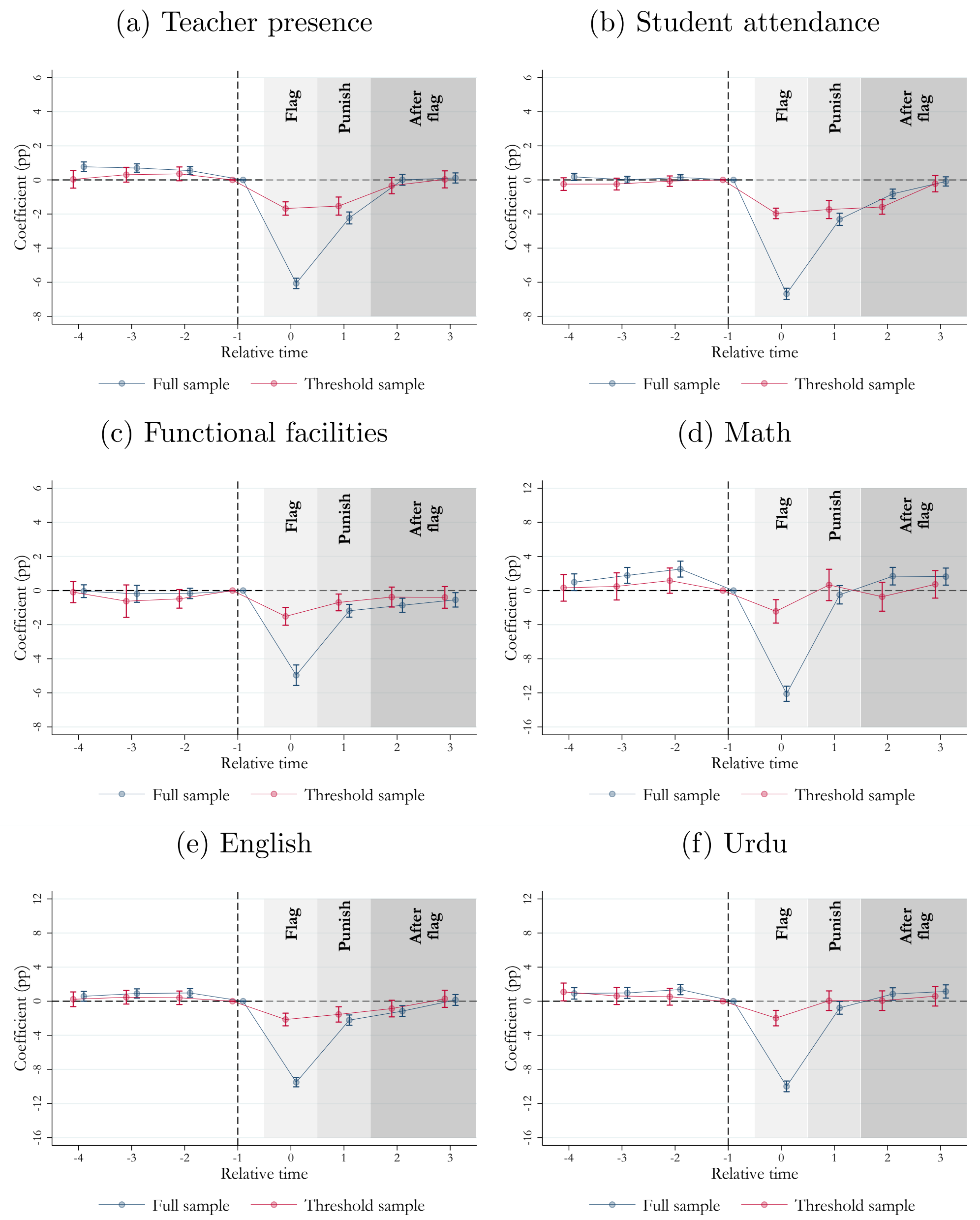
(a) Teacher presence

(b) Student attendance

(c) Functional facilities

Note: This figure presents results from estimating equation 3.1 for a stacked dataset including t additional post-periods. The blue and red coefficients present results for the full and threshold samples, respectively. The results are for flagging on the variable in the title of the panel. *Flag* is for the period in which the information is collected, and the markaz is flagged. *Punish* is for the period where the reports are distributed and the oversight meeting with the punishment occurs. *After flag* is for periods after the oversight meeting occurs. Error bars represent 95% confidence intervals.

Figure C.2.5: Event study - flagging effect on performance short stack



(a) Teacher presence

(b) Student attendance

(c) Functional facilities

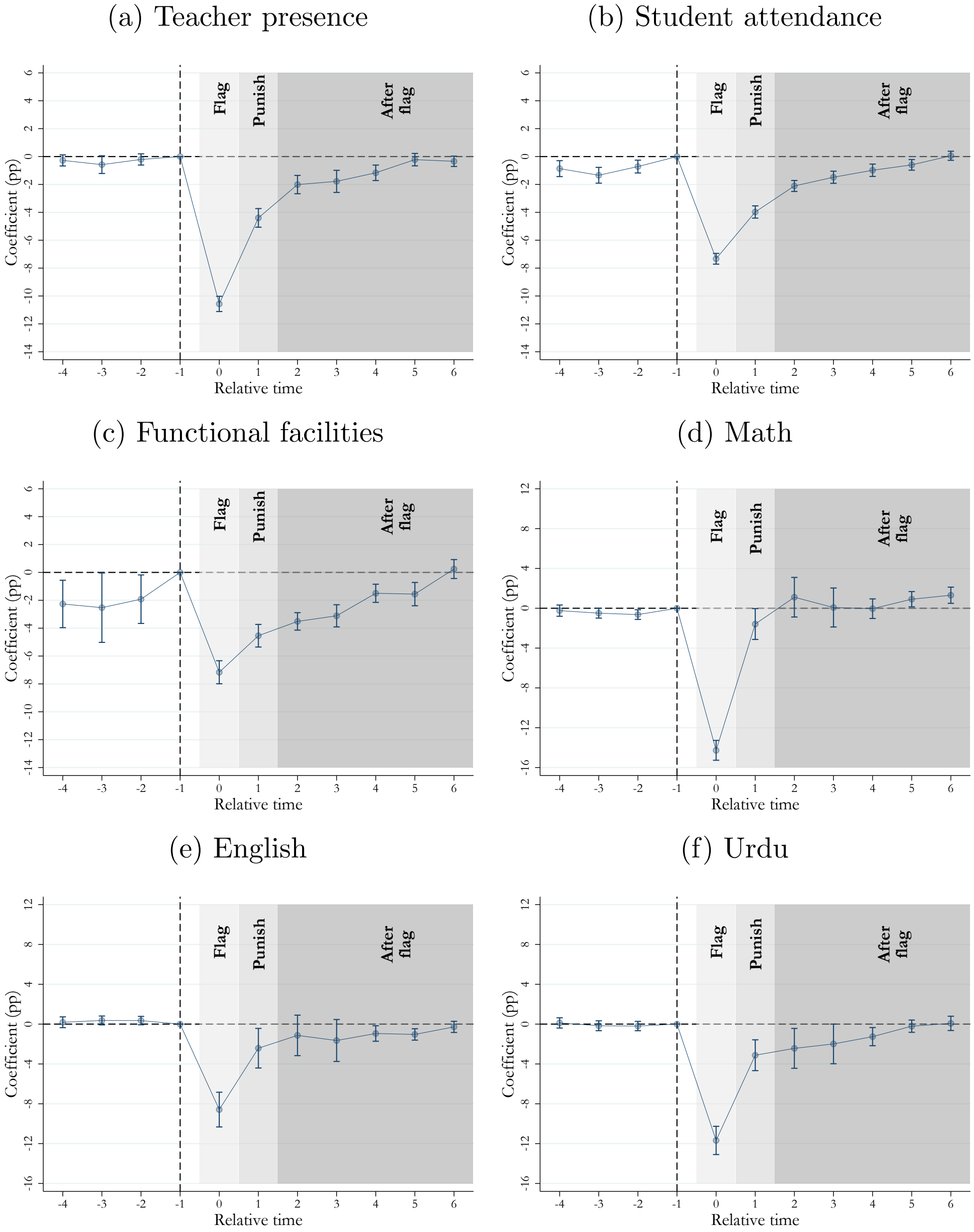
(d) Math

(e) English

(f) Urdu

Note: This figure presents event study graphs based on equation 3.1 using -1 as the base period, comparing schools in flagged and non-flagged maraakiz. The blue and red lines present results for the full and threshold samples, respectively. The results are for flagging on the variable in the title of the panel. *Flag* is for the period in which the information is collected, and the markaz is flagged. *Punish* is for the period where the reports are distributed and the oversight meeting with the punishment occurs. *After flag* is for periods after the oversight meeting occurs. Error bars represent 95% confidence intervals.

Figure C.2.6: Alternative specifications
Sun and Abraham (2021)



(a) Teacher presence

(b) Student attendance

(c) Functional facilities

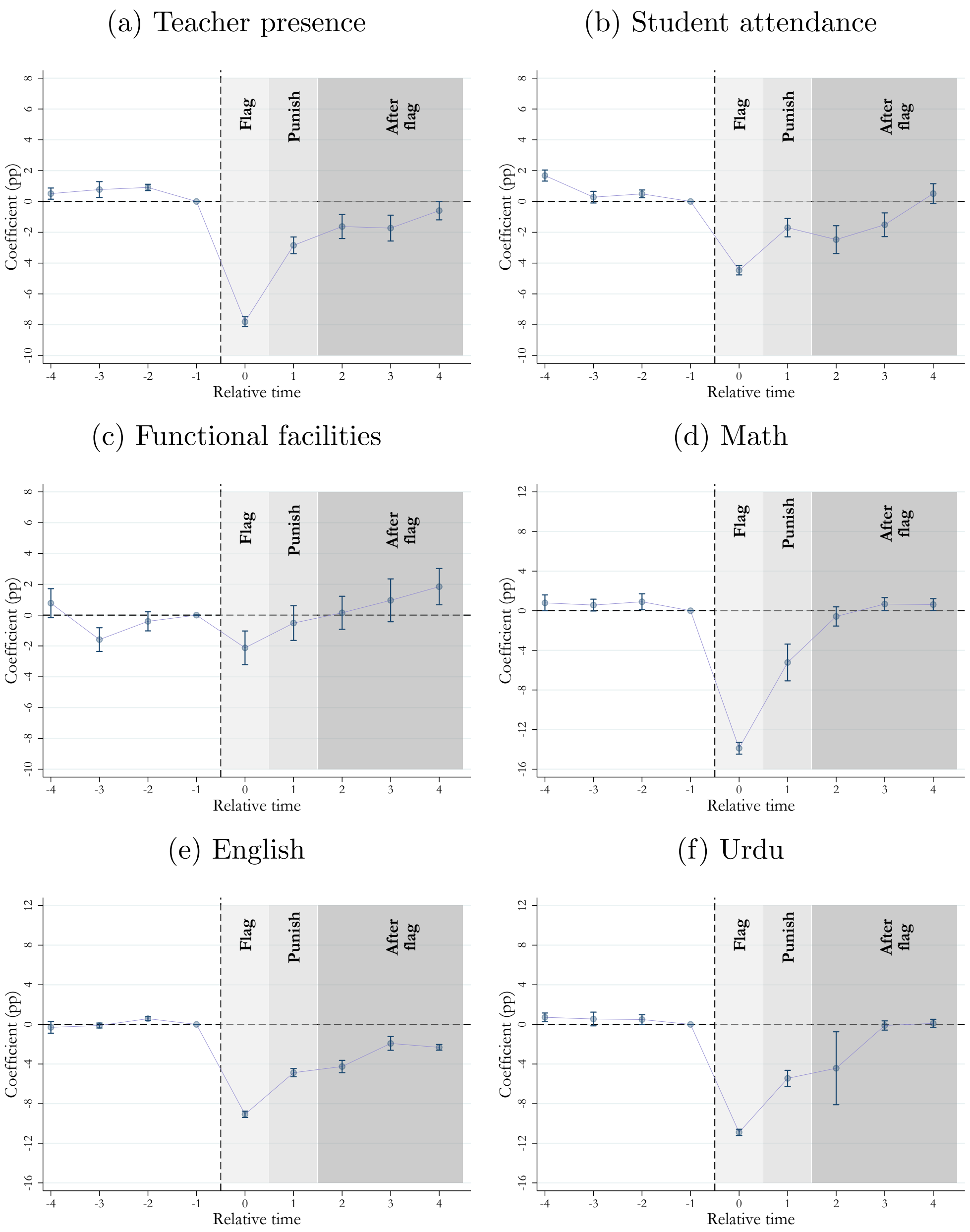
(d) Math

(e) English

(f) Urdu

Note: This figure presents the results from estimating an event study based on the Sun and Abraham (2021) difference-in-differences estimator, using -1 as the base period, comparing schools in flagged and non-flagged maraakiz. The results are for flagging on the variable in the title of the panel. *Flag* is for the period in which the information is collected, and the markaz is flagged. *Punish* is for the period where the reports are distributed and the oversight meeting with the punishment occurs. *After flag* is for periods after the oversight meeting occurs. Error bars at the 95 percent level are presented for each coefficient.

Figure C.2.7: Alternative specifications DID$_l$ De Chaisemartin and D'Haultfoeuille (2022)



(a) Teacher presence

(b) Student attendance

(c) Functional facilities

(d) Math

(e) English

(f) Urdu

Note: This figure presents event study graphs based on the DID$_l$ De Chaisemartin and D'Haultfoeuille (2022) difference-in-differences estimator, using -1 as the base period, and three placebo periods before the treatment, comparing schools in flagged and non-flagged maraakiz. The results are for flagging on the variable in the title of the panel. *Flag* is for the period in which the information is collected, and the markaz is flagged. *Punish* is for the period where the reports are distributed and the oversight meeting with the punishment occurs. *After flag* is for periods after the oversight meeting occurs. Error bars represent 95% confidence intervals.

C.2.3 Robustness of modeling approach

Table C.2.1 reports the average effect from estimating from equation 3.1 using flagging history fixed effects, comparing maraakiz that had the same flagging path before the negative shock. Flagging history is not a markaz attribute, so the term T_{mde} from equation 3.1 is not absorbed and the interactions can be compared against it. However, by conditioning on past flagging we are generating dependencies in the estimation that make our results more difficult to interpret. Here, we do not have a clean treatment period, and the pre-period now contains some maraakiz that have been flagged, such that the coefficient for T_{mde} is negative for all dependent variables. The average recovery to the mean of these maraakiz then yields a slightly positive coefficient on *After flag* in this specification. However, the net effect of these two coefficients yield qualitatively the same results as in our other tables.

We test for alternative margins of flagging. Figure C.2.8 test the effect of being orange flagged. The results suggest no positive effects after the negative shock. Additionally, Figure C.2.9 test for tehsil level red flagging. The results suggest a non-significant impact of flagging on performance. We also test for changes in the data pack structure involving an increase in the amount of data reported after December 2015 and January 2017. Table C.2.2 show the average results from equation 3.1, separating by data packs structure, with similar results as the effects on the after-flag period are always closer to zero or non-significant. Finally, Table C.2.3 presents heterogeneity by the coincidence between flagging and district meetings, from a modified version of equation 3.1 including the triple interaction between being flagged and being in a top/bottom district after the flagging. The results show no differential effect.

Table C.2.1: Monitoring effect on performance - markaz flagging - flagging history FE

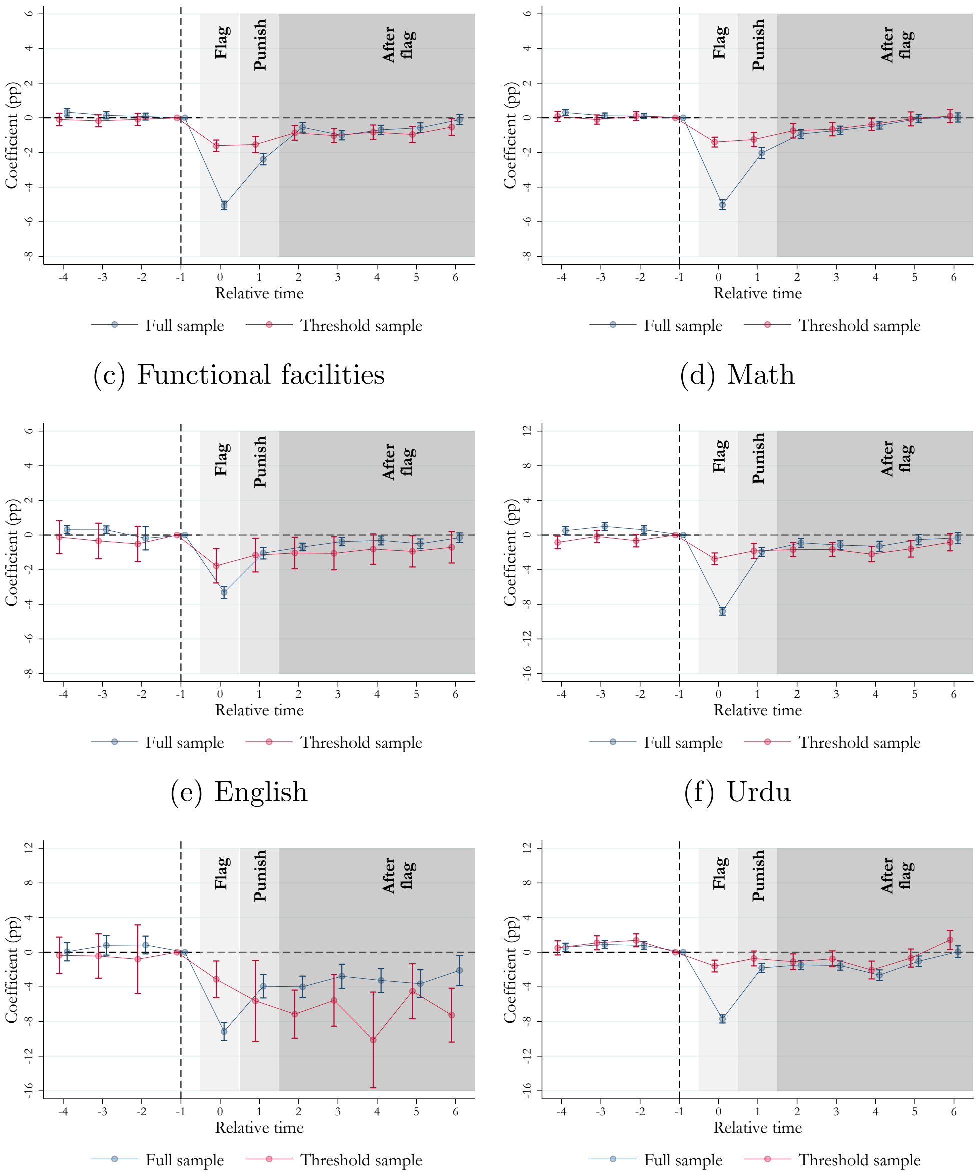
Panel A: School outcomes						
Dependent variable:	Teacher presence		Student attendance		Functional facilities	
T	-2.46***	-0.98***	-4.60***	-2.44***	-8.79***	-2.57***
	(0.097)	(0.10)	(0.15)	(0.14)	(0.27)	(0.12)
T×Flag	-4.47***	-1.11***	-3.45***	0.31**	0.93***	0.69***
	(0.12)	(0.13)	(0.14)	(0.16)	(0.13)	(0.12)
T×Punish	-0.76***	-0.65***	-0.27**	-0.13	2.24***	0.30**
	(0.10)	(0.15)	(0.13)	(0.18)	(0.14)	(0.13)
T×After flag	1.51***	0.35***	2.90***	1.09***	4.83***	1.08***
	(0.11)	(0.12)	(0.17)	(0.16)	(0.22)	(0.12)
N. of obs.	10,331,439	1,870,451	9,414,676	2,192,023	11,636,860	1,994,035
Mean Dep. Var. before	91.4	87.2	88.4	86.3	93.3	91.2
R^2	0.032	0.026	0.17	0.095	0.17	0.039
Panel B: Student scores						
Dependent variable:	Math		English		Urdu	
T	-3.50***	-0.71	-4.18***	-1.96***	-3.72***	-1.33***
	(0.35)	(0.43)	(0.16)	(0.18)	(0.24)	(0.30)
T×Flag	-12.7***	-2.17***	-7.65***	-2.39***	-9.88***	-1.91***
	(0.39)	(0.60)	(0.17)	(0.22)	(0.27)	(0.38)
T×Punish	-1.39***	-0.43	-1.40***	-1.71***	-0.80**	-0.63
	(0.46)	(0.76)	(0.17)	(0.26)	(0.32)	(0.53)
T×After flag	1.80***	0.19	1.79***	0.22	2.06***	0.80**
	(0.35)	(0.58)	(0.18)	(0.21)	(0.24)	(0.37)
N. of obs.	2,281,495	57,196	1,607,728	590,575	2,104,390	150,442
Mean Dep. Var. before	86.6	71.4	74.9	70.3	84.2	71.8
R^2	0.066	0.12	0.060	0.050	0.074	0.11
Sample	Full	Threshold	Full	Threshold	Full	Threshold
Flag history FE	Yes	Yes	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes	Yes	Yes
District time trends	Yes	Yes	Yes	Yes	Yes	Yes

Note: Results from estimating a modified version of equation 3.1, including flagging history FE instead of markaz FE. Flagging history is built from concatenating the flagging status in the three periods before the observed flagging. The school is the unit of observation for both panels. The first column for each outcome estimates for the full sample. The second column for each outcome estimates for the threshold sample, including schools in maraakiz that lie within the bandwidth obtained through regression discontinuity optimization methods. The flagging and threshold sample are based on the studied outcome. Teacher presence and student attendance are measured as the percentage of present teachers/students relative to the total teachers/students reported. The functional facilities variable is measured as the percentage of the school's functional infrastructure, including toilets, drinking water, boundary wall, and electricity. Math, English, and Urdu scores are measured as the percentage of correct answers in standardized tests. T equals 1 for schools in a flagged markaz. *Flag* equals 1 for the period in which the information is collected, and the markaz is flagged. *Punish* equals 1 for the period where the reports are distributed and the oversight meeting with the punishment occurs. *After flag* is equal to 1 for periods after the oversight meeting occurs. *Mean. Dep. Var before* shows the average outcome in the non-flagged maraakiz before the flagging occurs. Standard errors clustered by markaz, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figure C.2.8: Event study - flagging effect on performance orange threshold

(a) Teacher presence

(b) Student attendance



(c) Functional facilities

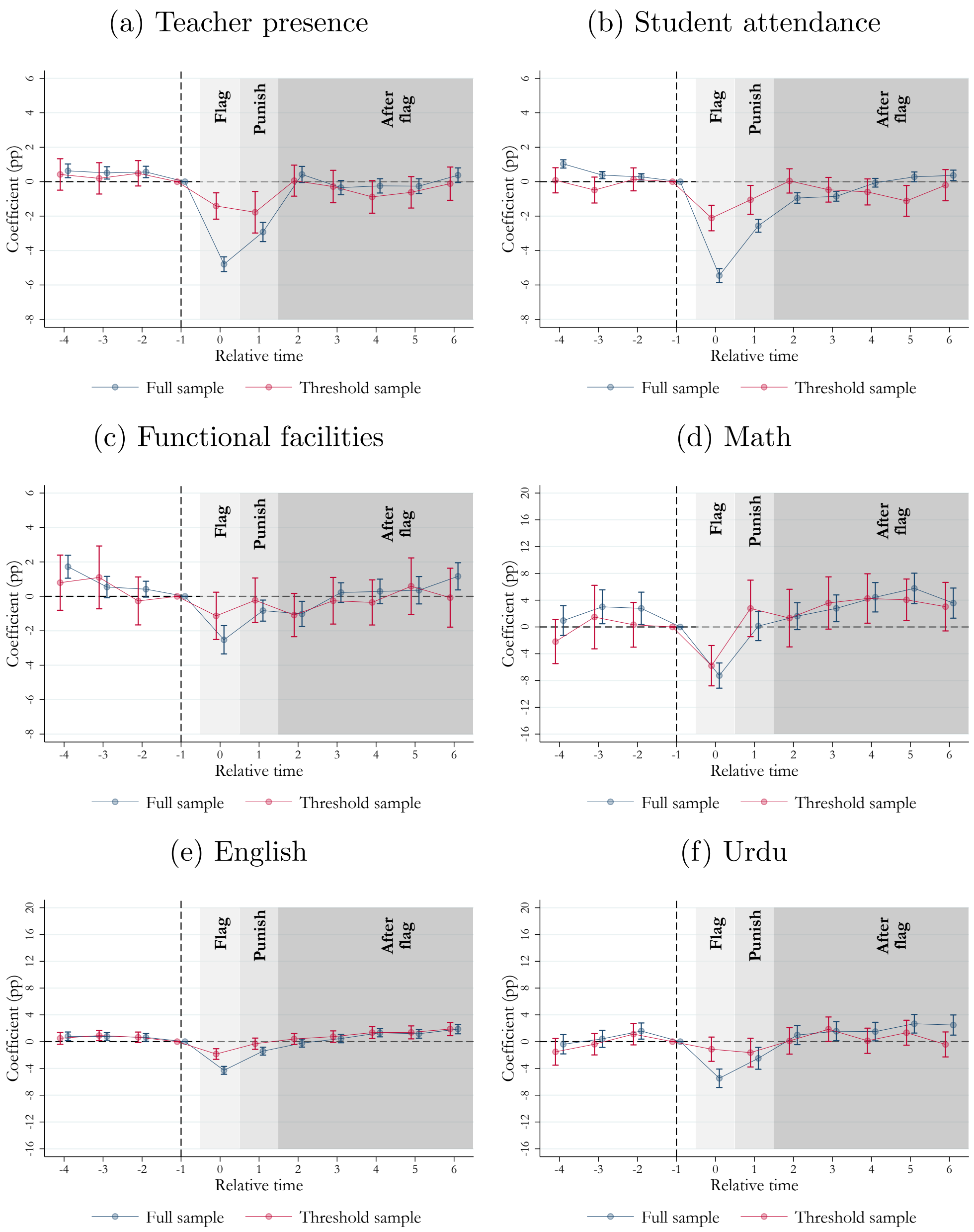
(d) Math

(e) English

(f) Urdu

Note: This figure presents event study graphs based on equation 3.1 using -1 as the base period, comparing schools in orange-flagged and non-flagged maraakiz. The blue and red lines present results for the full and threshold samples, respectively. The results are for flagging on the variable in the title of the panel. *Flag* is for the period in which the information is collected, and the markaz is flagged. *Punish* is for the period where the reports are distributed and the oversight meeting with the punishment occurs. *After flag* is for periods after the oversight meeting occurs. Error bars at the 95 percent level are presented for each coefficient.

Figure C.2.9: Event study - flagging effect on performance tehsil-wing flagging



(a) Teacher presence

(b) Student attendance

(c) Functional facilities

(d) Math

(e) English

(f) Urdu

Note: This figure presents study graphs based on equation 3.1 using -1 as the base period, comparing schools in flagged and non-flagged tehsil-wing. The blue and red lines presents results for the full and threshold samples, respectively. The results are for flagging on the variable in the title of the panel. *Flag* is for the period in which the information is collected, and the markaz is flagged. *Punish* is for the period where the reports are distributed and the oversight meeting with the punishment occurs. *After flag* is for periods after the oversight meeting occurs. Error bars at the 95 percent level are presented for each coefficient.

Table C.2.2: Monitoring effect on performance by datapack - markaz flagging

Panel A: Datapack 1						
Dependent variable:	Teacher presence		Student attendance		Functional facilities	
T×Flag	-5.37***	-1.42***	-6.99***	-2.00***	-4.73***	-1.38***
	(0.15)	(0.22)	(0.21)	(0.19)	(0.38)	(0.31)
T×Punish	-2.06***	-1.31***	-3.32***	-2.74***	-1.13***	-0.71**
	(0.15)	(0.26)	(0.22)	(0.36)	(0.19)	(0.32)
T×After flag	-0.69***	-0.55***	-1.19***	-1.62***	-0.54***	-0.37
	(0.12)	(0.20)	(0.10)	(0.21)	(0.18)	(0.32)
N. of obs.	4,960,055	383,685	2,848,511	444,614	4,852,832	350,862
Mean Dep. Var. before	92.3	87.2	91.4	87.5	96.8	94.1
R^2	0.024	0.028	0.10	0.092	0.063	0.072
Panel B: Datapack 2 (After December 2015)						
Dependent variable:	Teacher presence		Student attendance		Functional facilities	
T×Flag	-6.81***	-2.09***	-7.45***	-2.87***	-3.94***	-0.97
	(0.35)	(0.62)	(0.30)	(0.37)	(0.46)	(0.73)
T×Punish	-1.95***	-1.24*	-0.81***	-0.53	-1.72***	0.11
	(0.50)	(0.73)	(0.23)	(0.35)	(0.65)	(0.81)
T×After flag	-0.11	-0.56	-0.27	-0.073	-2.22**	-0.96
	(0.58)	(1.07)	(0.21)	(0.26)	(1.11)	(1.29)
N. of obs.	971,860	39,343	929,866	66,609	1,138,957	23,832
Mean Dep. Var. before	94.5	87.9	91.8	85.9	98.4	91.9
R^2	0.037	0.046	0.095	0.10	0.068	0.048
Panel C: Datapack 3 (After January 2017)						
Dependent variable:	Teacher presence		Student attendance		Functional facilities	
T×Flag	-11.2***	-3.84***	-8.00***	-2.38***	-5.74***	-1.67
	(0.41)	(0.42)	(0.42)	(0.50)	(0.62)	(1.35)
T×Punish	-5.50***	-3.46***	-2.02***	-0.90	-2.26***	-1.89
	(0.68)	(1.10)	(0.35)	(0.60)	(0.67)	(1.24)
T×After flag	0.57**	-0.082	-0.65***	-0.53*	-0.81*	-1.29
	(0.26)	(0.34)	(0.21)	(0.32)	(0.49)	(0.91)
N. of obs.	1,047,640	67,922	1,186,456	51,438	1,322,816	17,358
Mean Dep. Var. before	94.2	88.0	93.1	85.9	98.7	92.5
R^2	0.074	0.076	0.11	0.14	0.065	0.054
Sample	Full	Threshold	Full	Threshold	Full	Threshold
Markaz FE	Yes	Yes	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes	Yes	Yes
District time trends	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Results for school outcomes from estimating equation 3.1. The first column for each outcome shows estimates for the full sample, and the second column shows estimates for the threshold sample using optimal bandwidth obtained through regression discontinuity optimization methods. Teacher presence (student attendance) are measured as the percentage of present teachers (students) relative to the total teachers (students) reported. Functional facilities represents the percentage of the school's functional infrastructure, including toilets, drinking water, boundary wall, and electricity. *T* equals 1 for schools in a flagged markaz. *Flag* equals 1 for the period in which the information is collected, and the markaz is flagged. *Punish* equals 1 for the period where the reports are distributed and the oversight meeting with the punishment occurs. *After flag* is equal to 1 for periods after the oversight meeting occurs. *Mean. Dep. Var before* shows the average outcome in the non-flagged maraakiz before the flagging occurs. Standard errors clustered by markaz, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

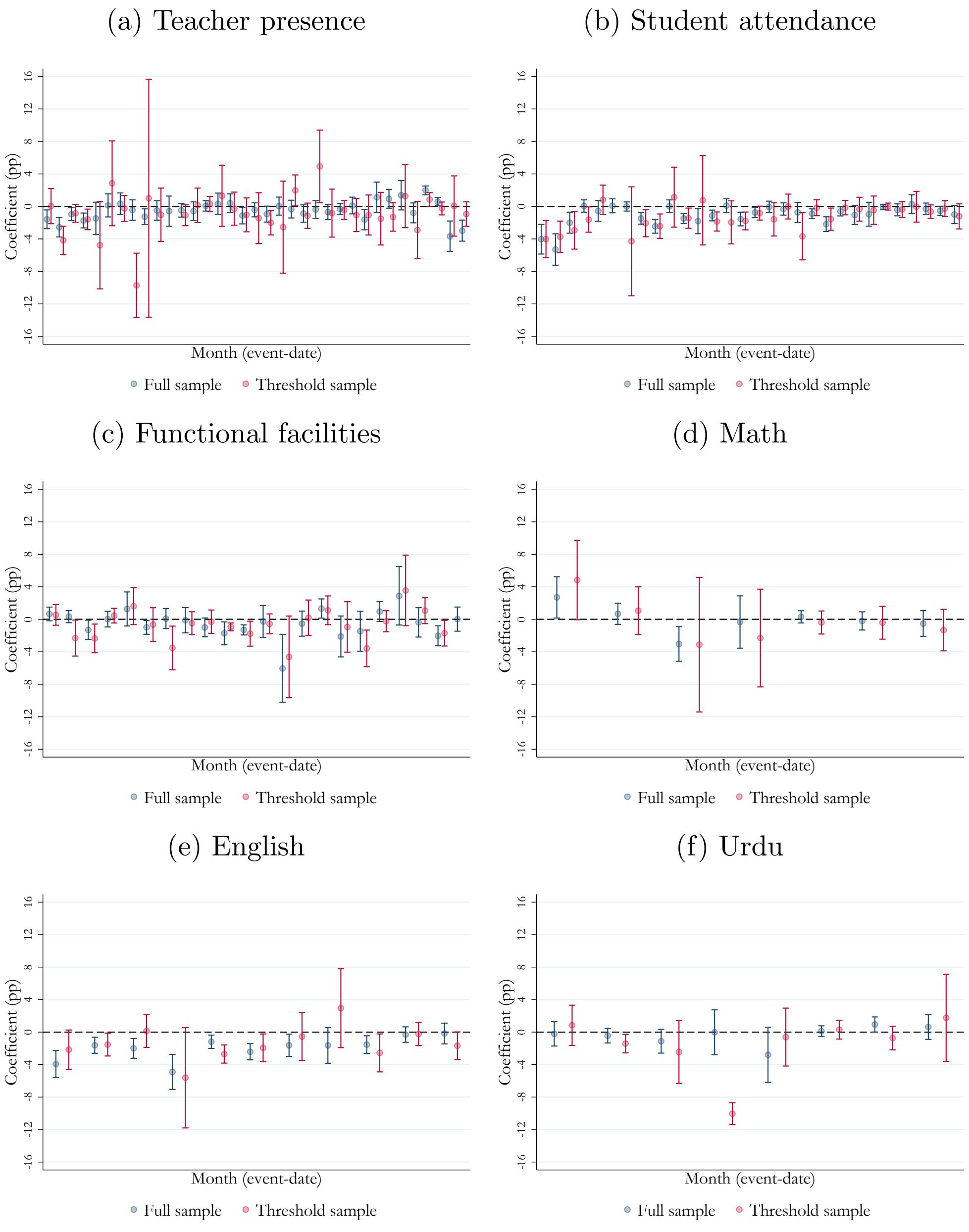
Table C.2.3: Monitoring effect on performance - district ranking and markaz flagging

Panel A: Bottom districts						
Dependent variable:	Teacher presence		Student attendance		Functional facilities	
Bottom×Flag×Meeting	-0.13	-0.61	-0.062	0.67	-0.47	-0.55
	(0.49)	(0.55)	(1.29)	(1.28)	(0.72)	(0.83)
Bottom×Meeting	0.56**	0.60**	1.01	1.13*	0.28	0.32
	(0.26)	(0.29)	(0.88)	(0.62)	(0.40)	(0.55)
Flag×Meeting	-4.92***	-4.14***	-4.92***	-5.50***	-0.85	-0.53
	(0.16)	(0.43)	(0.62)	(0.98)	(0.59)	(0.52)
Bottom×Flag×After meeting	0.79	0.31	-1.33	-0.48	-0.41	-0.051
	(0.54)	(0.65)	(1.00)	(1.09)	(0.68)	(0.71)
Bottom×After meeting	0.24	0.20	1.14*	1.49**	0.56	0.27
	(0.30)	(0.37)	(0.58)	(0.55)	(0.54)	(0.50)
Flag×After meeting	-0.39*	0.099	0.78***	0.29	1.25***	1.05**
	(0.20)	(0.48)	(0.20)	(0.59)	(0.26)	(0.47)
N. of obs.	3,063,835	583,417	3,063,410	583,248	3,009,844	565,920
Mean Dep. Var. before	91.4	90.1	88.8	86.0	92.5	90.0
R^2	0.028	0.033	0.13	0.16	0.17	0.20
Panel B: Top districts						
Dependent variable:	Teacher presence		Student attendance		Functional facilities	
Top×Flag×Meeting	0.47	0.29	-0.36	-3.17	1.25	3.18
	(0.53)	(0.91)	(1.57)	(2.02)	(0.91)	(3.22)
Top×Meeting	0.48	0.33	-1.28**	-0.64	-0.0098	-0.49
	(0.30)	(0.30)	(0.56)	(0.60)	(0.33)	(0.85)
Flag×Meeting	-4.79***	-5.22***	-5.27***	-2.96***	-0.76	-3.21
	(0.23)	(0.66)	(0.57)	(1.02)	(0.56)	(3.31)
Top×Flag×After meeting	-0.57	-1.48	1.17	-0.92	1.06	0.79
	(0.55)	(1.06)	(0.88)	(1.07)	(1.28)	(0.90)
Top×After meeting	0.81***	0.93***	-0.15	-0.30	-0.10	0.49
	(0.24)	(0.21)	(0.43)	(0.61)	(0.20)	(0.39)
Flag×After meeting	-0.049	0.49	0.37*	1.58**	1.29***	1.20**
	(0.15)	(0.67)	(0.18)	(0.71)	(0.26)	(0.54)
N. of obs.	3,111,642	682,461	3,111,048	682,369	3,036,557	672,780
Mean Dep. Var. before	91.0	91.9	87.4	90.0	91.6	92.7
R^2	0.029	0.028	0.13	0.13	0.17	0.18
Sample	Full	Threshold	Full	Threshold	Full	Threshold
District FE	Yes	Yes	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Results from estimating a modified version of equation 3.2, including the flagging status of markaz in the quarterly meeting as a third interactions term. The school is the unit of observation for both panels. The first column for each outcome estimates for the full sample. The second column for each outcome estimates for the threshold sample, including the schools in the five districts closer to the five in the bottom/top. The bottom/top status and threshold sample are based on the aggregate district performance. Teacher presence and student attendance are measured as the percentage of present teachers/students relative to the total teachers/students reported. The functional facilities variable is measured as the percentage of the school's functional infrastructure, including toilets, drinking water, boundary wall, and electricity. *Bottom* equals 1 for schools in the bottom five districts and Top equals 1 for the schools in the top five districts on the date of the quarterly meeting. *Meeting* equals 1 in the period of the quarterly meeting. *Mean. Dep. Var before* shows the average outcome in the non-top/bottom districts before the meeting occurs. Standard errors clustered by district, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

C.2.4 Other robustness checks

Figure C.2.10: Seasonality - monthly effects of flagging



(a) Teacher presence

(b) Student attendance

(c) Functional facilities

(d) Math

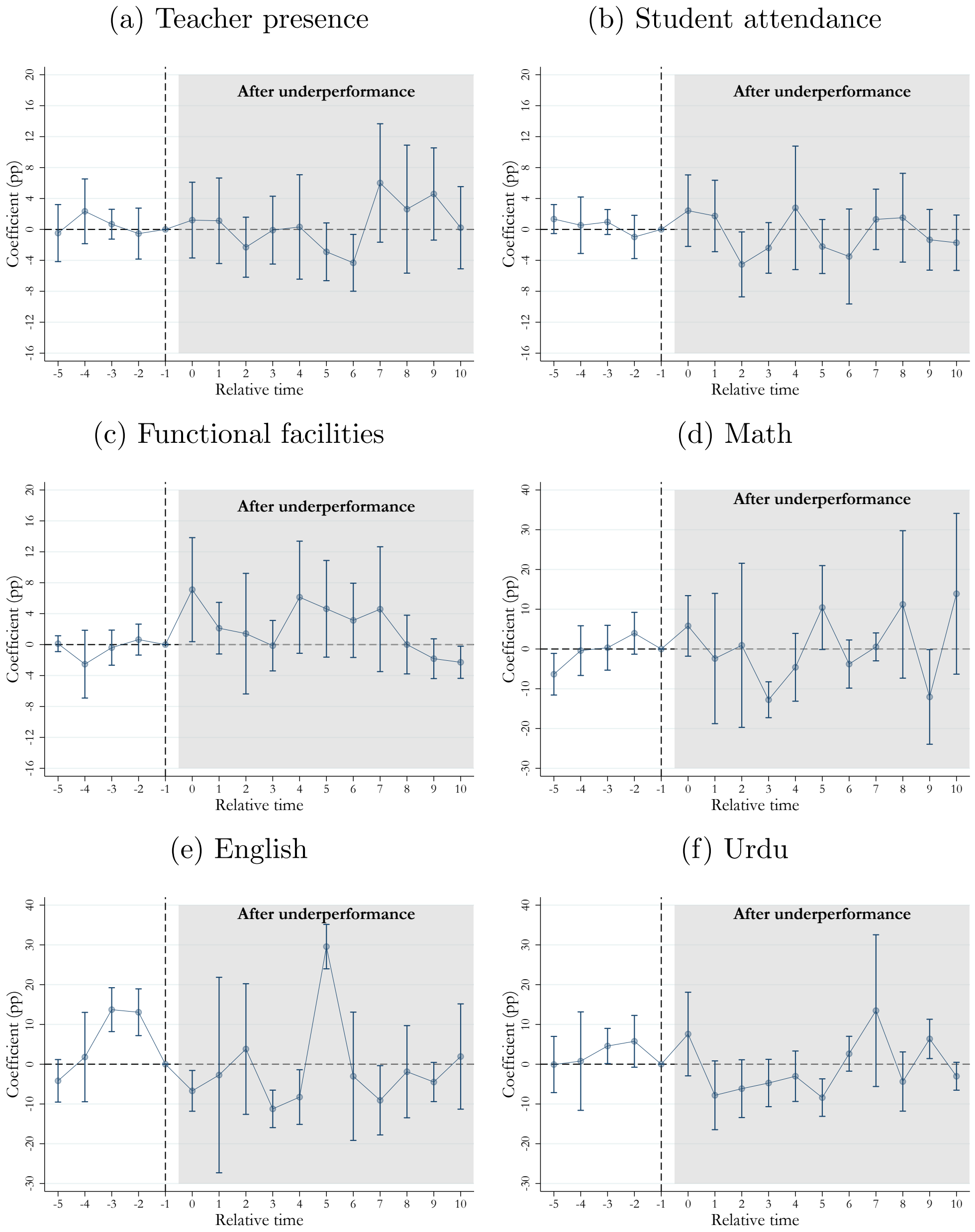
(e) English

(f) Urdu

Note: This figure presents results from the $AfterFlag$ coefficient by estimating equation 3.1 for each individual stack (event panel), comparing schools in flagged and non-flagged maraakiz in that particular event. The blue and red lines present results for the full and threshold samples, respectively. The results are for flagging on the variable in the title of the panel. Error bars represent 95% confidence intervals.

We test additional mechanisms by which results might be confounded. We estimated equation 3.1 for each specific event to test the robustness of the results to time shocks. Figure C.2.10 reports the coefficients for the $Afterflag$ period. In most of the event panels there appear to be non-significant results, which supports the evidence that on average the centralized monitoring scheme has not improved schools' performance. We tested the reversion to the mean hypothesis by identifying if there existed anticipation of the flagging. The premise follows the assumption that a markaz might start recovering before receiving the flagging if the person in charge knows they might be flagged at the end of the month. We estimated a daily event study where treatment starts once the average outcome of the visited schools on a particular day fell below the flagging threshold. In such a case, we assumed that the public officer might identify the potential flagging and react in the days afterward. The results in Figure C.2.11 suggests no reaction exists in response to being below the threshold for the first time in the month. Finally, we plotted the after-flag β coefficients by accumulating one month at a time an approach to estimate the effect of flagging conditional on the information that the public officer had available for each period. Figure C.2.12 plots the coefficients for each flagging variable. We note that the effect converges towards a null result all cases, and it was possible to identify negative or null effects since the very start of the scheme.

Figure C.2.11: Flagging anticipation


(a) Teacher presence

(b) Student attendance

(c) Functional facilities

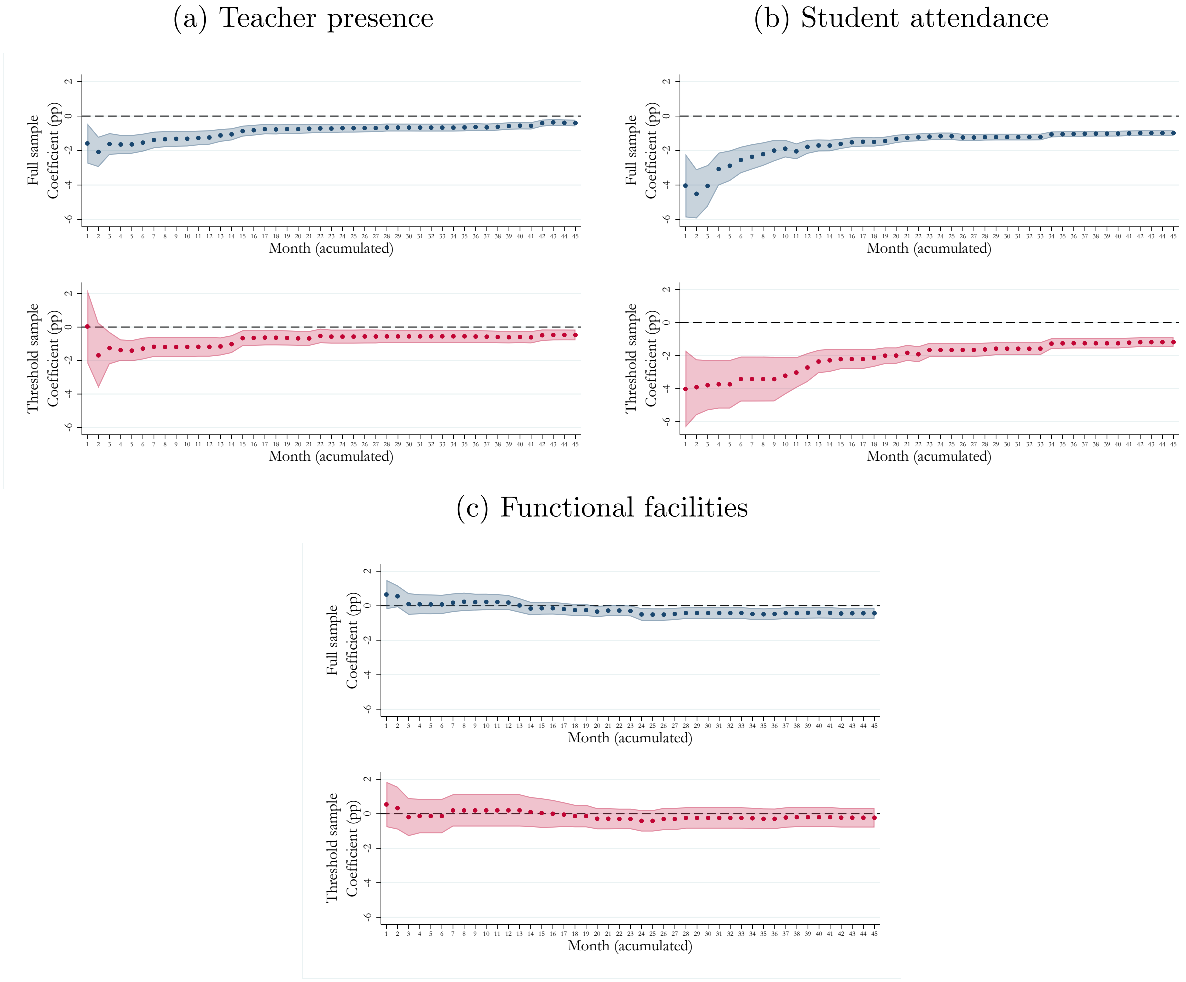
(d) Math

(e) English

(f) Urdu

Note: This figure presents results from estimating an event study for the daily average of the outcomes in the month of flagging, comparing maraakiz whose average of visited schools was below the threshold at some point in the month, against maraakiz that never underperformed. The base period consists of the day just before the average of the visited schools until that day lies below the flagging threshold. Thus, *After underperformance* is for the periods after the maraakiz was underperforming on average for the first time in the month of data collection (equivalent to *Flag* period in the main specification). Error bars at the 95% level are presented for each coefficient.

Figure C.2.12: Accumulated flagging effects by month
after flag accumulated effect



(a) Teacher presence

(b) Student attendance

(c) Functional facilities

Note: This figure presents results from the $AfterFlag$ coefficient by estimating equation 3.1 , accumulating one stack (event panel) at a time, comparing schools in flagged and non-flagged maraakiz. The specification accumulates all the months up until t. The blue coefficients presents results for the full sample, while the red coefficients presents results for the threshold sample, obtained through regression discontinuity optimization methods. The results are for flagging on the variable in the title of the panel. Error bars at the 95 percent level are presented for each coefficient.

C.2.5 Assessing the impacts of the scheme across distinct political environments

We assess whether flagging was different in places aligned with the ruling party, considering the pressure set on the scheme by Punjab Chief Minister's participation. We use Provincial

Assembly elections data for 2013 to define political alignment as in Callen, Gulzar and Rezaee (2020): an area is politically aligned if the winner of the constituency seat is from the same party as the Chief Minister. We match schools to electoral constituencies to define: i) maraakiz fully aligned: all winners have the same party as the chief minister, ii) not fully aligned, and iii) not aligned: no constituency with the same party as the chief minister. We estimate the following difference-in-differences specification:

$$\begin{gathered} Y_{mt} = \beta_1(FullyAligned_m \times AfterElection) + \\ \beta_2(NotFullyAligned_m \times AfterElection) + \alpha_m + \lambda_t + dt + \epsilon_{mt} \end{gathered} \tag{C.2}$$

α_m, λ_t are for markaz and time fixed effects, and dt is for district time trends. We limited the analysis to nine months before/after the elections. β_1 β_2 capture the effects of being politically aligned versus not before versus after the election. We also estimate the effect in the sample with high electoral competition, defined by close elections using optimal bandwidth procedures (Calonico, Cattaneo and Farrell (2020)).

Panel A in Table C.2.4 reports the effect on the probability of flagging. Panel B of Table C.2.4 reports the effect on the education outcomes. We do not observe any consistent results across outcomes: there are some effects on student attendance but they do not exist in the close elections sample for the probability of flagging, are very small (under 2 percentage points) for actual student attendance. Thus, we conclude that program implementation was not strongly connected to the ruling party differently from opposition-governed places.

Table C.2.4: Political alignment effect on probability of being flagged

Panel A - Dep. var: Flagging (=1):	Teacher presence		Student attendance		Functional facilities	
Fully aligned×After elections	0.015	0.015	-0.083***	-0.032	-0.014	-0.039
	(0.024)	(0.029)	(0.026)	(0.043)	(0.025)	(0.029)
Not fully aligned ×After elections	0.0024	-0.011	-0.072***	-0.050	-0.014	-0.0077
	(0.024)	(0.029)	(0.026)	(0.037)	(0.024)	(0.027)
N. of obs.	19,143	10,838	18,908	7,804	18,889	10,027
Mean Dep. Var. before	0.086	0.093	0.36	0.35	0.31	0.37
R^2	0.31	0.33	0.37	0.39	0.70	0.72
Panel B - Dep. var: Outcomes	Teacher presence		Student attendance		Functional facilities	
Fully aligned×After elections	-0.29	0.048	1.95***	1.66***	0.12	-0.16
	(0.44)	(0.56)	(0.33)	(0.43)	(0.52)	(0.69)
Not fully aligned×After elections	0.061	0.078	1.43***	1.39***	0.038	-0.67
	(0.44)	(0.58)	(0.32)	(0.38)	(0.57)	(0.77)
N. of obs.	19,141	8,913	19,141	10,277	19,132	9,744
Mean Dep. Var. before	92.0	91.9	86.2	86.0	92.8	92.0
R^2	0.30	0.33	0.54	0.57	0.73	0.77
Sample	Full	Close elections	Full	Close elections	Full	Close elections
Markaz FE	Yes	Yes	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes	Yes	Yes
District time trends	Yes	Yes	Yes	Yes	Yes	Yes

Note: Results from estimating equation C.2. Fully (not fully) aligned equals 1 for maraakiz where all (not all) the schools were in a aligned constituency. The control group are maraakiz not aligned. After elections equals 1 for months after May/2013 (elections date). Close elections sample is for the maraakiz with competitive elections, defined as the bandwidth obtained through RD optimization methods around the difference in the vote share between the party aligned and the party not aligned. Panel A report the results on the probability of being flagged for each outcome. Panel B report the results of being politically aligned on the value of the outcomes. Standard errors, clustered by markaz, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

C.2.6 Impacts on the machinery of the government

Table C.2.5: Monitoring effect on other outcomes - effort as mechanism

Panel A: School outcomes flagging						
Flagging variable:	Teacher presence		Student attendance		Functional facilities	
T×Flag	-0.0037	-0.012	-0.0031	-0.0036	-0.0085	-0.028*
	(0.0049)	(0.0092)	(0.0027)	(0.0052)	(0.0084)	(0.015)
T×Punish	-0.0013	-0.017*	-0.0012	-0.0056	0.013	-0.026
	(0.0058)	(0.0093)	(0.0036)	(0.0073)	(0.0088)	(0.016)
T×After flag	0.00034	-0.015**	0.0052*	0.0059	0.0054	-0.020
	(0.0042)	(0.0073)	(0.0032)	(0.0047)	(0.0062)	(0.013)
N. of obs.	6,208,175	436,428	4,400,630	501,589	6,588,404	356,783
Mean Dep. Var. before	0.97	0.92	0.97	0.95	0.97	0.95
R^2	0.16	0.26	0.20	0.24	0.18	0.25
Panel B: School scores flagging						
Flagging variable:	Math		English		Urdu	
T×Flag	0.011**	-0.0052	0.0067**	0.0040	0.0076**	0.0039
	(0.0044)	(0.0083)	(0.0029)	(0.0056)	(0.0037)	(0.0054)
T×Punish	0.021**	0.017	0.0064*	-0.00068	0.017**	0.0066
	(0.0098)	(0.020)	(0.0035)	(0.0061)	(0.0074)	(0.011)
T×After flag	0.00039	0.0098	0.000072	0.0016	-0.0039	-0.0043
	(0.0036)	(0.0073)	(0.0027)	(0.0044)	(0.0032)	(0.0049)
N. of obs.	1,922,793	45,750	706,106	128,094	1,705,174	103,434
Mean Dep. Var. before	0.98	0.97	0.98	0.97	0.98	0.97
R^2	0.24	0.25	0.23	0.27	0.23	0.30
Sample	Full	Threshold	Full	Threshold	Full	Threshold
Markaz FE	Yes	Yes	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes	Yes	Yes
District time trends	Yes	Yes	Yes	Yes	Yes	Yes

Note: Results from estimating equation 3.1. The dependent variable equals 1 if the schools received a visit by an AEO. The school is the unit of observation for both panels. The first column for each outcome estimates for the full sample. The second column for each outcome estimates for the threshold sample, including schools in maraakiz that lie within the bandwidth obtained through regression discontinuity optimization methods. Teacher presence and student attendance are measured as the percentage of present teachers/students relative to the total teachers/students reported. The functional facilities variable is measured as the percentage of the school's functional infrastructure, including toilets, drinking water, boundary wall, and electricity. Math, English, and Urdu scores are measured as the percentage of correct answers in standardized tests. T equals 1 for schools in a flagged markaz. *Flag* equals 1 for the period in which the information is collected, and the markaz is flagged. *Punish* equals 1 for the period where the reports are distributed and the oversight meeting with the punishment occurs. *After flag* is equal to 1 for periods after the oversight meeting occurs. *Mean. Dep. Var before* shows the average outcome in the non-flagged maraakiz before the flagging occurs. Standard errors clustered by markaz, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table C.2.6: Monitoring effect on other outcomes - gaming in bureaucratic visits

Panel A: Bigger schools						
Flagging variable:	Teacher presence		Student attendance		Functional facilities	
T×Flag	-0.0052	-0.012	-0.0030	-0.0035	-0.0098	-0.023
	(0.0049)	(0.0093)	(0.0031)	(0.0060)	(0.0090)	(0.015)
T×Punish	-0.0015	-0.018*	-0.00024	-0.0029	0.014	-0.019
	(0.0058)	(0.0096)	(0.0038)	(0.0078)	(0.0096)	(0.016)
T×After flag	0.00066	-0.014*	0.0071**	0.0070	0.0043	-0.014
	(0.0042)	(0.0076)	(0.0032)	(0.0046)	(0.0064)	(0.013)
N. of obs.	3,296,879	239,246	2,339,012	267,953	3,480,301	193,888
Mean Dep. Var. before	0.97	0.92	0.97	0.95	0.97	0.95
R^2	0.19	0.26	0.23	0.24	0.20	0.25
Panel B: Worst performing schools						
Flagging variable:	Teacher presence		Student attendance		Functional facilities	
T×Flag	-0.00058	-0.012	-0.0046	-0.0063	-0.019	-0.044**
	(0.0065)	(0.012)	(0.0031)	(0.0059)	(0.013)	(0.020)
T×Punish	0.0011	-0.011	-0.00041	-0.0011	0.0051	-0.042*
	(0.0077)	(0.012)	(0.0040)	(0.0080)	(0.014)	(0.024)
T×After flag	0.00063	-0.015*	0.0031	0.0034	-0.0025	-0.033*
	(0.0053)	(0.0091)	(0.0035)	(0.0057)	(0.0092)	(0.018)
N. of obs.	1,419,103	111,744	1,957,278	224,643	546,530	54,055
Mean Dep. Var. before	0.97	0.92	0.97	0.95	0.96	0.94
R^2	0.19	0.28	0.21	0.24	0.21	0.30
Panel C: Schools with most missing teachers						
Flagging variable:	Teacher presence		Student attendance		Functional facilities	
T×Flag	-0.00058	-0.012	-0.0041	-0.0098	-0.024**	-0.049**
	(0.0065)	(0.012)	(0.0037)	(0.0073)	(0.011)	(0.020)
T×Punish	0.0011	-0.011	0.00095	-0.0010	0.0081	-0.053*
	(0.0077)	(0.012)	(0.0048)	(0.010)	(0.015)	(0.028)
T×After flag	0.00063	-0.015*	0.0059	0.0036	0.0010	-0.030
	(0.0053)	(0.0091)	(0.0041)	(0.0064)	(0.0085)	(0.021)
N. of obs.	1,419,103	111,744	975,682	126,589	1,707,459	98,760
Mean Dep. Var. before	0.97	0.92	0.97	0.95	0.96	0.94
R^2	0.19	0.28	0.25	0.25	0.21	0.28
Sample	Full	Threshold	Full	Threshold	Full	Threshold
Markaz FE	Yes	Yes	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes	Yes	Yes
District time trends	Yes	Yes	Yes	Yes	Yes	Yes

Note: Results from estimating equation 3.1 on school outcomes flagging. The dependent variable equals 1 if the schools received a visit by an AEO. Panels A, B and C show estimates for above median-sized schools, below-median performance, and above-median teacher absenteeism in the maraakiz, respectively. The first column for each outcome shows estimates for the full sample, and the second for the threshold sample. Teacher presence (student attendance) are measured as the percentage of present teachers (students) relative to the total teachers (students) reported. Functional facilities represents the percentage of the school's functional infrastructure, including toilets, drinking water, boundary wall, and electricity. *T* equals 1 for schools in a flagged markaz. *Flag* equals 1 for the period in which the information is collected, and the markaz is flagged. *Punish* equals 1 for the period where the reports are distributed and the oversight meeting with the punishment occurs. *After flag* is equal to 1 for periods after the oversight meeting occurs. *Mean. Dep. Var before* shows the average outcome in the non-flagged maraakiz before the flagging occurs. Standard errors clustered by markaz, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table C.2.7: Monitoring effect on yearly school budget

Panel A: OLS estimates				
Dependent variable (in logs):	Total funds	Government funds	Non government funds	Total expenses
Num times flagged (teacher presence)	-0.0057	-0.0030	0.064***	-0.034
	(0.031)	(0.035)	(0.023)	(0.021)
N. of obs.	162,753	162,753	162,753	162,752
Mean Dep. Var. before	51205.8	46251.0	4954.8	66619.1
R^2	0.63	0.63	0.37	0.17
Num times flagged (student attendance)	0.030	0.071**	-0.021	0.025*
	(0.026)	(0.028)	(0.018)	(0.015)
N. of obs.	162,753	162,753	162,753	162,752
Mean Dep. Var. before	51205.8	46251.0	4954.8	66619.1
R^2	0.63	0.63	0.37	0.17
Num times flagged (functional facilities)	-0.0095	-0.012	0.029**	-0.0086
	(0.017)	(0.020)	(0.013)	(0.013)
N. of obs.	162,753	162,753	162,753	162,752
Mean Dep. Var. before	51205.8	46251.0	4954.8	66619.1
R^2	0.63	0.63	0.37	0.17
Panel B: IV estimates				
Dependent variable (in logs):	Total funds	Government funds	Non government funds	Total expenses
Num times flagged (teacher presence)	-0.10	-0.17	0.096	-0.12
	(0.17)	(0.19)	(0.11)	(0.094)
N. of obs.	162,753	162,753	162,753	162,752
F-stat	136.0	136.0	136.0	136.1
Mean Dep. Var. before	51205.8	46251.0	4954.8	66619.1
Num times flagged (student attendance)	-0.37	-0.46	0.082	0.053
	(0.55)	(0.64)	(0.33)	(0.28)
N. of obs.	162,753	162,753	162,753	162,752
F-stat	13.4	13.3	13.4	13.4
Mean Dep. Var. before	51205.8	46251.0	4954.8	66619.1
Num times flagged (functional facilities)	-0.94	-0.84	-0.63	0.39
	(1.04)	(1.08)	(0.69)	(0.51)
N. of obs.	162,753	162,753	162,753	162,752
F-stat	2.99	2.99	2.99	2.99
Mean Dep. Var. before	51205.8	46251.0	4954.8	66619.1
Markaz FE	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes
District time trends	Yes	Yes	Yes	Yes

Notes: Results from linear regression of log-transformed frequency of flagging by each variable in previous fiscal year, on the current fiscal year budget distribution. Total funds = Government funds + Non-government funds. Regressions include markaz and year fixed effects, and district time-trends. Panel B reports the second stage from an IV regression instrumenting the flagging frequency by the distance to the flagging threshold, akin to a fuzzy RD setup. The F-statistic comes from the first stage of the regression to test for weak instruments. Standard errors are clustered at the markaz level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table C.2.8: Monitoring effect on labor market - change of head teachers

Panel A: School outcomes flagging						
Flagging variable:	Teacher presence		Student attendance		Functional facilities	
T×Flag	0.0047*	0.0022	-0.0050**	-0.00011	0.0059	0.0098
	(0.0025)	(0.0042)	(0.0024)	(0.0050)	(0.0057)	(0.0091)
T×Punish	-0.00099	0.00059	-0.0039	0.0037	0.0013	0.0088
	(0.0026)	(0.0051)	(0.0025)	(0.0051)	(0.0051)	(0.0084)
T×After flag	-0.0015	-0.0016	-0.0038*	0.0022	0.0040	0.0070
	(0.0023)	(0.0038)	(0.0020)	(0.0037)	(0.0030)	(0.0058)
N. of obs.	6,979,870	490,971	4,965,559	562,830	7,409,613	398,961
Mean Dep. Var. before	0.060	0.058	0.059	0.056	0.057	0.069
R^2	0.094	0.10	0.099	0.097	0.084	0.080
Panel B: School scores flagging						
Flagging variable:	Math		English		Urdu	
T×Flag	-0.013	-0.028	-0.0050	-0.00022	0.00066	0.0057
	(0.0085)	(0.017)	(0.0051)	(0.010)	(0.0064)	(0.0091)
T×Punish	-0.025***	-0.016	-0.015***	-0.016	-0.019***	-0.017*
	(0.0073)	(0.013)	(0.0048)	(0.011)	(0.0050)	(0.0087)
T×After flag	-0.014***	-0.024***	-0.0078**	-0.013***	-0.0099**	-0.0032
	(0.0050)	(0.0091)	(0.0033)	(0.0049)	(0.0041)	(0.0060)
N. of obs.	2,198,503	53,361	810,829	147,630	1,950,809	119,592
Mean Dep. Var. before	0.061	0.093	0.071	0.090	0.061	0.063
R^2	0.13	0.13	0.13	0.15	0.13	0.16
Sample	Full	Threshold	Full	Threshold	Full	Threshold
Markaz FE	Yes	Yes	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes	Yes	Yes
District time trends	Yes	Yes	Yes	Yes	Yes	Yes

Note: Results from estimating equation 3.1. The dependent variable equals 1 if the head teacher is different from the one reported in $t-1$. The school is the unit of observation for both panels. The first column for each outcome estimates for the full sample. The second column for each outcome estimates for the threshold sample, including schools in maraakiz that lie within the bandwidth obtained through regression discontinuity optimization methods. Teacher presence and student attendance are measured as the percentage of present teachers/students relative to the total teachers/students reported. The functional facilities variable is measured as the percentage of the school's functional infrastructure, including toilets, drinking water, boundary wall, and electricity. Math, English, and Urdu scores are measured as the percentage of correct answers in standardized tests. *T* equals 1 for schools in a flagged markaz. *Flag* equals 1 for the period in which the information is collected, and the markaz is flagged. *Punish* equals 1 for the period where the reports are distributed and the oversight meeting with the punishment occurs. *After flag* is equal to 1 for periods after the oversight meeting occurs. *Mean. Dep. Var before* shows the average outcome in the non-flagged maraakiz before the flagging occurs. Standard errors clustered by markaz, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table C.2.9: Monitoring effect on labor markets - change of district officers

Dependent variable:	Change of DC			
	(1)	(2)	(3)	(4)
Bottom×Meeting	0.038	0.048		
	(0.048)	(0.062)		
Bottom×After meeting	-0.0075	0.0073		
	(0.026)	(0.037)		
Top×Meeting			0.037	0.075
			(0.058)	(0.051)
Top×After meeting			-0.0052	0.029
			(0.026)	(0.037)
N. of obs.	2,921	605	3,025	685
Mean Dep. Var. before	0.060	0.066	0.058	0.047
R^2	0.15	0.24	0.14	0.31
Flagging	Bottom	Bottom	Top	Top
Sample	Full	Threshold	Full	Threshold
District FE	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes

Notes: Results from estimating equation 3.2. The district is the unit of observation for both panels. The dependent variable equals 1 if the district commissioner is different from the one reported in $t-1$ The first column for each outcome estimates for the full sample. The second column for each outcome estimates for the threshold sample, including the five districts closer to the five in the bottom/top. *Bottom* equals 1 for schools in the bottom five districts and Top equals 1 for the schools in the top five districts on the date of the quarterly meeting. *Meeting* equals 1 in the period of the quarterly meeting. *Mean. Dep. Var before* shows the average outcome in the non-top/bottom districts before the meeting occurs. Standard errors clustered by district, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table C.2.10: Monitoring effect on labor markets - current position of district officers

Dependent variable:	Rank of current employment	
	(1)	(2)
Months in bottom districts	-1.21	
	(1.24)	
Months in top districts		0.39
		(1.14)
N. of obs.	82	81
Mean Dep. Var. before	2.74	2.79
R^2	0.010	0.0011

Notes: Results from estimating a linear regression of the number of months a district officer was ranked in the top/bottom in the quarterly meetings during its time in office on the rank of the current employment. Higher value in the rank account for better career trajectory for district officer. For details on the construction of the rank of employment variable see Appendix C.1.5. The data is aggregated as the district officer level. Bootstrapped standard errors in brackets. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.