

# Essays on Incentives in Bureaucracy: Experimental Evidence from Pakistan



Zahra Mansoor  
Wadham College  
University of Oxford

A thesis submitted for the degree of  
*Doctor of Philosophy in Public Policy*

Trinity 2020



# Acknowledgements

I am deeply grateful to my supervisor Julien Labonne for the constant motivation, enthusiasm, support, and supervision at every step of this thesis. I am also grateful to my second supervisor Stefan Dercon for his instrumental feedback and supervision at several critical junctures that helped shaped the direction of my research.

I would also like to express gratitude to Martin Williams for his advice during the course of this thesis, and to Clare Leaver, Simon Quinn, and Chris Woodruff for their detailed feedback during Transfer and Confirmation of Status.

I would like to acknowledge the Economic Development and Institutions (EDI), Center for the Study of African Economis (CSAE), and the Blavatnik School of Government (BSG) for supporting my research and fieldwork for Chapters 1 and 2. I also gratefully acknowledge funding from the World Bank i2i, World Bank Strategic Research Programme, and the International Growth Center for funding the research presented as Chapter 3.

Special thanks are due to my coauthors and the field research team in Pakistan that worked on Chapter 3. I am also thankful to the field survey firm, Research Consultants (Rcons), for survey implementation support.

This thesis would not have been possible without the support of several government counterparts at the Punjab Teacher Training Academy and the Punjab Agriculture Department. In particular, I am thankful for the numerous discussions and insights that helped improve the quality and policy relevance of my work.

To my Dphil colleagues - for intellectually stimulating conversations (related or unrelated to my thesis), the best support system that I could have hoped for, and the regular bouts of inspiration. You made these four years all the more rewarding. Thank you!

And finally, the biggest thanks to my parents, Khalid Mansoor and Noor-un-Nisa, who were a force behind me throughout this journey. My family and friends back home from Pakistan for their constant encouragement and confidence in my abilities. And last but not the least, my husband and my best friend, Ali Inam, who encouraged me to set out on this path and then supported me at every step imaginable and more.



# Abstract

Public sector performance is a key determinant for a country's economic growth and development. Yet, poor service delivery remains widespread across the developing world which is often attributed to weak monitoring and incentive systems. This thesis, written in a three-paper format, explores questions around how to improve the performance of frontline workers through the use of financial and non-financial incentives. I design two field experiments in collaboration with the Teacher Training Academy in Punjab Pakistan to explore the impact of employer recognition (a type of non-financial incentive) on the performance of public school teachers in in-service trainings. The first experiment serves as a pilot to the second experiment which is larger in scope and builds upon the findings of the pilot. These experiments are presented as Chapter 1 and Chapter 2 respectively. A third experiment is designed and implemented in collaboration with the Punjab Agriculture Department to evaluate the impact of a province-wide pay-for-performance programme on extension outreach and quality. This study is presented as Chapter 3.

In Chapters 1 and 2, I investigate whether non-financial incentives such as employer recognition can improve the performance of public school teachers in in-service trainings as measured through training test scores, and if so, what are the channels through which such tools operate?

In chapter 1 (pilot), I randomize 650 primary and secondary school teachers into three employer recognition schemes and a control group to understand the strength of different channels of employer recognition. The employer recognition incentive is a standard tournament-based incentive in which teachers are awarded a prestigious recognition certificate if they qualify. In T1 (Private arm), the certificate is given to the teacher privately; in T2 (Peer arm), the certificate is given in front of peers and colleagues; and in T3 (Career arm), the certificate is given privately but teachers are also told that their name would be added to an 'excellent teacher list' which will be shared with departmental leadership that can help them qualify for future career opportunities in their districts. Results show that the source of individual motivation of bureaucrats for entering the civil service mediates the effectiveness of employer recognition. We find that teachers who report *intrinsic* reasons for entering the service such as 'interested in the profession of teaching' exhibit a positive treatment response and those who report *extrinsic* reasons such as 'salary' show a negative treatment effect, with the two effects being significantly

different from each other. Further analysis shows that the positive treatment effect for intrinsic individuals comes from the “peer” and “career” channels of employer recognition, with no effects on the “private” channel. Treatment effects on endline self-efficacy beliefs show positive effects for intrinsic but negative effects for extrinsic individuals, indicating that self-efficacy beliefs may be an important causal mechanism for these effects. These results directly support the design of the study in Chapter 2.

In Chapter 2, I randomize 3,394 head teachers attending a mandatory government training into four different designs of recognition incentives that are tied to training performance and a control group. Treatment 1 (Peer arm) leverages the peer/collegial approval channel of recognition – trainees are told that those with the top score in the training post-test and the most improved score (over the pre-test) will be provided certificates in a district level ceremony which will be attended by their peers and colleagues in their district; Treatment 2 (Career arm) leverages the career-benefits channel of recognition – it is the same as Treatment 1 except trainees are told that those who qualify will receive the certificates privately but at the same time their name will be added to an ‘excellent teacher list’ (which will be shared with the School Education Department leadership) that can help them qualify for future opportunities in their districts. This leverages informal career incentives in the system such as postings to preferred schools, transfers to other lateral postings, or getting selected for promotion earlier once eligible; Treatments 3 and 4 (Public PLUS and Career PLUS) cross the first two treatments with a self-efficacy enhancing frame that aims to bolster teacher perceptions in their ability to do well in the training and their jobs more broadly. I find that employer recognition can improve teacher training performance if it is linked to tangible career benefits in the future but the positive effects can also backfire depending on how these incentives are framed. When a self-efficacy frame is added to either of the first two treatments, it improves teacher self-efficacy but makes teachers overconfident, reducing performance in the training. These findings have implications for how to design more powerful non-financial incentives that are effective for eliciting higher teacher effort, specifically in trainings and more broadly in schools and classrooms. At the same time, they point towards the sensitivity of such incentives to framing effects and suggest caution in how these are framed.

In Chapter 3, we partner with the Agriculture Extension Department to evaluate the impact of three alternative pay-for-performance schemes on extension outreach and quality. Extension service delivery is a typical setting where agents carry out tasks that include easily measurable and hard to measure dimensions, and hence the merits and demerits of objective versus subjective incentive systems are unclear. The performance schemes leverage the departmental digital monitoring system called AgriSmart and link incentives to objective metrics available on AgriSmart (Objective arm), supervisors’ own subjective evaluation (Subjective arm), or supervisors’ own subjective evaluation with an additional

element of top-down monitoring to align supervisors' incentives with the objectives of the principal (Subjective Plus arm). Early results show that while all treatments improve performance on AgriSmart metrics, Subjective Plus also shows positive and significant effects on farmer experience of extension outreach and quality, and farmer agricultural practices. In addition, we also find strong and positive effects on AD performance in their managerial tasks (such as assigning farmer training programmes to frontline extension agents) in the Subjective Plus arm, which form an important input into how frontline staff carry out extension outreach. Additional data collection will explore whether the results in Subjective Plus are driven by the effective use of discretion by supervisors in their evaluations, an improvement in AD performance and practices, or both. It will also explore mechanisms that explain why treatment effects observed on AgriSmart outreach indicators in the Objective and Subjective arm do not translate to farmer experience of extension outreach and quality.

This thesis generates three key findings. First, it highlights how to bolster teacher engagement, effort, and subsequently the acquisition of knowledge and skills in in-service trainings. It indicates that non-financial incentives that are linked to potential career benefits in the future are a feasible and cost-effective way for improving the impact of in-service trainings, where some of these career benefits such as postings to preferred schools or transfers to other lateral positions may be less formal and easier to exercise in the system as compared to promotions. More broadly, the results in Chapter 1 and Chapter 2 point towards the existence of such informal career incentives in the system and how they could be leveraged to make the design of non-financial incentives sharper. Second, the impact of non-financial incentives is mediated through individual sources of motivation and non-cognitive traits such as beliefs about ability and individual self-efficacy. This makes such incentives highly contextual, where the impact can often be negative depending on the group-type being incentivized. This implies a need for caution in how such incentives are designed and framed. Third, settings where tasks are multi-dimensional and quality of performance is hard to measure, incentive systems which give discretion to supervisors can be effective as long as supervisors incentives can be aligned with the objectives of the principal.



# Contents

<b>Introduction</b>	<b>1</b>
<b>1 The Double-Edged Sword of Non-Financial Incentives (Part I): Evidence from the Education Sector in Pakistan</b>	<b>17</b>
<b>2 The Double-Edged Sword of Non-Financial Incentives (Part II): Evidence from the Education Sector in Pakistan</b>	<b>77</b>
<b>3 Carrots or Sticks: The Impact of Incentives and Monitoring on the Performance of Public Extension Staff</b>	<b>155</b>
<b>Conclusion</b>	<b>247</b>



# Introduction

**Notes for the reader:** This thesis consists of a brief introduction, followed by three individual chapters, and a conclusion.

Chapter 1 and Chapter 2 are best read sequentially. Chapter 1 is a smaller pilot study with the Punjab Teacher Academy upon which the design of the main study in Chapter 2 builds upon. Chapter 3 is based on a different study with the Punjab Agriculture Extension Department. Each chapter is self-contained with all the relevant tables, figures, and appendices within that chapter.

The first two chapters are entirely my own work.

Chapter 3 is co-authored with Ghazala Mansuri and Garance Genicot. The project was initially set up by Ghazala Mansuri and I. Garance Genicot got involved in the second year of the implementation. I have done the entire regression analysis and writing for Chapter 3 on my own.

# Introduction

Public sector performance is a key determinant for a country's economic growth and development. Yet, developing countries continue to struggle with low public sector performance with sub-standard delivery of basic services. As such, absenteeism and low effort of frontline staff are now widely recognized as significant constraints to effective public service delivery (see Chaudhury et al., 2006). The past decade has seen a growing body of literature that explores underlying mechanisms for understanding the state of performance of frontline bureaucrats. Grounded in the standard principle-agent framework, a significant part of this literature focuses on the role of incentives, particularly financial incentives. More recently selection of bureaucrats, management practices of middle-tier bureaucrats, and intrinsic motivation have also garnered attention as potentially important determinants of performance (see Finan et al., 2015 for a review).

This thesis explores the overarching question of how to improve the performance of *existing* frontline workers through the use of financial and non-financial incentives. It explores this question through direct collaboration with two different public sector departments in Punjab, Pakistan - the Education Department and the Agriculture Department. In the collaboration with the Education Department, I examine the role of non-financial incentives for improving the performance of public school teachers in routine in-service trainings. With the Agriculture Department, I examine the role of financial incentives for improving the performance of frontline agriculture extension workers.

Non-financial incentives could be effective in pro-social settings like education where teachers may put a lower weight on financial incentives (Besley and Ghatak, 2005). Such incentives may also be less susceptible to crowding out of intrinsic motivation (Bénabou and Tirole, 2002) or inducing mistrust (Falk and Kosfeld, 2006). While a considerable empirical literature demonstrates the effectiveness of non-financial incentives - such as status rewards and recognition - in the private sector (see Frey and Neckermann, 2008; Markham et al., 2002; Luthans and Stajkovic, 1999; Kosfeld and Neckermann, 2011; Kosfeld et al., 2014), the evidence on their effectiveness and the channels through which they operate remains limited in the public sector with the exception of a few studies (see Ashraf, Bandiera, and Lee, 2014; Ashraf, Bandiera, and Jack, 2014; Gauri et al.,

2018). Another concern pertaining to incentives that are tournament-based (which non-financial incentives often tend to be) is that they can at times create negative effects such as damaging employee morale, self-efficacy, and self-esteem (see Lazear and Rosen, 1981; Ashraf, Bandiera, and Lee, 2014).<sup>1</sup> Overall, these gaps raise questions on the effectiveness of non-financial incentives in the public sector, the channels through which they operate, and also how to design them more effectively to address potential negative effects.

The existing evidence on financial incentives demonstrates that, when designed well, they can be a promising tool for improving performance (Muralidharan and Sundararaman, 2011; Duflo et al., 2012; Khan et al., 2016). However, some have argued for more caution. One key argument for this is that public sector tasks are often multi-dimensional which can lead to problems of multi-tasking and create perverse incentives by rewarding more on the sub-task that is easily measurable (Holmstrom and Milgrom, 1991). In other cases, challenges of measurement can dampen the effect of incentives leading to gaming, fraud, and collusion (Banerjee et al., 2008). Measures of subjective evaluation, which theoretically are a recourse in such cases as they can capture multiple dimensions, come with their own challenges. Such evaluations are mostly managed internally which can lead to wasteful time on ‘influence activities’ such as currying favors from supervisors, indulging in bribes, and becoming subject to supervisors biases (Gibbs et al., 2004; Prendergast, 2007; Neal, 2011). At the same time, subjective evaluations can also result in the effective use of local information for evaluations instead of patronage if the supervisor incentives are aligned with the objectives of the principal (Prendergast and Topel, 1996). While a large body of evidence has emerged on the impact of financial incentives across various contexts, more clarity is required on the role of financial incentives when tasks are multi-dimensional and they include both easy to measure and hard to measure dimensions, and where such tasks are carried out in contexts where supervisors may hold important private information about performance on tasks.

I address specific gaps in the academic literature on non-financial and financial incentives that

---

<sup>1</sup>In particular, this could happen due to two reasons. First, tournaments can create too much competition where workers may try to sabotage one another (Lazear and Rosen, 1981). Second, rankings involved in tournaments may challenge individual beliefs about ability which can result in lower effort as a response (Ashraf, Bandiera, and Lee, 2014)

are relevant to two different contexts - the Punjab Education Department and the Agriculture Department respectively.

Where education is concerned, existing evidence demonstrates that improving student learning outcomes remains a significant challenge across several developing country contexts (Glewwe and Muralidharan, 2016). Teachers are widely recognized as a central input into improving student learning outcomes (Das et al., 2012; Chetty et al., 2014). However, the evidence points towards challenges of teacher absenteeism, low time on task, and poor teacher quality as significant constraints to improving student learning outcomes (Chaudhury et al., 2006, Bold et al., 2017). Existing evidence on how to improve the effectiveness of public school teachers has largely focused on teacher recruitment, teacher characteristics that matter for performance, and teacher compensation and incentives (see Rockoff, 2004; Rivkin et al., 2005; Muralidharan and Sundararaman, 2011; Duflo et al., 2012; de Ree et al., 2018; Bau and Das, 2020). However, two critical gaps persist. First, there remains limited evidence on how public school teachers can learn more effectively on the job through in-service trainings which is the main vehicle for on-the-job skills and quality improvement for public school teachers. This is evident in the dearth of studies on the effectiveness of teacher in-service trainings or how to motivate teachers to learn more effectively in such trainings (Popova et al., 2016). Second, the literature on incentives largely focuses on financial incentives despite a role for non-financial incentives, especially given education is a pro-social department.

I tackle these gaps in the literature through designing and implementing two sets of experiments with the Teacher Training Academy in Punjab Pakistan, called the Quaid-e-Azam Academy for Educational Development (QAED). I first conduct a pilot study with 650 primary and secondary school teachers. Building upon the findings from the pilot, I conduct a larger experiment with 3,394 head teachers. Punjab, the provincial context of these studies, is Pakistan's most populous province. It employs a workforce of approximately 400,000 teachers spread across 52,000 schools.<sup>2</sup> QAED holds the mandate to offer in-service trainings to all public school teachers in Punjab to improve their skills and quality on the job. However, teacher quality remains a challenge and is widely seen as one of the main explanations for low student learning outcomes in the province.

---

<sup>2</sup>Annual School Census Data 2017.

Our baseline sample from the study in Chapter 1 presented below shows that more than 40% of teachers score less than 50% on grade 5 math problems.

In the pilot experiment (Chapter 1), I explore the following questions:

1. Can non-financial incentives such as employer recognition (that are tied to training test scores) improve teacher effort during training, and result in improved teacher knowledge as measured at the end of training? If yes, what are the main channels through which such incentives operate? We study this by designing three different employer recognition schemes.
2. How are the treatment effects of employer recognition schemes heterogeneous by two key moderators - source of bureaucrats' motivation (i.e. intrinsic, extrinsic, or pro-social) and personality traits of teachers?<sup>3</sup>

I randomize 650 primary and secondary school teachers attending a routine subject-based trainings at QAED (spread across four districts of Punjab) into three employer recognition schemes and a control group. The recognition incentive is a simple tournament-based incentive in which teachers are told that if they acquire the highest training score at the end of the training or the most improved score (over the pre-test), they can qualify for a prestigious certificate. In T1 (Private arm), the certificate is given to the teacher privately; in T2 (Peer arm), the certificate is given in front of peers and colleagues in a district-level ceremony; and in T3 (Career arm), the certificate is given privately but teachers are also told that their name would be added to an 'excellent teacher list' which will be shared with departmental leadership that can help them qualify for future career opportunities in their districts. The treatments are designed to leverage the following three main channels of employer recognition. First, agents may value employer recognition because of their motivation to acquire higher self-esteem and confidence. We call this the private channel. Second, agents may care about employer recognition because it allows them to gain peer-esteem or social respect (Frey, 1997; Besley and Ghatak, 2008). We call this

---

<sup>3</sup>This builds upon a nascent literature that demonstrates how non-cognitive traits such as source of motivation (intrinsic, extrinsic, or pro-social) or personality types can be important moderators for predicting heterogeneous impacts of incentives (see Callen et al., 2016; Deserranno, 2019; Lee, 2018).

the peer channel. Third, agents may value employer recognition because it can signal performance to supervisors which can in-turn help agents reap tangible career benefits in the future (Dewatripont et al., 1999; Ashraf, Bandiera, and Lee, 2014). We call this the career-benefits channel.

Results of our pilot experiment show that the source of individual motivation of bureaucrats for entering the civil service mediates the effectiveness of employer recognition. While the average treatment effects of each recognition scheme are null, we find that teachers who report *intrinsic* reasons for entering the service such as ‘interested in the profession of teaching’ exhibit a positive treatment response and those who report *extrinsic* reasons such as ‘salary’ show a negative treatment effect, with the two effects being significantly different from each other. Further analysis shows that the positive treatment effect for intrinsic individuals comes from the “peer” and “career” channels of employer recognition, with no effects on the “private” channel. Treatment effects on endline self-efficacy beliefs show positive effects for intrinsic but negative effects for extrinsic individuals, indicating that self-efficacy beliefs may be an important causal mechanism for these effects.

As a pilot study, this research points towards two important considerations for the design of the larger experiment. First, it indicates that more public or outward facing recognition arms may be more effective than private or inward facing recognition arms. This is inline with the existing literature on recognition which highlights the benefits of public versus private forms of recognition (as in Ashraf, Bandiera, and Lee, 2014; Ariely et al., 2009). Second, it points to the potential of improving the design of such recognition schemes through self-efficacy enhancing frames.

As a stand alone paper, this research points towards the potential negative effects of such incentives for teachers in in-service trainings, and also more broadly for teacher performance outside of training sessions. They highlight that the public school teacher workforce is heterogeneous by motivational orientations which has implications for how they respond to incentives. This implies caution against blanket administration of incentives given their potential to lower equilibrium effort depending on the non-cognitive traits of the workforce mix. More broadly, the results in this paper contribute to the literature on selection of bureaucrats and types of traits that

could be relevant for predicting performance. Our results show that motivational orientations of bureaucrats have a bearing on how individuals interact with employer recognition. This has further implications for how these sources of motivation may interact with other career incentives, opportunities for professional development, and other workplace features.

In the larger QAED experiment (Chapter 2), I explore the following main questions:

1. Can non-financial incentives such as employer recognition (that is tied to training test scores) improve teacher effort during training, and result in improved teacher knowledge as measured at the end of training? If yes, what is the impact of employer recognition that makes peer /collegial approval salient versus recognition that makes future tangible career benefits salient?
2. Does framing these recognition incentives with a self-efficacy enhancing frame improve their effectiveness?

I randomize 3,394 head teachers attending a mandatory government training into four different designs of recognition incentives that are tied to training performance and a control group. Treatment 1 (Peer arm) leverages the peer/collegial approval channel of recognition – trainees are told that those with the top score in the training post-test and the most improved score (over the pre-test) will be provided certificates in a district level ceremony which will be attended by their peers and colleagues in their district; Treatment 2 (Career arm) leverages the career-benefits channel of recognition – it is the same as Treatment 1 except trainees are told that those who qualify will receive the certificates privately but at the same time their name will be added to an ‘excellent teacher list’ (which will be shared with the School Education Department leadership) that can help them qualify for future opportunities in their districts. This leverages informal career incentives in the system such as postings to preferred schools, transfers to other lateral postings, or getting selected for promotion earlier once eligible; Treatments 3 and 4 (Public PLUS and Career PLUS) cross the first two treatments with a self-efficacy enhancing frame that aims to bolster teacher perceptions in their ability to do well in the training and their jobs more broadly.

I find that employer recognition can improve teacher training performance if it is linked to tangible career benefits in the future but the positive effects can also backfire depending on how these incentives are framed. When a self-efficacy frame is added to either of the first two treatments, it improves teacher self-efficacy but makes teachers overconfident, reducing performance in the training.

The results in this study have direct relevance for policy discussions on how to bolster teacher engagement, effort, and subsequently the acquisition of knowledge and skills in in-service trainings. Existing evidence, though limited, indicates that targeted instruction, provision of learning material alongside training, and linking teacher participation in training to incentives such as promotion or salary could be effective ways of improving the impact of in-service trainings (Popova et al., 2016). Our results indicate that non-financial incentives that leverage potential career benefits could also be feasible and cost-effective ways for improving the impact of in-service trainings, where some of these career benefits such as postings to preferred schools or transfers to other lateral positions may be less formal and easier to exercise in the system as compared to promotions. At the same time, our results indicate caution in how such incentives are framed and designed for use in teacher trainings. Our study is unique in framing incentives with a self-efficacy enhancing frame (and exogenously varying the latter) and showing the pitfalls of how such a framing could unrealistically raise teacher beliefs about ability, resulting in negative effects on performance.

The results in Chapter 2 and Chapter 1 highlight important differences which have implications for the use of non-financial incentives in teacher trainings in particular, and for teacher performance more broadly. The pilot was conducted with primary and secondary school teachers while the main experiment was conducted with head teachers. While the peer/collegial approval channel does not work for head teachers in Chapter 2, it works for primary and secondary school teachers in the pilot study for intrinsic teachers. Our qualitative discussions highlight that head teachers may not value social approval by peers/colleagues because they have already risen through the ranks and established peer-esteem. However, this may not be true for primary and secondary school teachers who are more junior. This indicates different types of non-financial incentives may work differently across in-service training programmes targeted at different cadres.

This highlights the highly contextual nature of these type of incentives (as in Gauri et al., 2018) and the sensitivity of such incentives to group-type and framing effects.

More broadly, the results in both Chapter 1 and Chapter 2 have implications for types of non-financial incentives that may be effective in addressing a slack in teacher effort in the school and the classroom. While it is true that formal career incentives in public bureaucracies are negligible given promotions are determined by seniority, our results indicate that head teachers have career concerns through other informal mechanisms and recognition incentives that are linked to such concerns can create an opportunity to reap these benefits (when and if the time comes). While formal incentive-based reforms can often be hard to implement, understanding the sources of different teacher motivations - informal career concerns in the system being one such source - and designing “soft” non-financial incentives around them is one way to address part of the weakness in incentive structures.

Chapter 3 of the thesis is based on a different contextual setting - the Agriculture Extension Department. Where public agriculture extension service is concerned, challenges of weak incentive and monitoring systems are common like many other critical frontline services. However, the problem is exacerbated in extension services due to the spatial spread of farmers and the nature of tasks of frontline extension agents that include both dimensions of outreach that are easy to measure (such as number of trainings provided) and dimensions of quality that are hard to measure (such as quality of advice given to farmers). This makes performance of public extension a classic agency problem where questions around how to best incentivize frontline extension staff are central, and where due to the nature of extension tasks, the known advantages and disadvantages of objective versus subjective incentive systems are unclear.

To address this question on the design of incentives in public extension (and the gaps in the literature on financial incentives when tasks are multi-dimensional as outlined above), we partner with the Agriculture Extension Department in Punjab Pakistan, and evaluate the impact of a province-wide pay-for-performance programme to answer the following questions:

1. Can performance incentives improve extension outreach and quality? If yes, is the impact of incentives that rely solely on objective (quantifiable) metrics larger or smaller than

incentives that allow for the use of supervisor discretion to assign incentives?

2. Is the use of discretion in subjective evaluations more effective if supervisor incentives are aligned with the objectives of the principal?

Leveraging the department's comprehensive digital performance management system called AgriSmart, the department designed an incentive programme with three different pay for performance schemes (and a control group) and randomized them across 126 tehsils in Punjab. In the first set of tehsils (Objective Arm), all field staff are assigned a bonus using only the metrics available from the AgriSmart portal. In the second set of tehsils (Subjective Arm), the supervisor (Assistant Director- AD) can assign a bonus to all staff under his purview and use the data from the portal as per his/her discretion. The third arm (Subjective plus) mirrors the second arm but adds an element of top-down monitoring to align the AD incentives with the objectives of the principal. In order to do this, a report on the performance of the AD's tehsil is shared with the top leadership of the department (Secretary, Director General, and Divisional Directors) every month. Early results show that while all treatments improve performance on AgriSmart metrics, Subjective Plus also shows positive and significant effects on farmer experience of extension outreach and quality, and farmer agricultural practices. In addition, we also find strong and positive effects on AD performance in their managerial tasks (such as assigning farmer training programmes to frontline extension agents) in the Subjective Plus arm, which form an important input into how frontline staff carry out extension outreach.

Additional data collection will explore why we observe treatment effects on measures of extension outreach and quality in the Subjective Plus arm across different measures (i.e. both AgriSmart measures and farmer reported measures of outreach and quality); and why the treatment effects on AgriSmart outreach measures in the Objective and Subjective arms do not translate to farmer reported measures of extension outreach and quality. Given the only difference in the Subjective Plus arm is the AD-layer of monitoring, we hypothesize two potential channels of AD behaviour change that could have resulted in improved outreach and quality. First, AD's own performance and practices could have changed which may have effected extension agents' performance. Second, AD's bonus assignment behaviour may have become more efficient which may have effected agents' performance. We will also explore mechanisms underlying how the

Objective and Subjective incentive systems may have worked to understand our results. This includes questions around whether the schemes encourage effort on different margins - i.e. on the extensive margin (i.e expanding outreach to more farmers) versus the intensive margin (i.e. repeating the same set of farmers more times) and how the schemes may have impacted staff satisfaction and motivation.

The existing results in this chapter contribute to several themes of literature. They contribute to the overall literature on performance pay in the public sector, where tasks have easy to measure and harder to measure dimensions, and where supervisors hold important private information about performance (see Burgess and Ratto, 2003; Manning et al., 2012 for surveys). Our results also contribute to the literature on rules-based bureaucracy versus discretion. While the challenges of discretion in subjective evaluation systems are well documented (Baker et al.,1994; Prendergast, 1999; Gibbs et al., 2004), we show circumstances under which they may or may not work well. Additional data collection on AD performance and practices will also contribute to the literature on how practices and performance of middle-management may relate to frontline service delivery (Rasul and Rogger, 2018).

The policy implications and further academic implications of my results are presented in more detail in each of my chapters and also in the conclusion of my thesis.

## References

- Ariely, Dan et al. 2009. “Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially.” *American Economic Review* 99 (1):544–555.
- Ashraf, Nava, Oriana Bandiera, and B. Kelsey Jack. 2014. “No margin, no mission? A field experiment on incentives for public service delivery.” *The Journal of Public Economics* 120:1.
- Ashraf, Nava, Oriana Bandiera, and Scott Lee. 2014. “Awards unbundled: evidence from a natural field experiment.” *Journal of economic behavior and organization* 100:44–63. URL <http://search.proquest.com/docview/1523807233/>.
- Baker, George et al. 1994. “Subjective performance measures in optimal incentive contrac.” *The Quarterly Journal of Economics* 109 (4):1125. URL <http://search.proquest.com/docview/210978266/>.
- Banerjee, Abhijit et al. 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 <http://search.proquest.com/docview/37031156/>.
- Bau, Natalie and Jishnu Das. 2020. “Teacher Value Added in a Low-Income Country †.” *American Economic Journal: Economic Policy* 12 (1):62–96.
- Besley, Timothy and Maitreesh Ghatak. 2005. “Competition and Incentives with Motivated Agents.” *American Economic Review* 95 (3):616–636.
- . 2008. “Status Incentives.” *American Economic Review* 98 (2):206–211.
- Bold, Tessa, Deon Filmer, Gayle Martin, Ezequiel Molina, Christophe Rockmore, Brian Stacy, Jakob Svensson, and Waly Wane. 2017. “What Do Teachers Know and Do? Does It Matter? : Evidence from Primary Schools in Africa.”
- Burgess, Simon and Marisa Ratto. 2003. “The Role of Incentives in the Public Sector Issues and Evidence.” *Oxford Review of Economic Policy* 19 (2):285–300.

- Bénabou, Roland and Jean Tirole. 2002. “Self-Confidence and Personal Motivation.” *The Quarterly Journal of Economics* 117 (3):871–915.
- Callen, Michael et al. 2016. “The Political Economy of Public Sector Absence: Experimental Evidence from Pakistan.” *NBER Working Paper Series* :22340URL <http://search.proquest.com/docview/1795921414/>.
- Chaudhury, Nazmul, Jeffrey Hammer, Michael Kremer, Karthik Muralidharan, and F. Halsey Rogers. 2006. “Missing in Action: Teacher and Health Worker Absence in Developing Countries.” *Journal of Economic Perspectives* 20 (1):91–116.
- Chetty, Nadarajan et al. 2014. “Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates.” 104 (9).
- Das, Jishnu et al. 2012. “Learning Levels and Gaps in Pakistan: A Comparison with Uttar Pradesh and Madhya Pradesh.” *Economic and Political Weekly* 47 (26/27):228–240.
- de Ree, Joppe, Karthik Muralidharan, Menno Pradhan, and Halsey Rogers. 2018. “Double for Nothing? Experimental Evidence on an Unconditional Teacher Salary Increase in Indonesia\*.” *The Quarterly Journal of Economics* 133 (2):993–1039.
- 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.
- Dewatripont, Mathias et al. 1999. “The economics of career concerns, Part I: comparing information structures.” *Review of economic studies* 66(1) (226):183–198. URL <http://search.proquest.com/docview/38702024/>.
- Duflo, Esther et al. 2012. “Incentives Work: Getting Teachers to Come to School.” *American Economic Review* 102 (4):1241–1278.
- Falk, Armin and Michael Kosfeld. 2006. “The Hidden Costs of Control.” *American Economic Review* 96 (5):1611–1630.

- Finan, Frederico et al. 2015. "The Personnel Economics of the State." *NBER Working Paper Series* :21825 URL <http://search.proquest.com/docview/1752001802/>.
- Frey, Bruno S. 1997. *Not just for the money : an economic theory of personal motivation*. Cheltenham: Edward Elgar.
- Frey, BS and S Neckermann. 2008. "Awards A View from Psychological Economics." *Zeitschrift Fur Psychologie-Journal Of Psychology* 216 (4):198–208.
- Gauri, Varun, Julian Jamison, Nina Mazar, Owen Ozier, Shomikho Raha, and Karima Saleh. 2018. "Motivating Bureaucrats Through Social Recognition: Evidence from Simultaneous Field Experiments." URL <http://search.proquest.com/docview/2063163854/>.
- Gibbs, Michael et al. 2004. "Performance Measure Properties and Incentives." *IDEAS Working Paper Series from RePEc* URL <http://search.proquest.com/docview/1698641182/>.
- Glewwe, P. and K. Muralidharan. 2016. *Improving Education Outcomes in Developing Countries*, vol. 5. 653–743.
- Holmstrom, Bengt and Paul Milgrom. 1991. "Multitask Principal - Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design." *Journal of Law, Economics, and Organization* 7:24. URL <http://search.proquest.com/docview/1300226943/>.
- Khan, Adnan Q. et al. 2016. "Tax farming redux: experimental evidence on performance pay for tax collectors.(Report)." 131 (1):219.
- Kosfeld, Michael and Susanne Neckermann. 2011. "Getting More Work for Nothing? Symbolic Awards and Worker Performance." *American Economic Journal: Microeconomics* 3 (3):86–99.
- Kosfeld, Michael et al. 2014. "Knowing that You Matter, Matters! The Interplay of Meaning, Monetary Incentives, and Worker Recognition." *SSRN Electronic Journal* .
- Lazear, Edward and Sherwin Rosen. 1981. "Rank-Order Tournaments as Optimum Labor Contracts." *The Journal of Political Economy* 89 (5):841. URL <http://search.proquest.com/docview/1290576073/>.

- Lee, Scott. 2018. "Intrinsic Incentives: A Field Experiment on Leveraging Intrinsic Motivation in Public Service Delivery." *SSRN Electronic Journal* .
- Luthans, F. and A.D. Stajkovic. 1999. "Reinforce for performance: The need to go beyond pay and even rewards." *Academy of Management Executive* 13 (2):49–57.
- Manning, Nick et al. 2012. *Public Sector Human Resource Practices to Drive Performance*. GET Note. World Bank, Washington, DC.
- Markham, Steven E. et al. 2002. "Recognizing Good Attendance: A Longitudinal, Quasi-Experimental Field Study." *Personnel Psychology* 55 (3):639–660.
- Muralidharan, Karthik and Venkatesh Sundararaman. 2011. "Teacher Performance Pay: Experimental Evidence from India." *Journal of Political Economy* 119 (1):39–77.
- Neal, Derek. 2011. *The Design of Performance Pay in Education [electronic resource]*. Working paper series (National Bureau of Economic Research : Online) ; no. 16710. Cambridge, Mass: National Bureau of Economic Research.
- Popova, Anna et al. 2016. "Training Teachers on the Job : What Works and How to Measure It."
- Prendergast, Canice. 1999. "The Provision of Incentives in Firms." *Journal of Economic Literature* 37 (1):7–63.
- . 2007. "The Motivation and Bias of Bureaucrats." *American Economic Review* 97 (1):180–196.
- Prendergast, Canice and Robert H. Topel. 1996. "Favoritism in Organizations." *Journal of Political Economy* 104 (5):958–978.
- 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.
- Rivkin, Steven G. et al. 2005. "Teachers, Schools, and Academic Achievement." *Econometrica* 73 (2):417–458.

Rockoff, Jonah E. 2004. "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data." *American Economic Review* 94 (2):247–252.

# 1

## The Double-Edged Sword of Non-Financial Incentives (Part I): Evidence from the Education Sector in Pakistan

# The Double-Edged Sword of Non-Financial Incentives (Pilot- Part I): Evidence from the Education Sector in Punjab, Pakistan

Zahra Mansoor<sup>§</sup>

October 2020

## ABSTRACT

This paper provides evidence that the source of motivation of frontline bureaucrats has implications for the impact of non-financial incentives such as employer recognition. We demonstrate this within the context of a pilot field experiment with the Punjab Teacher Training Academy in Pakistan where we embed different employer recognition schemes that are tied to training performance of public school teachers who attend a routine professional development training. We find that teachers who report *intrinsic* reasons for entering the service such as ‘interested in the profession of teaching’ exhibit a positive treatment response, teachers who report *pro-social* reasons such as ‘serving the community’ show no treatment response, whereas those who report *extrinsic* reasons such as ‘salary’ show a negative treatment effect. The impact of employer recognition for intrinsic and extrinsic teachers is significantly different which implies that such non-financial incentives have the potential to lower equilibrium effort depending on the heterogeneity in the source of bureaucrats’ motivation in the workforce mix. Further analysis shows that the positive treatment effect for intrinsic teachers comes from more outward/public facing versus inward/private facing recognition schemes. We also find teacher self-efficacy beliefs to be an important causal mechanism for these effects, with positive treatment effects on endline self-efficacy for intrinsic and negative treatment effects on endline self-efficacy for extrinsic teachers. The findings of this pilot study feed into the larger experiment presented as Chapter 2 of this thesis.

**Acknowledgements:** This research was made possible through numerous insights from stakeholders at QAED. In particular, I am thankful to Mr. Waseem Shirazi, Mr. Nadeem Hussain, and Mr. Adnan Bashier at the Quaid-e-Azam Academy for Educational Development . I gratefully acknowledge funding from Economic Development Institutions (EDI), Center for the Study of African Economis (CSAE), and the Blavatnik School of Government (BSG).

\* Dphil Candidate, Blavatnik school of Government

# 1 Introduction

A motivated and high-performing bureaucracy is central to effective public service delivery, economic growth, and development. This is especially important in the education sector where despite huge investments in various school-level inputs, learning outcomes have remained low (Glewwe and Muralidharan, 2016). Teachers are widely recognized as a key input into improving student learning outcomes (Rivkin et al., 2005; Chetty et al., 2014). However, across several developing country contexts low teacher quality, teacher absenteeism, and low time on task remain critical constraints to improving student learning outcomes (Bold et al., 2017).

Thus, it is not surprising that public sector education departments spend a significant proportion of their budgets on in-service and pre-service teacher training programmes as an investment into teacher quality.<sup>1</sup> At the same time, several governments have tested the impact of various types of financial incentives to elicit higher teacher effort. However, open questions remain on both fronts. Popova et al. (2016) evaluate 26 in-service teacher training programmes and show that trainings that have a subject focus, include follow-up visits, provision of textbooks alongside trainings, or the use of incentives for trainees are associated with positive impacts on student learning outcomes. However, they caveat these findings given the small sample size in their study and highlight that despite the scale of investments in in-service trainings, the evidence on their effectiveness or how to improve teacher engagement and effort within these trainings remains sparse. Where financial incentives are concerned, while the evidence shows that when designed well they can be effective (Muralidharan and Sundararaman, 2011; Duflo et al., 2012), opponents argue that they can often be distortionary via creating perverse incentives or crowding out intrinsic motivation (Glewwe et al., 2010; Bénabou and Tirole, 2003). Non-financial incentives offer a recourse - especially in pro-social settings like education where teachers may put a lower weight on financial incentives (Besley and Ghatak, 2005). However, evidence on the effectiveness of non-financial incentives or the channels through which they operate remains limited in the public sector with the exception of a few studies (see Ashraf, Bandiera, and Lee, 2014; Ashraf,

---

<sup>1</sup>For example, two thirds of World Bank programmes with an education component had elements of teacher professional development training (Popova et al., 2016).

Bandiera, and Jack, 2014; Gauri et al., 2018).<sup>2</sup>

To address these gaps in the literature, I design a pilot field experiment in collaboration with the teacher training academy (Quaid-e-Azam Academy of Educational Development) in Punjab Pakistan to understand the impact of non-financial incentives, in particular employer recognition, in improving teacher performance in in-service trainings. In doing so, the study asks two key questions. First, whether employer recognition can improve teacher engagement and acquisition of knowledge in in-service trainings? If yes, what are the main channels through which it operates? In addition, the study also aims to test the role of non-cognitive traits in how teachers respond to our recognition schemes. A nascent literature is beginning to demonstrate how non-cognitive traits such as source of motivation (intrinsic, extrinsic, or pro-social) or personality types can be important moderators for predicting heterogeneous impacts of incentives (see Callen et al., 2016; Lee, 2018). In our pilot experiment, we test how treatment effects of recognition schemes may be heterogeneous by two key moderators - source of bureaucrats' motivation and personality traits of teachers. To operationalize sources of motivation, we categorize teachers into three broad motivational orientations for entering the civil service - 'intrinsic' reasons such as interest/enjoyment/self-enrichment in the profession, 'pro-social' reasons such as serving the community or the country, and 'extrinsic' reasons such as salary or any other external feature of the environment (Deci et al., 1989; Amabile et al., 1995). We capture personality traits using the Big Five Inventory (John et al., 2008).

The study is situated in the education sector of Punjab Pakistan. Punjab is the most populous province of the country employing a workforce of approximately 400,000 teachers spread across 52,000 schools.<sup>3</sup> Student learning outcomes in Punjab have shown little progress. For example, the ASER (2019) report shows that more than 40% of children in grade 5 have not reached grade 2 levels of learning in literacy and numeracy. In addition, teacher quality and effort remain critical concerns. In our experimental sample, more than 40% of teachers scored less than 50% on grade 5 math problems in their training pre-test indicating the scope for potential gains by bolstering teacher effort in trainings. However, there are no incentives or other motivational tools that can

---

<sup>2</sup>A significant literature has looked at the impact of non-financial incentives in the private sector (see Stajkovic and Luthans, 1997; Markham et al., 2002; Frey and Neckermann, 2008).

<sup>3</sup>Annual School Census Data 2017.

help bolster teacher participation and engagement in trainings. More broadly, anecdotal evidence highlights that the current incentive structure for teachers within the education department remains unbalanced with too much monitoring but very few formal financial or non-financial rewards.

The pilot field experiment is implemented in direct collaboration with the Quaid-e-Azam Academy of Educational Development (QAED) by embedding the recognition incentives within a specific type of in-service training called professional development (PD) days. These are one-day in-service training days in which primary and secondary school teachers are offered training in specific topics of Maths, English, Urdu, and Science. In our experiment, teachers attending the PD day training across four districts of Punjab are randomized into different designs of recognition schemes. The treatments are designed to leverage the following three main channels of employer recognition. First, agents may value employer recognition because of their motivation to acquire higher self-esteem and confidence. We call this the private channel. Second, agents may care about employer recognition because it allows them to gain peer-esteem or social respect (Frey, 1997; Besley and Ghatak, 2008). We call this the peer channel. Third, agents may value employer recognition because it can signal performance to supervisors which can in-turn help agents reap tangible career benefits in the future (Dewatripont et al., 1999; Ashraf, Bandiera, and Lee, 2014). We call this the career-benefits channel. The recognition incentive is a simple tournament-based incentive in which teachers are told that if they acquire the highest training score or the most improved score (over the pre-test) in the training, they can qualify for a prestigious certificate (authorized and undersigned by the QAED headquarters). In T1 (Private arm), the certificate is given to the teacher privately; in T2 (Peer arm), the certificate is given in front of peers and colleagues; and in T3 (Career arm), the certificate is given privately but teachers are also told that their name would be added to an ‘excellent teacher list’ which will be shared with departmental leadership (such as QAED leadership and Secretary Education) that can help them qualify for future career opportunities in their districts.

The study reveals the following findings. We find no treatment effects on training test scores for any of the recognition treatments. However, we find that treatment effects are heterogeneous by our two key moderators - source of motivation (i.e. motivational orientation) and personality

traits. Teachers who report *intrinsic* reasons for entering the service such as ‘interested in the profession of teaching’ exhibit a 0.16 standard deviations increase in training scores (significant at the 10% level), teachers who report *pro-social* reasons such as ‘serving the community’ show no treatment response, and teachers those who report *extrinsic* reasons such as ‘salary’ show a 0.20 standard deviations decrease in training scores (significant at the 10% level). The treatment effects for the intrinsic and the extrinsic individuals are significantly different from each other at the 5% level of significance. The positive treatment effect for intrinsic individuals is consistent across all arms but primarily comes from the Peer and the Career arms (0.24 and 0.21 standard deviations higher test score at the 10% level of significance respectively). The negative treatment effect for the extrinsically motivated individuals is also consistent across all arms but is the most stark for the Career arm. The treatment coefficients for personality types move in opposite directions as motivational orientation, with positive treatment effects for above median personality and negative treatment effects for below median personality. However, the effects for above median or below median personality types are not significantly different from each other.

Supplementary evidence helps explore mechanisms for understanding the opposite treatment effects for intrinsic and extrinsic teachers. We hypothesize two potential explanations for the observed effects. First, we are interested in understanding how extrinsic and intrinsic individuals value the external and internal benefits of a recognition certificate. Second, we are interested in understanding whether our recognition schemes effect the endline measures of our secondary outcomes such as self-efficacy and locus of control for intrinsic and extrinsic teachers in different ways, which could have had subsequent (and different) effects on effort of teachers across these two groups.

With regards to how extrinsic and intrinsic teachers value the external and internal benefits of employer recognition, we find that extrinsic treated teachers in the Peer and Career arms are 28% points and 18% points less likely to report valuing a certificate for external reasons (such as social respect or potential career benefits) in comparison to the control group. This effect (in both the Peer and Career arms) is significantly different from intrinsic teachers. This result is puzzling for two reasons. First, we would expect extrinsic teachers to place a higher value on the external benefits of the certificate as compared to the control group. Second, we would expect

extrinsic teachers in the Peer and Career arms (which make external benefits of the certificate salient) to be more likely to report valuing the certificate for external reasons as compared to extrinsic teachers in the more inward facing Private arm. However, our results show that extrinsic teachers are less likely to report valuing the certificate for external reasons and this effect primarily comes from the outward facing Peer and Career arms. Our qualitative discussions with 10 randomly selected extrinsically motivated teachers highlight that these teachers did not believe non-financial rewards could benefit them in any tangible way. They argued that non-financial rewards like recognition were simply offered for optics with no real value to the recipients. Based on this line of reasoning, our results suggest that extrinsic teachers who primarily care about financial incentives or salary strongly disagreed that non-financial rewards such as recognition could bring any tangible external benefit. This internal disagreement within extrinsic teachers may have had a backlash in the form of lower effort in response to the treatment.

The opposite effects for intrinsic and extrinsic teachers are also explained by treatments effects for intrinsic and extrinsic teachers on individual self-efficacy. While treated intrinsic individuals show a significant increase of 0.29 standard deviations in self-efficacy, extrinsic individuals show a significant decrease of 0.27 standard deviations. A similar trend is observed across all treatment arms with the effects on self-efficacy for intrinsic and extrinsic individuals being significantly different from each other for all recognition schemes. This offers additional explanation for why effort may have crowded out for extrinsic teachers but crowded-in for intrinsic teachers.

The results in this paper contribute to the emerging literature on strategies for improving teacher effort in in-service trainings. The results highlight that the public school teacher workforce is heterogeneous by motivational orientations which has implications for how they respond to incentives. Our results also speak to the literature on the potential negative effects of incentives which argues that administering incentives can at times crowd out effort by decreasing employee morale, perceived competence, or self-efficacy (as highlighted by Lazear and Rosen, 1981; Deci, Koestner, and Ryan, 1999). Our results uniquely show that this “crowding-out” effect only holds for extrinsic individuals, which highlights that whether incentives crowd out effort or not also depends on such traits of individuals. Overall our results highlight that there are important heterogeneities in non-cognitive traits of the workforce which can help explain

why in some contexts incentives work but in others they do not. Therefore, depending on the workforce mix, such incentives have the potential to lower equilibrium effort. This requires caution in their design and against their blanket administration both in in-service trainings and more widely to elicit higher teacher effort in the classroom.

More broadly, the results in this paper contribute to the literature on selection of bureaucrats which underscores the importance of types of traits that could be relevant for potential screening methodologies. These results point to the importance of considering motivational orientations of bureaucrats at the time of selection. In doing so, it also highlights that intrinsic motivations can be of many different types where intrinsic motivation from serving the community (i.e. pro social motivation) can be different from intrinsic motivation derived from interest in the job. The source of these motivations is likely to have a bearing on how individuals interact with career incentives, opportunities for professional development, and other workplace features.

As a pilot study, this research also points towards two important considerations for the design of the larger study. First, it indicates that more public or outward facing recognition arms may be more effective than private or inward facing recognition arms. This is inline with the existing literature on recognition which highlights the benefits of public versus private forms of recognition (as in Ashraf, Bandiera, and Lee, 2014; Ariely et al., 2009). Second, it points to the potential of improving the design of such recognition schemes through self-efficacy enhancing frames. While the benefits of self-efficacy enhancing interventions has been documented for a range of social and economic outcomes such as job search, employment, and health seeking behaviours (see Eden and Aviram, 1993; Haushofer, John, and Orkin, 2019), their potential remains to be tested in combination with incentives. We address these questions in our larger experiment presented as Chapter 2 of the thesis.

This paper is organized as follows. Section 2 presents the theory and related literature backing our experimental design. Section 3 describes the background and context of our experimental setting. Section 4 describes the experimental design and randomization. Section 5 presents key data sources. Section 6 presents the empirical strategy and main results. Section 7 concludes.

## 2 Theory and Related Literature

In this section, we outline the theory which supports our experimental design and the key questions and hypotheses.

The design of the treatment arms is based on three different underlying theories of why agents may value employer recognition. First, agents may value employer recognition because of their motivation to acquire higher self-esteem and confidence. We call this the private channel. Second, agents may care about employer recognition because it allows them to gain peer-esteem or social respect (Frey, 1997; Besley and Ghatak, 2008). We call this the peer channel. Third, agents may value employer recognition because it can signal performance to supervisors which can in turn help agents reap tangible career benefits in the future (Dewatripont et al., 1999; Ashraf, Bandiera, and Lee, 2014). We call this the career-benefits channel.

Our experimental design includes three different recognition treatments where each treatment makes one of these channels salient. While the existing empirical literature documents the impact of public versus private forms of employer recognition (as in Ashraf, Bandiera, and Lee, 2014) or the impact of employer versus community recognition (as in Gauri et al., 2018), no existing study looks at the different channels of employer recognition and how the strength of each varies from the other. A unique contribution of our study is therefore to make the career-benefits channel explicit in the design and to study the strength of each of these channels.

While the existing literature does document positive impacts of recognition, part of the evidence also shows how recognition rewards which tend to be tournament based can often have negative effects by damaging employee morale, self-efficacy, or beliefs about ability which can result in lower effort (Lazear and Rosen, 1981; Deci, Koestner, and Ryan, 1999; Bénabou and Tirole, 2003; Ashraf, Bandiera, and Lee, 2014). While the ‘crowding out’ of pro-social motivation (and effort) by incentives has mostly been documented in the case of financial incentives (see Titmuss, 1970; Gneezy and Rustichini, 2000) and appears to be relatively less common for non-financial incentives such as recognition (see Ashraf, Bandiera, and Jack, 2014), what is less clear is how effort may be crowded out by negative effects on these other non cognitive traits such as individual

self-efficacy, beliefs about ability, and employee morale.

Our first set of hypotheses directly follow from the theory outlined above and are as follows:

**Hypothesis I:** Recognition that leverages the motivation for individual self-esteem and self confidence improves teacher training performance as measured through post-training test scores relative to the control group.

**Hypothesis II:** Recognition that leverages the motivation for peer-esteem and peer approval improves teacher training performance as measured through post-training test scores relative to the control group.

**Hypothesis III:** Recognition that leverages the motivation for tangible career benefits in the future improves teacher training performance as measured through post-training test scores relative to the control group.

**Hypothesis IV:** Receiving any of the treatments does not crowd out pro-social motivation relative to the control group.

**Hypothesis V:** Receiving any of the treatments reduces self-efficacy and shifts locus of control from internal to external relative to the control group.

In addition to the average treatment effects of our recognition schemes, we also test heterogeneous treatment effects by non-cognitive traits of teachers. Economists and psychologists are increasingly recognising how non-cognitive traits are important predictors of a range of social and economic outcomes (Borghans et al., 2008; Besley and Ghatak, 2018). Some empirical work specifically highlights how non-cognitive traits such as source of motivation (i.e. motivational orientation) and personality traits may predict varying responses to incentives. For example, Callen et al. (2016) show how personality types predict responses of Punjab health inspectors to a monitoring intervention, Deserranno (2019) shows that higher pro-social motivation tends to be correlated with better performance of health workers in Uganda, and Lee (2018) shows how intrinsically (as opposed to extrinsically) motivated health workers in India respond more to intrinsic incentives. In our pilot, we focus our attention on two key types of non-cognitive

traits by which we expect our treatment impact to vary: source of motivation (i.e. motivational orientation) and personality traits of teachers.

While a small but growing empirical evidence shows how motivational orientation or personality types may matter for how agents respond to incentives, theory on the subject is limited. Mainly, two different theoretical conceptualisations help explain how differences in such traits can result in varying responses to incentives. Callen et al. (2016) and Almlund et al. (2011) model personality traits as a way of identifying different types of individuals, where individuals with better traits are better types who can produce more with a fixed amount of effort (either due to higher efficiency or lower costs of effort). Lee (2018), on the other hand, presents a standard utility maximizing framework in which agents can derive extrinsic or intrinsic returns from effort when incentives are rolled out. He allows agents to have different preferences over their extrinsic and intrinsic returns to effort depending on their own motivational orientation. The key implication of the framework is that whether or not agents respond to intrinsic incentives depends on their valuation of the marginal utility from intrinsic incentives. While the two conceptualizations look at personality types or motivational orientation, they argue that agents with different types of non-cognitive traits can respond to incentives in different ways either due to different costs for exerting the same amount of effort or due to different preferences over returns from incentives.

While our experiment does not explicitly calibrate cost of effort or individual preferences underlying their returns from recognition, we add to the existing empirical literature by studying whether treatment impacts for recognition incentives are heterogeneous by these non-cognitive traits. In particular, we test the following secondary hypotheses:

**Hypothesis VI:** Better personality types respond more to the different recognition schemes as compared to the worse types.

**Hypothesis VII:** Intrinsically motivated teachers will respond more to recognition schemes that rely on internal rewards versus extrinsically motivated teachers.

**Hypothesis VIII:** Extrinsically motivated teachers will respond more to the recognition schemes that rely on external rewards versus recognition schemes that rely on internal rewards.

While we recognise power limitations in testing hypotheses VII and VIII, we intend to do this analysis for its exploratory value in our pilot.<sup>4</sup> Studying the heterogeneous response of motivational orientation by the three different recognition schemes is especially important. This is because while the private channel of recognition has no external benefit, the peer and career-benefits channels include an element of external benefit such peer/collegial respect or potential career benefits. Given extrinsically motivated teachers are more likely to respond to treatments arms with external benefits, we expect teachers with different motivational orientations to respond differently across the three different treatment arms.

Ultimately, we also aim to use the findings from this study to be able to better calibrate individual preferences for such rewards, including how such preferences would vary by the non-cognitive traits that we aim to study.

### 3 Context

In this section, we describe the broader context of the Education Department in Punjab and our specific experimental context of the Teacher Training Academy, Quaid-e-Azam Academy of Educational Development (QAED).

#### 3.1 Education Department in Punjab, Pakistan

This experiment is placed within the context of the Punjab education sector which employs a workforce of approximately 400,000 teachers across 52,000 schools that serve 11 million children across the province.<sup>5</sup> Amongst efforts to improve education outcomes in the province, the education budget has doubled in the last 5 years, the government has set up an extensive school monitoring programme (including yearly and monthly audits), and instituted a merit-based teacher

---

<sup>4</sup>Note that hypothesis VI can also be tested for the pooled treatment since we do not hypothesize different outcomes by treatment

<sup>5</sup>Annual School Census Data 2017

recruitment strategy.<sup>6</sup> Despite these efforts, learning outcomes have remained poor. The ASER (2019) report shows that nearly 40% of children in grade 5 have not reached grade 2 levels of learning in literacy and numeracy.

Anecdotal evidence indicates that stringent monthly monitoring measures have eroded teacher motivation, and teacher effort in the classroom remains low if not worsened. While the extensive school monitoring system in Punjab may have addressed part of the agency problem, it is primarily linked to input indicators like attendance (which is a poor indicator of teacher effort in the classroom) or school enrolment (which is not directly under teacher control). In addition, promotions tend to be based on seniority and there are no mechanisms to reward good performers. This has resulted in an incentive system that relies heavily on negative reinforcements and no rewards. Given discouraging existing evidence on the effectiveness of financial incentives in improving student learning outcomes in the Punjab Education Department (see Barrera-Osorio and Raju, 2017), there is direct policy interest in understanding the role of non-monetary incentives in improving teacher motivation and effort.

### 3.2 Punjab Teacher Training Academy

Our experiment is set-up in collaboration with the Quaid-e-Azam Academy for Educational Development in Punjab (QAED), the provincial teacher training academy which holds the mandate for all professional development needs of public school teachers in Punjab and thus improving teacher quality on the job.

Amongst many different types of professional development trainings, one type of training that QAED conducts is a one-day needs-based training in subject content, which the department calls ‘professional development’ (PD) day trainings. The goal of PD day is to identify student learning objectives where student test scores are low and offer targeted trainings to teachers in those areas. The training model in Punjab is a cascaded model where trainings are first provided by provincial-level trainers to district-level trainers, followed by district-level trainers to cluster-

---

<sup>6</sup>I-SAPS (Institute of Social and Policy Sciences) report on Public Financing in Education in Pakistan ([http://i-saps.org/upload/report\\_publications/docs/1496496299.pdf](http://i-saps.org/upload/report_publications/docs/1496496299.pdf)).

level trainers, and finally by cluster-level trainers to primary school teachers.<sup>7</sup> The PD Day trainings are also offered through this cascaded approach (See Figure 1).

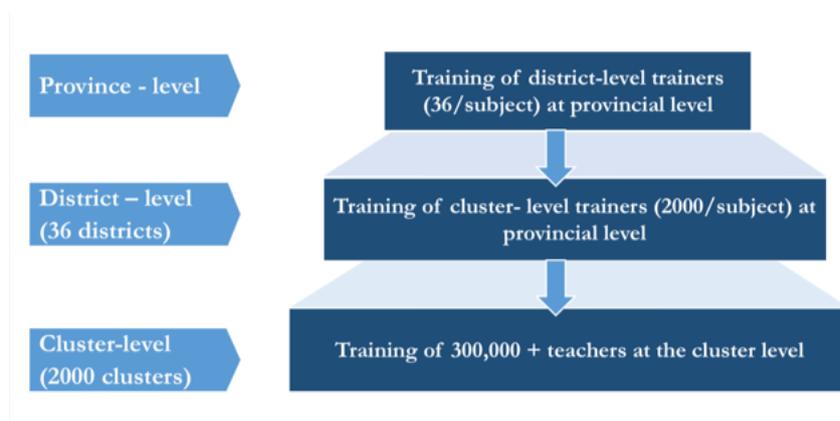


Figure 1: Punjab Cascaded Model of Training

Despite investments in various professional development trainings, there is little rigorous evidence on whether teachers engage and learn in these trainings and how their learning is translated to the classroom. Data from our baseline training test scores shows that there is indeed potential to improve teacher quality via improving their subject-based knowledge further. On average, around 40% of all teachers who attended trainings in our target districts score less than 50% on Grade 5 problems in Maths, English, and Science (see Figure 2). However, at present there are no formal incentives or motivational strategies to encourage higher teacher engagement in these trainings.

Our experiment is embedded within the QAED subject-based PD Day trainings at the district-level that took place in January and March 2017 for cluster-level trainers who are primary and secondary school teachers.<sup>8</sup> The trainings were offered by district-level master trainers for each subject. While PD Day trainings are offered for all subjects and across all district of Punjab, the experiment focuses on four priority subjects of the department - English, Urdu, Mathematics,

<sup>7</sup>The ‘cluster’ is the smallest administrative boundary which is specific to the education sector and includes around 12-17 schools on average.

<sup>8</sup>The trainings are thus relevant for not only improving subject-based knowledge of these cluster-level trainers for their own teaching but also for how well the training is cascaded the training further down to primary school teachers

and Science - in four districts of Punjab spread across the north, south, and central regions of the province.

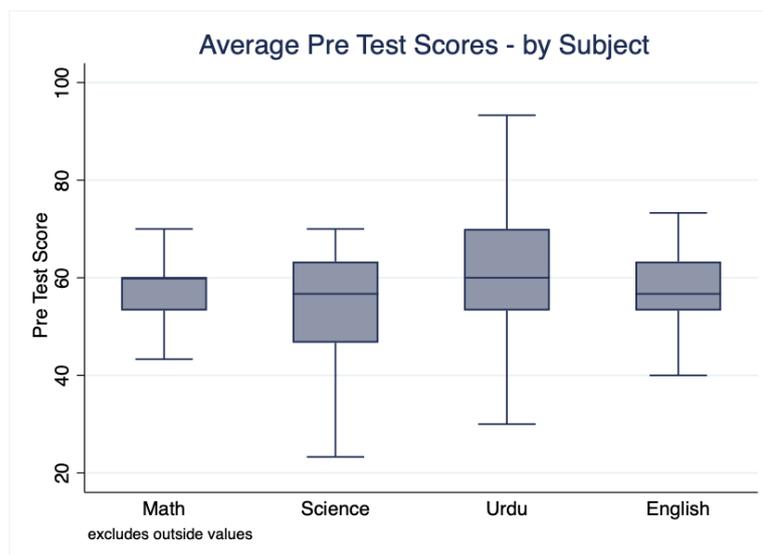


Figure 2: Pre Test Scores by Subject Type

## 4 Experimental Design and Randomization

### 4.1 Randomization and Sample

Our experiment is implemented in four districts of Punjab (Lahore, Faisalabad, Attock, and Bahawalpur) as shown in Figure 3. We focus on trainings in four subjects, where each subject training is spread over two days in each of our districts. This gives us a total of 32 training days (4 districts x 4 subjects x 2 days). Stratifying at the training day level, we randomize teachers in each session across the three treatment groups and a control group.

Teachers in our sample are on average 34 years of age with 8-9 years of experience in teaching, 55% of the teachers are male, and the average training baseline test score is 58% (See Table A.2 for details).

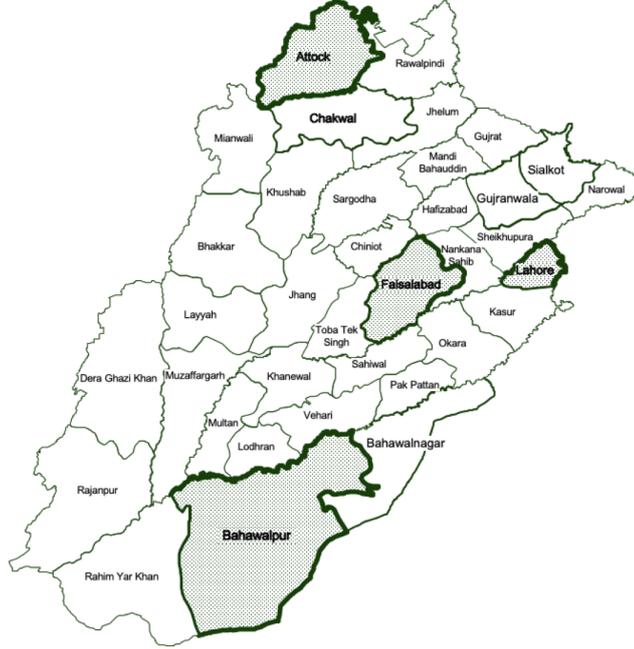


Figure 3: Districts for the Experiment

## 4.2 Treatment Arms

As mentioned in Section 2, the three recognition schemes leverage different underlying theories for why agents may care about employer recognition. The recognition incentive is a typical tournament-based incentive. Each group to which the treatment is administered to is told that the teacher who scores the highest score in the training post-test or shows the maximum improvement over the pre-test (highest value-added) can qualify for a prestigious departmental certificate. This qualification criteria encourages participation across the distribution (as in Ashraf, Bandiera, and Lee, 2014). Given each group has 5-6 teachers on average, 30-40% of the individuals in the group can qualify for the reward. While each treatment group receives the same recognition certificate, the way in which the reward is *distributed* and *framed* is varied across treatments. A key feature of our design is that the training setting allows us to incentivize teacher training test scores as the outcome measure which is one-dimensional in effort and relatively easier to measure and incentivize than teacher effort in the classroom. This allows us to avoid the standard issues of multi-tasking in the design of performance-linked incentives

(Holmstrom and Milgrom, 1991).

Design details of the three recognition schemes are as follows:

- 1) Private Recognition: In this arm, the certificate won by the teacher is distributed privately. This treatment leverages the individual's motivation for self-esteem to acquire employer recognition.
- 2) Peer Recognition: In this arm, the certificate won by the teacher is distributed at a public ceremony at the district training academy which is attended by the teacher's peers and colleagues in the district. This leverages the individual's motivation for peer/collegial approval and peer esteem.
- 3) Career Recognition: In this arm, the certificate won by the teacher is distributed privately as in the private recognition arm but at the same time career concerns are made salient. The teacher is also informed that if he/she wins, his/her name will be included in an 'Excellent Teacher List' which will be shared with the departmental leadership such as the district training head, QAED provincial head, and Secretary Education, which may help secure future opportunities in the department. This arm leverages individual's motivation for receiving informal career benefits through employer recognition such as transfer to a school of liking, transfer to another more preferable lateral position of the same pay and grade, or getting selected for promotional appointments earlier once eligible.<sup>9</sup>

All treatment groups are informed that the winning teachers will receive their certificates within two months once all PD Day trainings have concluded.

---

<sup>9</sup>While it is reasonable to argue that the peer recognition arm could also make career concerns salient, we believe that Peer arm makes social respect and peer-esteem more salient and the Career arm makes career concerns more salient.

### 4.3 Sequencing of Training Day Activities

In this section, we outline the sequencing of the PD Day training activities to outline where and how we embed the experimental design in the regular PD day schedule.

The standard PD Day training includes lessons on three different topics over a six hour training day, with an hour and a half devoted to each lesson. Lesson topics are based on Grade 5 student learning objectives and are narrow enough to be able to cover in the allocated time. Examples of lesson topics include two digit fractions in Maths, use of pronouns in English, or reflection and refraction in Science. Each training day begins at around 8:00 am and ends at 2:00 pm and is conducted by the assigned QAED master trainer.

To embed the experimental design within this training, we modify the training day to collect baseline and endline data from teachers, conduct training pre and post tests, and embed the administration of our recognition schemes. The activities in the PD Day follow the following sequence during our experiment:

1. Registration: Upon arrival to the training, each teacher is handed out a teacher registration survey by the team of enumerators. The registration survey serves as our baseline survey.
2. Randomization: Once the registration form is complete, the enumerator collects the registration form, and randomly assigns a token to each teacher. The token numbers are pre-mapped onto either one of the treatments or the control group. Hence, teachers are automatically randomised into one of the treatments based on their token number. At this stage, teachers are only distributed their token numbers but are not informed of their treatment group.
3. Pre-test: Once the registration and token distribution is complete, the QAED master trainer conducts a pre-test which is based on the lessons to be taught that day. Once the test is complete, it is collected by the master trainer. Teachers do not learn about their baseline test performance. This serves as our baseline test score measure.<sup>10</sup>

---

<sup>10</sup>We explain the format of the tests in more detail in Section 5

4. Administration of recognition schemes: Following the pre-test, the QAED master trainer divides the classroom into different groups based on the assigned token numbers, where each token number represents a treatment group. Teachers are told that they are divided into groups for a “group activity” and should continue to sit in the same seating arrangement for the rest of the training. An enumerator conducts the said group activity with each of the groups following a pre-determined script based on the treatment group. All scripts contain two different elements - First, a “neutral” messaging component which is the same across all treatments and the control group; and second, an “incentive component” which is only administered to the treatment groups. The incentive component of each script varies by type of treatment.
5. Training: Once the activity concludes, the QAED master trainer delivers the training. Teachers continue to sit in the same arrangement as the group-work activity. This reduces the likelihood of spillovers.
6. Post-test: Once the training concludes, teachers appear for a training post-test.
7. Feedback survey: At the end of the training, there is a post-training feedback survey through which we collect our additional endline measures.

## 5 Performance Measurement and Data Sources

### 5.1 Primary and Secondary Outcomes

Our primary outcome of interest is teacher training test scores, that we collect at baseline at the start of the training (pre-test) and at endline after the training concludes (post-test). The questions in the tests are tightly linked to the subject content that is taught in the training. While the pre-test and post-test include questions on the the same learning constructs, they include slight differences in the content of the questions. For example, in a Math test the same question about 2-digit subtraction would be included in the pre and post-test but the numbers

would be different. In an English test a question on identifying pronouns would remain the same but the options would be different.

To ensure the validity of items that were used in the tests, we encouraged the department to use items developed by examination experts at the Punjab Examination Commission (PEC). The items developed by PEC were further verified by an external group of item writers for quality control. To build in reliability, we tried to eliminate guesswork by encouraging trainees to choose ‘I don’t know’ if they did not know an answer.<sup>11</sup>

In addition to training test scores, we also collect additional data to measure the impact of our recognition treatments on a selected set of secondary outcomes that we capture through our post-training endline survey. Our key secondary outcomes include teacher perceptions on departmental support, credibility of the department, job commitment, teacher self-efficacy, and locus of control.

Self-efficacy captures individuals’ judgement of their own capabilities to achieve a certain goal. It originates from the Social Cognitive Theory developed by Bandura (1986) in which he argues “what people think, believe, and feel affects how they behave”, and that individuals’ own assessment of their abilities strongly influences the courses of action that they take. Where locus of control is concerned, deCharms (1968) distinguished between internal and external locus of control, where in the former the individual views himself as the “origin” of his/her behaviour whereas in the latter the individual views his/her behaviour as being outside control. We capture both self-efficacy and locus of control through prior tested scales.<sup>12</sup>

After our analysis of treatment effects on our primary and secondary outcomes, we also held qualitative discussions with 10 randomly selected extrinsic and 10 randomly selected intrinsic teachers to understand our results.

---

<sup>11</sup>We explored introducing negative marking for choosing wrong answers but did not incorporate that due to the department’s concerns around creating too much pressure for the trainees.

<sup>12</sup>Job commitment was captured using tested scales from Angle and Perry (1981); Self-efficacy was measured using a 4-point tested scale by TALIS as in Fackler and Malmberg (2016).

## 5.2 Measures for Heterogeneous Treatment Effects

### 5.2.1 Measures of Non-Cognitive Traits

We collect measures of non-cognitive traits for our two key moderators - motivational orientation and personality traits.

Teacher source of motivation (or motivational orientation) for entering the civil service is captured by asking respondents the question: “What most influenced you to enter the service?” as in Rasul and Rogger (2018). While the public administration and (more recently) the economics literature commonly focus on one type of intrinsic motivation which is characterized by existing levels of altruism and/or pro-social motivation, the psychology literature, points to the importance of ‘types’ or ‘degrees’ of intrinsic and extrinsic motivations. In their self-determination theory, Deci et al. (1989) present a framework where individuals have different motivational orientations ranging from extrinsic to intrinsic. Amabile et al. (1995) develop an inventory of questions to capture extrinsic and intrinsic motivations, and highlight that such motivations can have several components - extrinsic motivations can include the motivation for money, recognition, or any other external feature of the environment. Similarly, intrinsic motivation can be driven from pure enjoyment or interest, self-enrichment, or the need to serve the community where the latter is commonly termed as pro-social motivation.

To operationalize sources of motivation, we categorize teachers into three broad motivational orientations for entering the civil service based on their response to the question presented above - ‘intrinsic’ reasons such as interest/enjoyment/self-enrichment in the profession, ‘pro-social’ reasons such as serving the community or the country, and ‘extrinsic’ reasons such as salary or any other external feature. While individuals could have both intrinsic and extrinsic motivations at the same time, our measure makes teachers force rank their *primary* motivation which allows us to identify “types” of individuals along these different dimensions.<sup>13</sup> To map the responses of individuals into these three categories, we follow the categorization of extrinsic versus intrinsic

---

<sup>13</sup>It is also intuitive that individuals vary in their types of motivational orientations at work – the general categorization of individuals into either ‘sell outs’ who choose a job for extrinsic reasons such as money or ‘true to self’ who work in low-paying but internally gratifying jobs is fairly common.

motives as in Amabile et al. (1995) and Gagné et al. (2010). For example, an individual who joined the service for financial benefits would be categorized as extrinsic, an individual who joined the service for interest in the job or the pure joy of it would be categorized as intrinsic, and an individual who joined the service to serve their country or community would be categorized as pro-social. Table A.2 shows that our sample of teachers is approximately equally spread across these three different motivational orientations, with 38% of the teachers reporting intrinsic reasons, 33% reporting extrinsic reasons, and 28% reporting pro-social reasons.

We capture personality traits using the Big Five Inventory (John et al., 2008). This captures individual personality across five dimensions: agreeableness, conscientiousness, openness, extraversion, and neuroticism (or emotional stability). Due to time constraints during the training, we use the 10-point short BFI inventory instead of the original 44-item scale, which has been shown to retain significant levels of reliability and validity (Rammstedt and John, 2007; Soto and John, 2017). To compute our measure of the personality index, we calculate z-scores of each dimension and calculate the mean of all z-scores (as in Callen et al., 2016).

Further details on these measures is given in Appendix B.

### **5.2.2 Basic Teacher Profile**

We also collect data on a range of basic teacher profile and job characteristics. These include measures such as age, gender, years of experience, salary, number of teachers known in the training session, number of times a teacher got promoted, time expected till next promotion, and visibility to senior leadership.

## **5.3 Balance Checks**

We conduct tests for equality across three categories of measures – basic teacher characteristics such as age, gender, salary, and education; the baseline measure of performance which is the pre-test training score; and non-cognitive traits which includes measures of motivational orientation

and personality types. Table A.3 shows that the the pooled treatment is balanced against the control for all variables, except the personality trait extraversion.<sup>14</sup>

We also conduct tests of equality for all treatments – Private, Peer, and Career - against the control group. Table A.4 shows that treatments are balanced against the control group for all variables except extraversion for two treatments and salary for one of the treatments. The joint F tests for all treatments have a p-value of greater than 0.50.

Attrition is not a concern in our sample given the experiment is embedded within a single day of the training. Spillovers are controlled by ensuring that teachers in the same treatment continue to sit with teachers in the same treatment group.

## 6 Empirical Strategy and Results

### 6.1 Average Treatment effects on Post-Test Scores

In this section we present average treatment effects of our three recognition schemes. For identification of these effects, we do an ANOCOVA estimation as shown below:

$$y_{igs} = \alpha + \gamma \underline{y}_{igs} + \sum_j \beta_j T_{igs}^j + \zeta X_{igs} + \nu_s + \epsilon_{igs} \quad (1)$$

where  $y_{igs}$  is the post-test score for teacher  $i$  in group  $g$  (group-level at which treatment is administered) and stratum  $s$  (training day) and  $\underline{y}_{igs}$  is the pretest score which is our baseline measure for the ANCOVA estimation. Both training pre-test and post-test are normalized by the mean and standard deviation of the control group allowing us to observe treatment effects in standard deviation units.  $T_{igs}^j$  is the treatment indicator that takes the value 1 if teacher  $i$  is assigned to treatment  $j$  and 0 otherwise;  $X_{igs}$  is a vector of teacher controls; and  $\epsilon_{igs}$  is the error term. We include strata fixed effects (training day level),  $\nu_s$ , to control for all time-invariant

---

<sup>14</sup>The joint F test has a p-value of greater than 0.95

training day characteristics such as day and time when the training was held and master trainer effects. We assume errors to be correlated within groups in which treatments are administered, and hence we cluster errors at the group-level within each stratum.

Table 1 shows the average treatment effects on post-test scores for the pooled treatment (columns 1 and 2) and across our three different treatments (columns 3 and 4). We find that the average treatment effects for all treatments are null. In addition, the magnitude of the coefficients on each of the treatments is also small suggesting that there was little movement in post-test scores across treatments (see Figure 4 below).

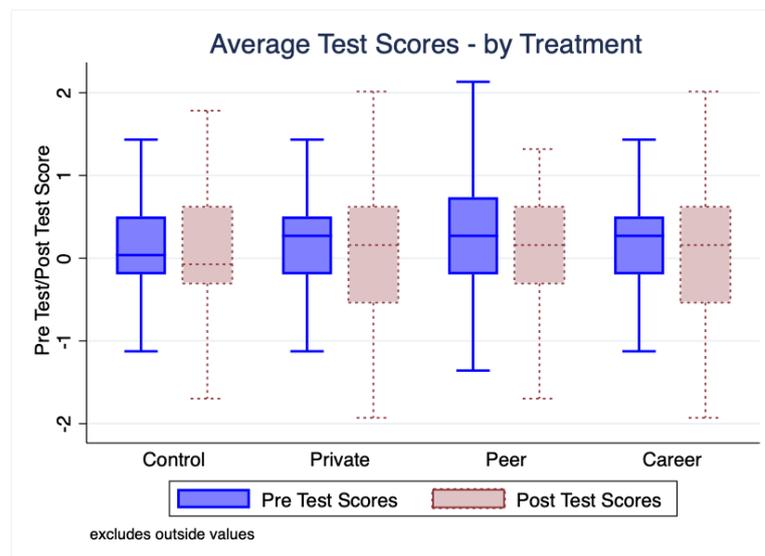


Figure 4: Mean of Pre and Post Test Scores - By Treatment

## 6.2 Heterogeneous Treatment effects by Moderators

In this section, we present heterogeneous treatment effects by our two key moderators - motivational orientation of individuals and their personality traits. While average treatment effects are null, we expect that the treatment effects may be heterogeneous by these moderators. As outlined in Section 2, individuals with different traits can respond differently to such incentives due to either being more efficient which lowers their cost of effort or due to different valuations of the benefits of such rewards.

### 6.2.1 Heterogeneous Treatment effects by Motivation Type

To estimate the heterogeneous treatment impact by different motivation types of individuals, we estimate equation 2 below:

$$y_{igs} = \alpha + \gamma \underline{y}_{igs} + \sum_k \rho_k M_{igs}^k + \sum_k \beta_k M_{igs}^k * T_{igs} + \zeta X_{igs} + \nu_s + \epsilon_{igs} \quad (2)$$

where  $M_{igs}^k$  is motivation  $k$  for teacher  $i$  in group  $g$  and stratum  $s$  ( $k=1$  if teacher reports intrinsic reasons for entering service,  $k=2$  if teacher reports extrinsic reasons, and  $k=3$  if teacher reports pro-social reasons).  $\beta_k$  is the heterogeneous treatment impact of the motivation type  $k$ . Since teachers are either in the intrinsic, extrinsic, or pro-social category, this specification allows us to compare the treatment coefficients of each of the motivational orientations directly. To estimate heterogeneous treatment effects for each recognition scheme, we use the same specification but restrict the sample to the treatment in question and the control group.

Table 2 shows that the impact of the treatment is heterogeneous by motivational orientation. Column 1 presents the heterogeneous treatment effects for the pooled treatment. Individuals who are intrinsically motivated (by interest/enjoyment in the job) show a 0.16 standard deviation higher test score when administered any treatment. In comparison, individuals who are extrinsically motivated (by either salary or any other extrinsic feature) show a 0.20 standard de-

viation lower test score in response to any treatment. Individuals who are pro-socially motivated (by motivation to serve the country or community) show no response to the treatment. The treatment effects for the intrinsic and the extrinsic individuals are significant at the 10% level, and significantly different from each other at the 5 % level of significance.

Columns 3-8 show that the positive treatment effect for intrinsic individuals is consistent across all arms but primarily comes from the Peer and the Career arms, with coefficients of 0.24 and 0.21 respectively significant at the 10% level. The negative treatment effect for the extrinsically motivated individuals is also consistent across all arms but is the most stark for the Career arm. Figure 5 depicts these treatment effects graphically.<sup>15</sup>

These results highlight the following main points. First, while the treatments work for intrinsic teachers, they do not work for extrinsic teachers. The treatment effect for intrinsic teachers also points attention towards different types of intrinsic motivations. While the existing economics and public administration literature largely focuses on one specific type of intrinsic motivation which is ‘pro-sociality’ borne out of the desire to serve one’s country or the community, the psychology literature also emphasizes other types of intrinsic motivations such as interest or enjoyment in the task itself, self-satisfaction, and self enrichment (Amabile et al., 1995). Our analysis distinguishes between intrinsic motivation from interest/enjoyment in the job and pro-sociality and shows that these differences matter for how teachers respond to our recognition schemes.

Second, when the treatment works for intrinsic teachers, the effect comes from more public/external facing recognition schemes (Peer and Career arms). This is also in line with the existing empirical evidence on the subject which shows that public-facing recognition rewards tend to have higher incentive power than private-facing rewards (Ashraf, Bandiera, and Lee, 2014; Markham et al., 2002; Ariely et al., 2009).

Third, when the treatment works negatively for extrinsic teachers, the effects are negative for both the Peer and Career arm but are more stark for the latter. This is puzzling since the

---

<sup>15</sup>we acknowledge that our sample is not powered enough for the by-treatment heterogeneous treatment effects but since this is a pilot study, we present these for their exploratory value.

Career arm is more extrinsic in design which should have elicited a smaller negative (or even positive) effect for extrinsic individuals. Our qualitative discussions with 10 randomly selected extrinsically motivated teachers highlighted that these teachers did not believe non-financial rewards could benefit them in any tangible way. They argued that non-financial rewards like recognition were simply offered for optics with no real value to the recipients. Based on this line of reasoning, one possible explanation for the stronger negative response in the Career arm could be that treatments that emphasized potential external benefits of the schemes created more cognitive dissonance (or internal disagreement) for extrinsic teachers resulting in a larger negative response. This is explored further in Section 6.3 where we discuss mechanisms for these effects.

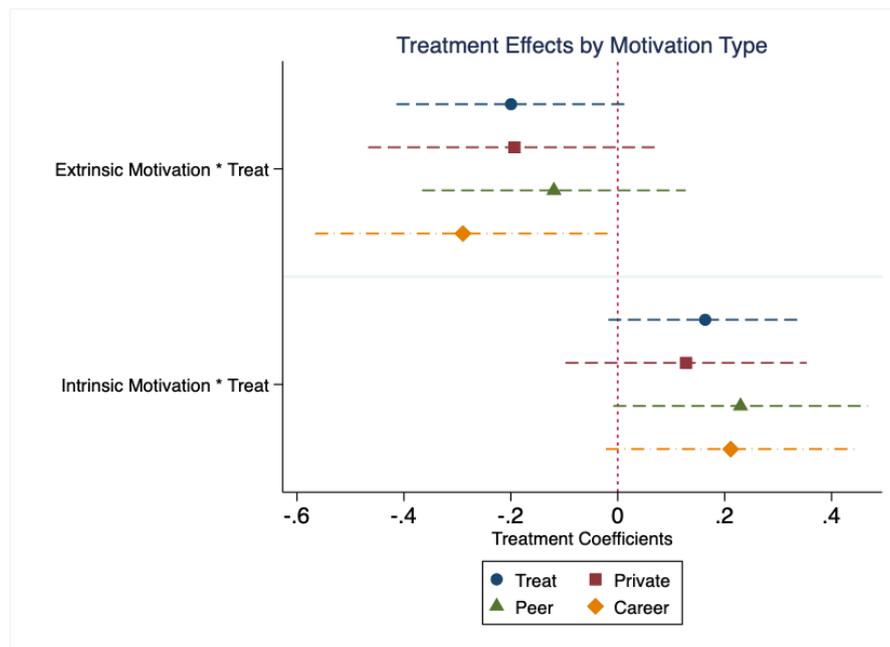


Figure 5: Heterogeneous Treatment Effects - By Motivation Type

### 6.2.2 Heterogeneous Treatment effects by Personality

To explore heterogeneous treatment effects by personality traits, we estimate the following:

$$y_{igs} = \alpha + \gamma \underline{y}_{igs} + \rho P_{igs}^1 + \beta_1 P_{igs}^1 * T_{igs} + \beta_2 P_{igs}^2 * T_{igs} + \zeta X_{igs} + \nu_s + \epsilon_{igs} \quad (3)$$

where  $P_{igs}^1$  is the below median category of the personality index,  $P_{igs}^2$  is the above median category, and  $\beta_1$  and  $\beta_2$  are the coefficients of interest. As in the specifications above, we include stratum (i.e. training day) fixed effects and cluster errors at the group-level within each stratum. To assess heterogeneous treatment effects for each incentive scheme, we use the same specification but restrict the sample to the treatment in question and the control group.

Table 3 shows that the treatment coefficients for personality types also move in opposite directions as intrinsic and extrinsic teachers, with positive treatment coefficients for teachers who have above median personality traits and negative treatment coefficients for teachers who have below median personality traits. The opposite treatment effects, however, are smaller in comparison to the difference between intrinsic and extrinsic teachers and not significantly different from each other (except for the Career arm).

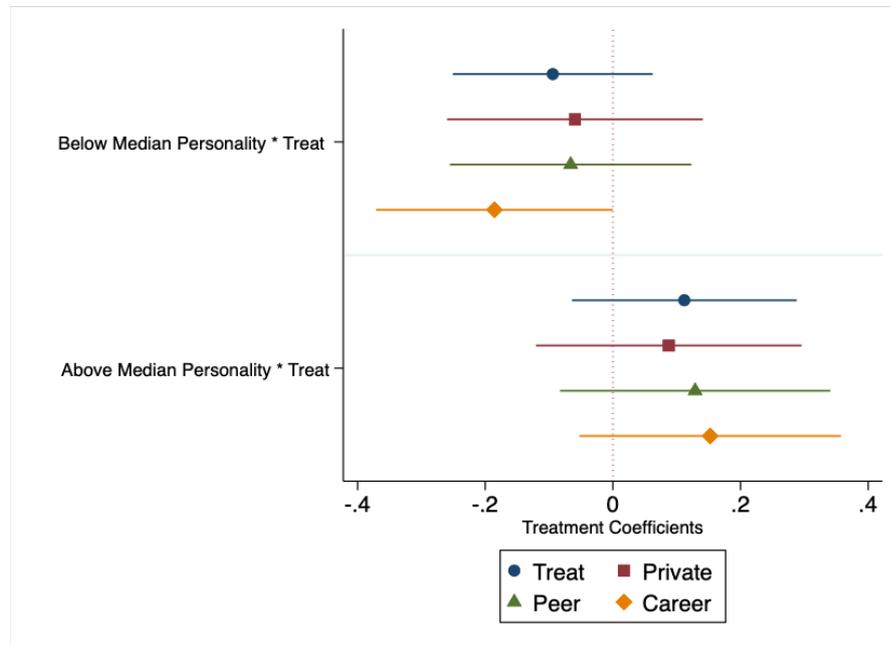


Figure 6: Heterogeneous Treatment Effects - By Personality Type

### 6.3 Mechanisms

In this section, we explore mechanisms underlying the opposite treatment effects for intrinsic and extrinsic teachers. We hypothesize two potential explanations for the observed effects. First, we are interested in understanding if intrinsic and extrinsic teachers value the internal and external benefits of recognition rewards differently. An interpretation of this based on a standard utility maximising framework would be that intrinsic and extrinsic individuals have different preferences and assign different weights to the marginal extrinsic and intrinsic benefit received from these rewards. Second, we are interested in understanding whether our recognition schemes effect our endline measures of self-efficacy and locus of control differently for intrinsic and extrinsic teachers, which could have had subsequent (and different) effects on teacher effort across these two groups.

### 6.3.1 Value of Benefits from Recognition Incentives

We try to explore if the value that extrinsic and intrinsic individuals place on the benefits of these rewards is different. Building upon the line of inquiry suggested by our qualitative discussions in Section 6.2.1, we explore if extrinsic teachers are less likely to recognise external benefits of non-financial rewards.

While we do not have the data to carefully calibrate the internal and external value that agents receive from the different recognition schemes, we use one of the questions from our endline survey to make progress on this front. For the subset of teachers who report that a departmental recognition certificate would be useful for them (which is around 99% of the sample equaling 644 teachers), the survey includes a follow-up question which asks the primary reason for why a departmental recognition certificate would be valuable to them. We map the responses to this question for all 644 teachers into either intrinsic benefits, extrinsic benefits, or mixed reasons for valuing the certificate.<sup>16</sup> Intrinsic benefits include internal reasons such as self-confidence or self-satisfaction, whereas extrinsic benefits include external benefits such as social respect or potential tangible career benefits. Following this mapping, we construct a dummy variable that is equal to one if teachers report strictly external reasons for valuing the certificate. We construct a similar dummy variable that is equal to one if teachers report strictly internal reasons for valuing the certificate.

Table 4 shows that extrinsic treated individuals are 19% points less likely to report external reasons for valuing the certificate (significant at the 5% level) in comparison to the control group (columns 1 and 2). Intrinsic treated teachers on the other hand have a coefficient of 0.01 which is statistically insignificant. The coefficients on the extrinsic and intrinsic teachers are statistically different from each other at the 5% level. A similar trend is observed across the Peer and Career arms. Extrinsic teachers in the Peer and Career arms are 28% and 18% points less likely to report external reasons for valuing the certificate in comparison to the control group (columns 6 and 8). The coefficients on the extrinsic and intrinsic teachers in the Peer and Career arms are statistically different from each other (at the 1% and 10% level respectively), suggesting that

---

<sup>16</sup>The mapped responses approximately fall equally into these three categories.

extrinsic teachers value external benefits of recognition less than intrinsic teachers.

These results lend support to our qualitative discussions and suggest that part of the reason why extrinsic teachers respond negatively to the treatments is due to their disagreement with the fact that such rewards can bring any tangible external benefit to them. Hence, the treatments which emphasized these types of external benefits (Peer and Career arms versus the private arm) appear to have heightened this internal disagreement which could have resulted in a larger negative response.

A competing explanation could be that extrinsic teachers are also more likely than intrinsic teachers to place a lower value on intrinsic benefits of recognition such as self-confidence or self-satisfaction. In Table A.6 in the appendix, we show that intrinsic and extrinsic teachers do not value the intrinsic benefits from the certificate differently confirming that the differences mainly come from potentially different valuations of the external benefits from the certificates. It is important to note that this analysis is unable to offer insights into why the treatment effects are positive and significant for intrinsic teachers.

### **6.3.2 Treatment Effects on Self-Efficacy**

We are also interested in understanding how our recognition schemes effect our secondary outcomes such as self-efficacy and locus of control for intrinsic and extrinsic teachers. This mechanism is very much in line with the literature on the negative effects of incentives which argues that incentives can often reduce effort by reducing employee morale (Lazear and Rosen, 1981), threatening individual beliefs of perceived competence and self-efficacy (Deci, Koestner, and Ryan, 1999), or adversely effecting perceptions of ability (Bénabou and Tirole, 2003). As outlined in Section 2, these negative effects are often viewed as potential ways through which incentives can “crowd-out” effort (Deci, Koestner, and Ryan, 1999; Bénabou and Tirole, 2003). Our analysis, hence, aims to explore if self-efficacy beliefs are impacted differently for treated intrinsic and extrinsic teachers, which could have resulted in crowding-out of effort for intrinsic and extrinsic teachers in different ways.

Table 5 shows that the average treatment effects on self-efficacy are null except for a positive and marginally significant effect on the Private arm. Table A.8 in the appendix also shows average treatment effects on other secondary outcomes such as intrinsic motivation and locus of control which are also null.

Table 6 shows significant and opposite treatment effects on endline self-efficacy for intrinsic and extrinsic individuals, both for the pooled treatment and across our three different recognition schemes. While treated intrinsic individuals show a significant increase of 0.29 standard deviations in self-efficacy, extrinsic individuals show a significant decrease of 0.27 standard deviations (column 1). A similar trend is observed across all treatments with the effects on self-efficacy for intrinsic and extrinsic individuals being significantly different from each other (column 3, 5, & 7).

These results indicate two points. First, self-efficacy appears to be an important channel through which crowding-in or crowding-out of effort takes place. These results are in line with the predictions of Deci, Koestner, and Ryan (1999), Bénabou and Tirole (2003), and Lazear and Rosen (1981) who argue that incentives may reduce individual self-efficacy and/or morale which can in-turn crowd out effort. Second, and an important distinction between our result and the predictions of the existing literature, the negative effects on self-efficacy only hold true for extrinsic teachers. This highlights that whether in equilibrium effort is crowded out or not (through negative effects on non-cognitive traits such as morale or individual self-efficacy) depends on the heterogeneity of individuals in the workforce.

## 7 Conclusion

We present experimental evidence from a pilot randomized controlled trial that is implemented in collaboration with the Punjab Teacher Training Academy called the Quaid-e-Azam Academy of Educational Development. We embed three different recognition schemes within a routine professional development training of primary and secondary school teachers that incentivize teachers to perform well in the training by tying the recognition reward to training test scores

- a private recognition scheme that awards high performers in teacher training privately; peer recognition that rewards high performers in a public ceremony attended by peers and colleagues; and career recognition that awards high performers in the teacher training privately but makes career benefits of recognition salient. Each treatment leverages a different underlying theory for why agents may value employer recognition.

The experiment shows null average treatment effects of the recognition schemes on training performance of teachers. However, we find heterogeneous treatment effects by the teachers' source of motivation - teachers who report 'intrinsic' reasons for entering the service such as 'interested in the profession of teaching' exhibit a positive treatment response and those who report extrinsic reasons such as 'salary' show a negative treatment effect, with the two effects being significantly different from each other. The treatment effects for each of the recognition schemes follow the same trend, however both the positive effect for intrinsic teachers and the negative effect for extrinsic teachers primarily come from the more outward-facing Peer and Career arms. The stronger positive effect for the more outward (or public) versus inward (or private) facing arms is inline with the existing literature which highlights that more public versus private recognition rewards appear to work better (Markham et al., 2002; Ashraf, Bandiera, and Lee, 2014). Further analysis shows that the negative effects for extrinsic teachers come from two channels. First, extrinsic teachers value the external benefits of recognition such as peer/social approval and potential career-benefits in the future much lesser than intrinsic individuals. Recognition schemes which make such external benefits salient (such as the Peer and Career arms) appear to cause further internal disagreement within extrinsic teachers resulting in a larger negative response. Second, treatment effects on endline self-efficacy show positive effects for intrinsic but negative effects for extrinsic individuals, with the effects being significantly different from each other. The same trend is observed across all recognition schemes. This offers additional explanation as to why effort may have crowded-out for extrinsic individuals but crowded in for intrinsic individuals.

Overall, these results highlight that there is heterogeneity in the sources of individual motivation that interact differently with incentives and with other non-cognitive traits. This indicates that recognition incentive schemes are highly contextual where the impacts on equilibrium effort depend on such heterogeneities in the workforce mix.

The results of this pilot have direct implications for the design of our main experiment on the impact of different recognition schemes which is presented as Chapter 2 of the thesis. In particular, we draw the following lessons from this pilot for our larger experiment. First, when the treatments work for intrinsic individuals, the effects primarily come from the Peer and Career arms which are more outward/public facing in design. Based on this, we drop the Private arm in our larger experiment. Second, self-efficacy appears to be an important causal mechanism through which effort is either crowded-in or crowded out. Framing existing incentive schemes with a self-efficacy enhancing frame therefore offers potential for improving the design of these schemes. We incorporate this in the experimental design of Chapter 2.

More broadly, the results in this pilot have direct relevance for how public sectors think about characteristics that are deemed important for screening mechanisms. While personality traits and public sector motivation have been looked at as potentially important traits (Callen et al., 2016; Deserranno; 2019), this paper shows that intrinsic versus extrinsic motivations within the workplace could also be important determinants of how individuals interact with career incentives, professional development opportunities, and overall performance on the job. In that particular respect, there is a need for improving measurement for capturing different types of intrinsic motivations.

While we view motivational orientations as stable traits in this paper, it is not entirely clear if intrinsic motivation cannot be shaped. Though limited, emerging evidence shows that intrinsic motivation could be shaped by changing existing mental frames of how individuals perceive themselves and their environment (see Akerlof and Kranton, 2005; Tanguy et al., 2014; Beaman et al., 2012; Collier, 2016). Further research needs to explore potential intrinsic incentives through which intrinsic motivation can be shaped.

Table 1: Main Treatment Effects on Post Test Scores

	(1)	(2)	(3)	(4)
Any Treatment	-0.001 (0.056)	0.006 (0.057)		
Private			-0.004 (0.072)	0.003 (0.073)
Peer			0.021 (0.069)	0.029 (0.070)
Career			-0.023 (0.067)	-0.016 (0.068)
Observations	650	648	650	648
Adjusted R <sup>2</sup>	0.070	0.067	0.067	0.064
PDS LASSO controls	No	Yes	No	Yes
Training Day Fixed Effects	Yes	Yes	Yes	Yes

Notes: All regressions are an Ancova estimation with baseline measures of the dependent variable. The regressions also include training day fixed-effects with errors clustered at the group level at which the treatment is administered. Standard errors are presented in parenthesis. The dependent variable is the incentivised post test scores normalized with the mean and standard deviation of the control group. Estimates are significant at the \*10%, \*\*5%, and \*\*\*1% level

Table 2: Heterogeneous Treatment Effects by Motivation Type, Post Test Scores

	Any Treat		Private		Peer		Career	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Intrinsic Motivation	-0.105 (0.244)	-0.115 (0.243)	-0.608* (0.328)	-0.623* (0.329)	-0.413 (0.254)	-0.416 (0.258)	-0.221 (0.153)	-0.238 (0.150)
Extrinsic Motivation	0.139 (0.244)	0.133 (0.242)	-0.323 (0.302)	-0.325 (0.303)	-0.162 (0.228)	-0.158 (0.232)	-0.009 (0.132)	-0.023 (0.131)
Pro-social Motivation	0.030 (0.237)	0.020 (0.236)	-0.442 (0.317)	-0.456 (0.319)	-0.277 (0.238)	-0.274 (0.243)	-0.101 (0.143)	-0.124 (0.142)
Intrinsic Motivation * Treat	0.164* (0.092)	0.174* (0.093)	0.128 (0.114)	0.143 (0.114)	0.230* (0.120)	0.239* (0.122)	0.212* (0.118)	0.221* (0.119)
Extrinsic Motivation * Treat	-0.200* (0.108)	-0.199* (0.108)	-0.193 (0.138)	-0.179 (0.137)	-0.119 (0.124)	-0.119 (0.124)	-0.290** (0.139)	-0.285** (0.140)
Pro-social Motivation * Treat	0.003 (0.103)	0.016 (0.105)	0.090 (0.126)	0.115 (0.130)	-0.078 (0.145)	-0.064 (0.147)	-0.060 (0.126)	-0.052 (0.129)
P-Value Difference								
Intrinsic*Treat - Extrinsic*Treat	0.02**	0.02**	0.11	0.11	0.07*	0.07*	0.02**	0.02**
Observations	649	647	321	321	361	360	351	350
Adjusted R <sup>2</sup>	0.072	0.070	0.152	0.154	0.074	0.069	0.071	0.067
PDS LASSO controls	No	Yes	No	Yes	No	Yes	No	Yes
Training Day Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: All regressions are an Ancova estimation with baseline measures of the dependent variable. The regressions also include training day fixed-effects with errors clustered at the group level at which the treatment is administered. Standard errors are presented in parenthesis. The dependent variable is the incentivised post test scores. Column 1 & 2 present heterogeneous treatment effects by pooling all treatments, column 3 & 4 presents heterogeneous treatment effects for the private treatment, column 5 & 6 for the peer treatment, and column 7 & 8 for the career treatment. Post-test and pre-test scores are normalized with the mean and standard deviation of the control group. Observed treatment effects are robust to the inclusion of controls as presented in this table. Estimates are significant at the \*10%, \*\*5%, and \*\*\*1% level

Table 3: Heterogeneous Treatment Effects by Personality Type, Post Test Score

	Any Treat		Private		Peer		Career	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Below Median Personality	0.186 (0.116)	0.185 (0.116)	0.205 (0.135)	0.211 (0.134)	0.172 (0.128)	0.168 (0.128)	0.124 (0.121)	0.126 (0.121)
Below Median Personality * Treat	-0.095 (0.078)	-0.097 (0.079)	-0.064 (0.099)	-0.061 (0.100)	-0.069 (0.094)	-0.077 (0.094)	-0.188** (0.092)	-0.202** (0.095)
Above Median Personality * Treat	0.110 (0.089)	0.116 (0.088)	0.087 (0.105)	0.091 (0.105)	0.128 (0.106)	0.142 (0.107)	0.150 (0.102)	0.146 (0.103)
P-Value Difference								
Above med - below med *Treat	0.11	0.09*	0.37	0.37	0.20	0.15	0.03**	0.03**
Observations	650	648	321	321	361	360	352	351
Adjusted R <sup>2</sup>	0.072	0.068	0.149	0.141	0.072	0.066	0.070	0.065
PDS LASSO controls	No	Yes	No	Yes	No	Yes	No	No
Training Day Fixed Effects	Yes	Yes						

Notes: All regressions are an Ancova estimation with baseline measures of the dependent variable. The regressions also include training day fixed-effects with errors clustered at the group level at which the treatment is administered. Standard errors are presented in parenthesis. The dependent variable is the incentivized post test scores. Column 1 & 2 present heterogeneous treatment effects by pooling all treatments, column 3 & 4 presents heterogeneous treatment effects for the private treatment, column 5 & 6 for the peer treatment, and column 7 & 8 for the career treatment. Post-test and pre-test scores are normalized with the mean and standard deviation of the control group. Observed treatment effects are robust to the inclusion of controls as presented in this table. Estimates are significant at the \*10%, \*\*5%, and \*\*\*1% level

Table 4: Heterogeneous Treatment Effects by Motivation Type, Certificate External Value

	Any Treat		Private		Peer		Career	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Intrinsic Motivation	0.072 (0.309)	0.074 (0.310)	-0.859*** (0.136)	-0.858*** (0.137)	-0.751*** (0.174)	-0.744*** (0.175)	0.125 (0.310)	0.134 (0.310)
Extrinsic Motivation	0.238 (0.314)	0.237 (0.314)	-0.690*** (0.141)	-0.687*** (0.142)	-0.559*** (0.176)	-0.552*** (0.178)	0.290 (0.315)	0.294 (0.315)
Pro-social Motivation	0.169 (0.310)	0.169 (0.310)	-0.737*** (0.149)	-0.736*** (0.151)	-0.648*** (0.182)	-0.637*** (0.184)	0.252 (0.312)	0.257 (0.312)
Intrinsic Motivation * Treat	0.011 (0.054)	0.006 (0.054)	-0.056 (0.079)	-0.055 (0.080)	0.083 (0.068)	0.079 (0.068)	0.014 (0.062)	0.008 (0.061)
Extrinsic Motivation * Treat	-0.185** (0.072)	-0.185** (0.072)	-0.077 (0.094)	-0.080 (0.094)	-0.283*** (0.082)	-0.286*** (0.083)	-0.175* (0.089)	-0.181** (0.090)
Pro-social Motivation * Treat	-0.129* (0.074)	-0.133* (0.075)	-0.105 (0.098)	-0.102 (0.100)	-0.179** (0.082)	-0.190** (0.084)	-0.122 (0.098)	-0.137 (0.099)
P-Value Difference								
Intrinsic*Treat - Extrinsic*Treat	0.02**	0.03**	0.86	0.83	0.01***	0.01***	0.08*	0.08*
Observations	644	644	319	319	358	358	347	347
Adjusted R <sup>2</sup>	0.008	0.004	0.019	0.013	0.046	0.039	0.006	0.003
PDS LASSO controls	No	Yes	No	Yes	No	Yes	No	Yes
Training Day Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: All regressions include training day fixed-effects with errors clustered at the group level at which the treatment is administered. Standard errors are presented in parenthesis. The dependent variable is a dummy variable equals to 1 if the trainee reported extrinsic value from a certificate such as social respect or receiving tangible benefits as the certificate's main value. Estimates are significant at the \*10%, \*\*5%, and \*\*\*1% level

Table 5: Average Treatment Effects on Self-efficacy

	(1)	(2)	(3)	(4)
Any Treatment	-0.005 (0.074)	-0.001 (0.074)		
Private			0.231* (0.122)	0.244** (0.123)
Peer			-0.125 (0.080)	-0.118 (0.081)
Career			-0.075 (0.086)	-0.080 (0.085)
Observations	631	631	631	631
Adjusted R <sup>2</sup>	-0.002	-0.002	0.011	0.011
PDS LASSO controls	No	Yes	No	Yes
Training Day Fixed Effects	Yes	Yes	Yes	Yes

Notes: All regressions include training day fixed-effects with errors clustered at the group level at which the treatment is administered. Standard errors are presented in parenthesis. The dependent variable is self-efficacy beliefs of teachers normalized by the mean and standard deviation of the control group. Estimates are significant at the \*10%, \*\*5%, and \*\*\*1% level

Table 6: Heterogeneous Treatment Effects by Motivation Type, Self-efficacy

	Any Treat		Private		Peer		Career	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Intrinsic Motivation	0.136 (0.625)	0.129 (0.624)	0.884*** (0.178)	0.701*** (0.152)	1.073*** (0.257)	1.042*** (0.261)	0.104 (0.655)	0.096 (0.656)
Extrinsic Motivation	0.370 (0.632)	0.361 (0.630)	1.125*** (0.167)	0.959*** (0.121)	1.342*** (0.249)	1.313*** (0.253)	0.385 (0.664)	0.366 (0.666)
Prosocial Motivation	0.384 (0.632)	0.366 (0.630)	1.146*** (0.195)	0.975*** (0.146)	1.291*** (0.264)	1.255*** (0.266)	0.346 (0.664)	0.316 (0.665)
Intrinsic Motivation * Treat	0.292** (0.145)	0.292** (0.145)	0.715* (0.361)	0.690** (0.346)	0.105 (0.137)	0.117 (0.137)	0.192 (0.126)	0.194 (0.126)
Extrinsic Motivation * Treat	-0.278*** (0.104)	-0.271** (0.105)	-0.291 (0.175)	-0.263 (0.181)	-0.370*** (0.101)	-0.360*** (0.105)	-0.152 (0.118)	-0.141 (0.119)
Prosocial Motivation * Treat	-0.119 (0.131)	-0.108 (0.127)	0.115 (0.217)	0.132 (0.215)	-0.066 (0.153)	-0.067 (0.150)	-0.263 (0.180)	-0.235 (0.174)
P-Value Difference								
Intrinsic*Treat - Extrinsic*Treat	0.01***	0.01***	0.02**	0.02**	0.01***	0.01***	0.04**	0.06*
Observations	630	630	311	311	349	349	338	338
Adjusted R <sup>2</sup>	0.008	0.007	0.024	0.047	0.021	0.026	0.004	0.008
PDS LASSO controls	No	Yes	No	Yes	No	Yes	No	Yes
Training Day Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: All regressions include training day fixed-effects with errors clustered at the group level at which the treatment is administered. Standard errors are presented in parenthesis. The dependent variable is self-efficacy beliefs of teachers normalized by the mean and standard deviation of the control group. Estimates are significant at the \*10%, \*\*5%, and \*\*\*1% level

## References

- Akerlof, George A and Rachel E Kranton. 2005. "Identity and the Economics of Organizations." *Journal of Economic Perspectives* 19 (1):9–32.
- Almlund, Mathilde et al. 2011. "Personality Psychology and Economics." *NBER Working Paper Series* :16822URL <http://search.proquest.com/docview/1687830854/>.
- Amabile, Teresa M., Karl G. Hil, Beth A. Hennessey, and Elizabeth M. Tighe. 1995. "The Work Preference Inventory: Assessing intrinsic and extrinsic motivational orientations": Correction." *Journal of Personality and Social Psychology* 68 (4):580–580.
- Angle, H.L. and J.L. Perry. 1981. "An Empirical Assessment of Organizational Commitment and Organizational Effectiveness." *Administrative Science Quarterly* 26:1–14.
- Ariely, Dan et al. 2009. "Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially." *American Economic Review* 99 (1):544–555.
- ASER. 2019. "Annual Status of Education Report - ASER-PAKISTAN 2019." URL <https://palnetwork.org/wp-content/uploads/2020/02/Annual-Status-of-Education-Report-ASER-PAKISTAN-2019.pdf>.
- Ashraf, Nava, Oriana Bandiera, and B. Kelsey Jack. 2014. "No margin, no mission? A field experiment on incentives for public service delivery." *The Journal of Public Economics* 120:1.
- Ashraf, Nava, Oriana Bandiera, and Scott Lee. 2014. "Awards unbundled: evidence from a natural field experiment." *Journal of economic behavior and organization* 100:44–63. URL <http://search.proquest.com/docview/1523807233/>.
- Bandura, Albert. 1986. *Social foundations of thought and action : a social cognitive theory*. Prentice-Hall series in social learning theory. Englewood Cliffs, N.J.: Prentice-Hall.
- Barrera-Osorio, Felipe and Dhushyanth Raju. 2017. "Teacher performance pay: Experimental evidence from Pakistan." *Journal of Public Economics* 148 (C):75–91.

- Beaman, Lori, Esther Duflo, Rohini Pande, and Petia Topalova. 2012. "Female leadership raises aspirations and educational attainment for girls: a policy experiment in India." *Science (New York, N.Y.)* 335 (6068):582.
- Besley, Timothy and Maitreesh Ghatak. 2005. "Competition and Incentives with Motivated Agents." *American Economic Review* 95 (3):616–636.
- . 2008. "Status Incentives." *American Economic Review* 98 (2):206–211.
- . 2018. "Prosocial Motivation and Incentives." *Annual Review of Economics* 10:411. URL <http://search.proquest.com/docview/2090263812/>.
- Bold, Tessa, Deon Filmer, Gayle Martin, Ezequiel Molina, Christophe Rockmore, Brian Stacy, Jakob Svensson, and Waly Wane. 2017. "What Do Teachers Know and Do? Does It Matter? : Evidence from Primary Schools in Africa."
- Borghans, L, Angela Lee Duckworth, James J. Heckman, and Bas Ter Weel. 2008. "The Economics and Psychology of Personality Traits." *The Journal of human resources* 43 (4):972–1059. URL <https://muse.jhu.edu/article/466652>.
- Bénabou, Roland and Jean Tirole. 2003. "Intrinsic and Extrinsic Motivation." *Review of Economic Studies* 70 (3):489–520.
- Callen, Michael et al. 2016. "The Political Economy of Public Sector Absence: Experimental Evidence from Pakistan." *NBER Working Paper Series* :22340URL <http://search.proquest.com/docview/1795921414/>.
- Chetty, Nadarajan et al. 2014. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." 104 (9).
- Collier, Paul. 2016. "The cultural foundations of economic failure: A conceptual toolkit." *Journal of Economic Behavior and Organization* 126:5–24.
- deCharms, Richard. 1968. *Personal Causation:: The Internal Affective Determinants of Behavior*. New York: Academic Press.

- Deci, Edward, Richard Koestner, and Richard Ryan. 1999. “A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation.” *Psychological Bulletin* 125 (6):627–668. URL <http://search.proquest.com/docview/203472909/>.
- Deci, Edward et al. 1989. “Self-Determination in a Work Organization.” *Journal of Applied Psychology* 74 (4):580. URL <http://search.proquest.com/docview/1290460137/>.
- 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.
- Dewatripont, Mathias et al. 1999. “The economics of career concerns, Part I: comparing information structures.” *Review of economic studies* 66(1) (226):183–198. URL <http://search.proquest.com/docview/38702024/>.
- Duflo, Esther et al. 2012. “Incentives Work: Getting Teachers to Come to School.” *American Economic Review* 102 (4):1241–1278.
- Eden, Dov and Arie Aviram. 1993. “Self-Efficacy Training to Speed Reemployment: Helping People to Help Themselves.” *Journal of Applied Psychology* 78 (3):352. URL <http://search.proquest.com/docview/1290321664/>.
- Fackler, Sina and Lars-Erik Malmberg. 2016. “Teachers’ self-efficacy in 14 OECD countries: Teacher, student group, school and leadership effects.” *Teaching and Teacher Education* 56.
- Frey, Bruno S. 1997. *Not just for the money : an economic theory of personal motivation*. Cheltenham: Edward Elgar.
- Frey, BS and S Neckermann. 2008. “Awards A View from Psychological Economics.” *Zeitschrift Fur Psychologie-Journal Of Psychology* 216 (4):198–208.
- Gagné, Marylène, Jacques Forest, Marie-Hélène Gilbert, Caroline Aubé, Estelle Morin, and Angela Malorni. 2010. “The Motivation at Work Scale: Validation Evidence in Two Languages.” *Educational and Psychological Measurement* 70 (4):628–646.

- Gauri, Varun, Julian Jamison, Nina Mazar, Owen Ozier, Shomikho Raha, and Karima Saleh. 2018. “Motivating Bureaucrats Through Social Recognition: Evidence from Simultaneous Field Experiments.” URL <http://search.proquest.com/docview/2063163854/>.
- Glewwe, P. and K. Muralidharan. 2016. “Improving Education Outcomes in Developing Countries: Evidence, Knowledge Gaps, and Policy Implications.” In *Handbook of the Economics of Education*, vol. 5. Elsevier, 653–743.
- Glewwe, Paul et al. 2010. “Teacher Incentives.” *American Economic Journal: Applied Economics* 2 (3):205–227.
- Gneezy, Uri and Aldo Rustichini. 2000. “A Fine is a Price.” *The Journal of Legal Studies* 29 (1):1–17.
- Haushofer, J, A John, and K Orkin. 2019. “Can Simple Psychological Interventions Increase Preventive Health Investment?” .
- Holmstrom, Bengt and Paul Milgrom. 1991. “Multitask Principal - Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design.” *Journal of Law, Economics, and Organization* 7:24. URL <http://search.proquest.com/docview/1300226943/>.
- John, Oliver et al. 2008. *Paradigm shift to the integrative big five trait taxonomy: History, measurement, and conceptual issues*, vol. Vol. 3. 114–158.
- Lazear, Edward and Sherwin Rosen. 1981. “Rank-Order Tournaments as Optimum Labor Contracts.” *The Journal of Political Economy* 89 (5):841. URL <http://search.proquest.com/docview/1290576073/>.
- Lee, Scott. 2018. “Intrinsic Incentives: A Field Experiment on Leveraging Intrinsic Motivation in Public Service Delivery.” *SSRN Electronic Journal* .
- Markham, Steven E. et al. 2002. “Recognizing Good Attendance: A Longitudinal, Quasi-Experimental Field Study.” *Personnel Psychology* 55 (3):639–660.
- Muralidharan, Karthik and Venkatesh Sundararaman. 2011. “Teacher Performance Pay: Experimental Evidence from India.” *Journal of Political Economy* 119 (1):39–77.

- Popova, Anna et al. 2016. “Training Teachers on the Job : What Works and How to Measure It.”
- Rammstedt, Beatrice and Oliver P John. 2007. “Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German.” *Journal of Research in Personality* 41 (1):203–212.
- 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.
- Rivkin, Steven G. et al. 2005. “Teachers, Schools, and Academic Achievement.” *Econometrica* 73 (2):417–458.
- Soto, Christopher J and Oliver P John. 2017. “Short and extra-short forms of the Big Five Inventory–2: The BFI-2-S and BFI-2-XS.” *Journal of Research in Personality* 68 (C):69–81.
- Stajkovic, AD and F Luthans. 1997. “A meta-analysis of the effects of organizational behavior modification on task performance, 1975-95.” *Academy Of Management Journal* 40 (5):1122–1149.
- Tanguy, B, Stefan Dercon, K Orkin, and AS Taffesse. 2014. “The future in mind: Aspirations and forward-looking behaviour in rural Ethiopia.” .
- Titmuss, Richard Morris. 1970. *The gift relationship : from human blood to social policy*. New York, NY: Vintage Books.

# Appendix

## Appendix A: Tables

A. 1: Psychological Traits and Workplace Outcomes, Punjab Survey

	(1)	(2)	(3)
	Any Job Satisfaction	Job Commitment	Ranked Non-financial higher
Intrinsic Motivation	0.234*** (0.020)	0.241*** (0.020)	0.015 (0.010)
Extrinsic Motivation	-0.019 (0.018)	-0.041** (0.018)	-0.026** (0.010)
Prosocial Motivation	0.219*** (0.051)	0.146*** (0.057)	-0.021 (0.029)
Personality Index	0.255*** (0.029)	0.279*** (0.030)	-0.013 (0.015)
Observations	8419	8419	8419
Adjusted R <sup>2</sup>	0.104	0.086	0.002
PDS LASSO controls	Yes	Yes	Yes
School Fixed Effects	Yes	Yes	Yes

Notes: Each column is a regression of a different dependent variable on a set of teacher controls and non-cognitive traits of interest. Teacher controls include: teacher age, gender, academic qualification, marital status, salary. Each regression includes school fixed effects and errors are clustered at the school level. Intrinsic and extrinsic motivational orientations are measured via the workplace motivation scale used and tested by Gagne et al (2001) based on the Self Determination Theory of Deci and Ryan (1975). Pro-social motivation is measured using the public service motivation (PSM) scale by Perry (1996, 1997). The Personality Index calculated through the short 10-point BFI inventory. The index is calculated by estimating the z-score of each personality dimension and averaging across all dimensions as in Callen et al (2015). Estimates are significant at the \*10%, \*\*5%, and \*\*\*1% level

## A. 2: Descriptive Statistics

	(1)	(2)
	Mean	Sd
<b>Basic teacher characteristics</b>		
Age	34.97	7.88
Gender (=1 if male)	0.55	
Years of Experience	8.69	6.95
Years of Education	15.90	1.28
Married (=1 if married)	0.73	
<b>Baseline Score</b>		
Pre Test Scores (normalised)	0.11	0.92
Pre Test Scores (Percent)	0.58	0.13
<b>Non-Cognitive Traits</b>		
<i>Personality Traits</i>		
Overall Index	0.37	1.03
Openness	0.03	0.84
Extraversion	0.01	1.02
Conscientiousness	-0.02	1.16
Agreeableness	-0.02	0.95
Neuroticism	-0.01	0.83
<i>Motivation Type Proportions</i>		
Intrinsic Motivation	0.38	0.49
Extrinsic Motivation	0.33	0.47
Pro-social Motivation	0.28	0.45
Observations	650	

### A. 3: Randomization Balance - Pooled Treatment

	(1)	(2)	(3)
	Control	Treatment	P-value difference
<b>Basic teacher characteristics</b>			
Age	35.66 (0.93)	34.68 (0.72)	0.23
Gender (=1 if male)	0.55 (0.93)	0.55 (0.72)	0.91
Salary	39411 (1178)	38591 (705)	0.51
Years of Experience	9.04 (0.76)	8.55 (0.59)	0.52
Years of Education	15.84 (0.12)	15.93 (0.05)	0.51
Married (=1 if married)	0.77 (0.03)	0.71 (0.03)	0.14
<b>Baseline Score</b>			
Pre Test Scores	0.06 (0.10)	0.13 (0.08)	0.37
<b>Non-Cognitive Traits</b>			
<i>Personality Traits</i>			
Openness	0.05 (0.04)	0.02 (0.03)	0.43
Extraversion	-0.17 (0.12)	0.08 (0.04)	0.05**
Conscientiousness	-0.03 (0.09)	-0.03 (0.05)	0.90
Agreeableness	0.04 (0.02)	-0.05 (0.05)	0.14
Neuroticism	0.04 (0.02)	-0.03 (0.04)	0.16
<i>Motivation Type Proportions</i>			
Intrinsic Motivation	0.39 (0.03)	0.38 (0.03)	0.95
Extrinsic Motivation	0.33 (0.03)	0.33 (0.02)	0.98
Pro-social Motivation	0.28 (0.03)	0.28 (0.02)	0.87
Joint F-Test			0.98
Observations	192	458	

Notes: The first two columns report the mean and standard deviation in the treatment and control groups, with the third column showing equality for each variable. Estimates are significant at the \*10%, \*\*5%, and \*\*\*1% level

#### A. 4: Randomization Balance - All Treatments

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Control	Private	Peer	Career	C-Pvt	C-Peer	C-Career
<b>Basic teacher characteristics</b>							
Age	35.66 (0.93)	34.74 (0.69)	35.08 (0.93)	34.21 (0.92)	0.30	0.53	0.16
Gender (=1 if male)	0.55 (0.04)	0.50 (0.05)	0.55 (0.04)	0.58 (0.05)	0.34	0.96	0.55
Salary	39,338 (1244)	39,132 (1361)	40,101 (1389)	36,301 (1031)	0.91	0.68	0.06*
Years of Experience	9.04 (0.79)	8.54 (0.68)	8.79 (0.67)	8.30 (0.84)	0.56	0.77	0.42
Years of Education	15.84 (0.12)	16.02 (0.02)	15.92 (0.10)	15.88 (0.10)	0.16	0.64	0.85
Married (=1 if married)	0.77 (0.03)	0.71 (0.04)	0.70 (0.04)	0.73 (0.05)	0.32	0.16	0.43
<b>Baseline Score</b>							
Pre Test Scores	0.06 (0.10)	0.14 (0.10)	0.08 (0.10)	0.17 (0.09)	0.49	0.88	0.12
<b>Non-Cognitive Traits</b>							
<i>Personality Traits</i>							
Openness	0.06 (0.03)	0.06 (0.03)	0.04 (0.03)	-0.05 (0.11)	0.94	0.69	0.35
Extraversion	-0.17 (0.12)	0.09 (0.05)	0.09 (0.09)	0.06 (0.02)	0.05**	0.09*	0.06*
Conscientiousness	-0.03 (0.08)	-0.06 (0.13)	-0.07 (0.09)	0.07 (0.02)	0.87	0.80	0.21
Agreeableness	0.04 (0.02)	-0.10 (0.13)	-0.09 (0.10)	0.03 (0.01)	0.30	0.21	0.71
Neuroticism	0.04 (0.02)	0.04 (0.02)	-0.04 (0.08)	-0.08 (0.09)	0.94	0.38	0.23
<i>Motivation Type Proportions</i>							
Intrinsic Motivation	0.39 (0.03)	0.39 (0.05)	0.33 (0.04)	0.43 (0.04)	0.97	0.22	0.32
Extrinsic Motivation	0.33 (0.03)	0.33 (0.04)	0.37 (0.04)	0.29 (0.04)	0.87	0.39	0.31
Pro-social Motivation	0.28 (0.03)	0.27 (0.04)	0.29 (0.04)	0.27 (0.03)	0.84	0.76	0.84
Joint F-Test					0.93	0.95	0.52
Observations	192	129	169	160			

Notes: The first four columns report the mean and standard errors of the three recognition treatments and the control group. The last three columns show equality of means between the control group and the treatment group for each variable. Estimates are significant at the \*10%, \*\*5%, and \*\*\*1% level

A. 5: Average Treatment Effects on Certificate Valued for Extrinsic Reasons

	(1)	(2)	(3)	(4)
Any Treatment	-0.099*** (0.037)	-0.099*** (0.038)		
Private			-0.078 (0.052)	-0.078 (0.052)
Peer			-0.137*** (0.046)	-0.144*** (0.047)
Career			-0.075* (0.044)	-0.070 (0.043)
Observations	645	626	645	626
Adjusted R <sup>2</sup>	0.008	0.007	0.008	0.007
PDS LASSO controls	No	Yes	No	Yes
Training Day Fixed Effects	Yes	Yes	Yes	Yes

Notes: All regressions include training day fixed-effects with errors clustered at the group level at which the treatment is administered. Standard errors are presented in parenthesis. The dependent variable is a dummy variable equals to 1 if the trainee reported extrinsic value from a certificate such as self confidence or self-satisfaction. Estimates are significant at the \*10%, \*\*5%, and \*\*\*1% level

A. 6: Heterogeneous Treatment Effects by Motivation Type, Certificate Internal Value

	Any Treat		Private		Peer		Career	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Intrinsic Motivation	0.028 (0.273)	0.021 (0.275)	0.571*** (0.066)	0.570*** (0.066)	0.460** (0.199)	0.451** (0.200)	-0.017 (0.235)	-0.015 (0.231)
Extrinsic Motivation	0.008 (0.282)	0.005 (0.284)	0.558*** (0.068)	0.553*** (0.067)	0.417** (0.202)	0.414** (0.202)	-0.031 (0.245)	-0.026 (0.242)
Prosocial Motivation	0.036 (0.279)	0.026 (0.282)	0.576*** (0.077)	0.570*** (0.082)	0.468** (0.206)	0.458** (0.208)	-0.046 (0.245)	-0.044 (0.243)
Intrinsic Motivation * Treat	-0.028 (0.055)	-0.022 (0.054)	0.060 (0.076)	0.057 (0.077)	-0.010 (0.079)	-0.004 (0.077)	-0.106 (0.065)	-0.102 (0.065)
Extrinsic Motivation * Treat	0.043 (0.072)	0.045 (0.072)	-0.079 (0.087)	-0.075 (0.088)	0.201** (0.087)	0.201** (0.088)	-0.060 (0.084)	-0.060 (0.085)
Prosocial Motivation * Treat	0.152** (0.070)	0.164** (0.072)	0.119 (0.095)	0.124 (0.099)	0.135 (0.088)	0.144 (0.090)	0.193** (0.093)	0.203** (0.095)
P-Value Difference								
Intrinsic*Treat - Extrinsic*Treat	0.46	0.49	0.25	0.28	0.12	0.13	0.70	0.72
Observations	644	644	319	319	358	358	347	347
Adjusted R <sup>2</sup>	0.010	0.007	0.000	-0.012	0.012	0.007	0.010	0.001
PDS LASSO controls	No	Yes	No	Yes	No	Yes	No	Yes
Training Day Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: All regressions include training day fixed-effects with errors clustered at the group level at which the treatment is administered. Standard errors are presented in parenthesis. The dependent variable is a dummy variable equals to 1 if the trainee reported intrinsic value from a certificate such as self confidence or self-satisfaction. Estimates are significant at the \*10%, \*\*5%, and \*\*\*1% level

A. 7: Average Treatment Effects by Ability Distribution

	Full Sample		Extrinsically Motivated		Intrinsically Motivated	
	(1)	(2)	(3)	(4)	(5)	(6)
Pre Test Score * Treat	-0.040 (0.101)		0.057 (0.150)		0.005 (0.129)	
1st Quartile * Treat		0.145 (0.115)		0.036 (0.222)		0.205 (0.169)
2nd Quartile * Treat		-0.241** (0.093)		-0.097 (0.190)		-0.194* (0.102)
3rd Quartile * Treat		0.219 (0.177)		-0.537* (0.320)		0.623** (0.280)
4th Quartile * Treat		-0.031 (0.110)		-0.327 (0.227)		0.076 (0.166)
Observations	650	649	216	216	249	249
Adjusted R <sup>2</sup>	0.069	0.089	0.064	0.091	0.089	0.121
PDS LASSO controls	No	No	No	No	No	No
Training Day Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes

Notes: All regressions include training day fixed-effects with errors clustered at the group level at which the treatment is administered. Standard errors are presented in parenthesis. Estimates are significant at the \*10%, \*\*5%, and \*\*\*1% level

A. 8: Average Treatment Effects on additional Non-Cognitive Traits

	(1)	(2)	(3)
	Intrinsic Mot	Self-efficacy	External Locus
Private	0.116 (0.093)	0.231* (0.122)	-0.079 (0.066)
Peer	-0.128 (0.098)	-0.125 (0.080)	0.120 (0.135)
Career	-0.024 (0.088)	-0.075 (0.086)	-0.018 (0.094)
Observations	643	631	631
Adjusted R <sup>2</sup>	0.003	0.011	-0.000
PDS LASSO controls	No	No	No
Training Day Fixed Effects	Yes	Yes	Yes

Notes: All regressions include training day fixed-effects with errors clustered at the group level at which the treatment is administered. Standard errors are presented in parenthesis. Estimates are significant at the \*10%, \*\*5%, and \*\*\*1% level

A. 9: Heterogeneous Treatment Effects on Non-cognitive traits

	Self-efficacy			External Locus			Intrinsic Motivation		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Private	Peer	Career	Private	Peer	Career	Private	Peer	Career
Intrinsic Motivation	0.884*** (0.178)	1.073*** (0.257)	0.104 (0.655)	-0.243*** (0.086)	-0.357** (0.161)	0.079 (0.196)	0.988*** (0.344)	1.122*** (0.275)	-0.201 (0.450)
Extrinsic Motivation	1.125*** (0.167)	1.342*** (0.249)	0.385 (0.664)	-0.384*** (0.051)	-0.412*** (0.083)	0.004 (0.170)	1.001*** (0.331)	1.152*** (0.262)	-0.177 (0.459)
Prosocial Motivation	1.146*** (0.195)	1.291*** (0.264)	0.346 (0.664)	-0.382*** (0.076)	-0.535*** (0.197)	-0.005 (0.176)	0.731** (0.348)	0.837*** (0.277)	-0.480 (0.473)
Intrinsic Motivation * Treat	0.715* (0.361)	0.105 (0.137)	0.192 (0.126)	-0.169 (0.116)	0.071 (0.211)	0.018 (0.171)	0.193 (0.161)	-0.245 (0.181)	-0.186 (0.157)
Extrinsic Motivation * Treat	-0.291 (0.175)	-0.370*** (0.101)	-0.152 (0.118)	-0.048 (0.051)	0.290 (0.202)	-0.034 (0.043)	-0.083 (0.226)	-0.311* (0.179)	-0.167 (0.201)
Prosocial Motivation * Treat	0.115 (0.217)	-0.066 (0.153)	-0.263 (0.180)	-0.032 (0.076)	0.030 (0.129)	-0.021 (0.067)	0.188 (0.194)	0.203 (0.195)	0.495*** (0.171)
Observations	311	349	338	311	349	338	318	356	346
Adjusted R <sup>2</sup>	0.024	0.021	0.004	-0.000	-0.007	-0.013	0.005	0.004	0.009
PDS LASSO controls	No	No	No	No	No	No	No	No	No
Training Day Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: All regressions include training day fixed-effects with errors clustered at the group level at which the treatment is administered. Standard errors are presented in parenthesis. Estimates are significant at the \*10%, \*\*5%, and \*\*\*1% level

## Appendix B: Instruments for Non-Cognitive Traits

The non-cognitive traits which serve as our main moderators for analysing heterogeneous treatment effects include motivational orientation and personality traits.

In this section, we provide additional details on the scales used for measurement.

### *Motivational Orientation:*

Rasul and Rogger (2015) capture pro-social motivation for civil servants in Nigeria through the following question: “What most influenced you to take up a career in the service?” The proportion of staff that answered “The chance to serve Nigeria” serves as their intrinsic motivation measure. Based on the psychology literature, we post-code the responses to the same question for public school teachers in our experiment by mapping each response to either intrinsic motivation derived from the interest/enjoyment in the job, pro-social motivation for serving the country or community, or extrinsic motivation derived from salary or other external benefits.

To map these responses into appropriate categories, we refer to the motivational orientation at workplace scale tested by Gagne et al (2010) and Amiable et al (2004). Their items which measure intrinsic and extrinsic reasons for working are given below:

Intrinsic items:

1. Because I enjoy this work very much
2. Because I have fun doing my job
3. For the moments of pleasure that this job brings me

Extrinsic items:

1. Because this job affords me a certain standard of living
2. Because it allows me to make a lot of money do this job for the paycheck

*Personality Measure:*

Personality traits are measured using the short Big Five Inventory as given below. Each statement in the question below is ranked on a likert scale of five ranging from strongly agree to strongly disagree.

I see myself as someone who..

1 is reserved

2 is generally trusting

3 tends to be lazy

4 is relaxed, handles stress well

5 has few artistic interests

6 is outgoing, sociable

7 tends to find fault with others

8 does a thorough job

9 gets nervous easily

10 has an active imagination

Each statement maps onto the five different dimensions of personality as follows:<sup>17</sup>

Extraversion: 1R, 6

Openness: 5R, 10

Agreeableness: 2, 7R

Conscientiousness: 3R, 8

Neuroticism: 4R, 9

---

<sup>17</sup>R stands for items that are reverse scored

## Appendix C: Sample Test Questions

Quaid-e- Azam Academy for Educational Development, Punjab – Professional Development Day

### Topic 1: Unitary Method (Concept of unitary method, direct and inverse proportion)

Q1. If cost of one notebook is Rs. 3200, then cost of 12 such notebooks is:

اگر ایک نوٹ بک کی قیمت 3200 روپے ہو تو اس طرح کی 12 نوٹ بکس کی قیمت ہے

A. Rs.36000 <input type="checkbox"/>	B. Rs.3200 <input type="checkbox"/>	C. Rs. 38400 <input type="checkbox"/>	D. Rs.38500 <input type="checkbox"/>	E. Don't know <input type="checkbox"/>
---	--	--	---	---

### SECTION- A

Q2. In 2:6 : x: 3 , the value of "x" is:

2:6 : x: 3 میں، "x" کی قیمت ہے

A. 1 <input type="checkbox"/>	B. 2 <input type="checkbox"/>	C. 5 <input type="checkbox"/>	D. 6 <input type="checkbox"/>	E. Don't Know <input type="checkbox"/>
----------------------------------	----------------------------------	----------------------------------	----------------------------------	---

Q3. If the price of 15 locks is Rs. 1275 then the price of three such lock will be:

A. Rs.71 <input type="checkbox"/>	B. Rs.85 <input type="checkbox"/>	C. Rs.255 <input type="checkbox"/>	D. Rs.200 <input type="checkbox"/>	E. Don't Know <input type="checkbox"/>
--------------------------------------	--------------------------------------	---------------------------------------	---------------------------------------	---

اگر 15 تلوں کی قیمت 1275 روپے ہے۔ پھر اس طرح کے 3 تلوں کی قیمت ہوگی

### Topic 2: Angles, Triangles, Quadrilateral

Q1. What type of triangle is  $\Delta ABC$ , where  $AB = BC = 11$  cm? (Note: cm denotes centimeter.)

AB = BC = 11 cm جبکہ، کس قسم کی مثلث ہے،

A. equilateral triangle مساوی الاضلاع مثلث <input type="checkbox"/>	B. isosceles triangle متمائل ساقین مثلث <input type="checkbox"/>	C. Obtuse angled triangle منفرجہ زاویہ مثلث <input type="checkbox"/>	D. scalene triangle مختلف زاویہ مثلث <input type="checkbox"/>	E. Don't Know <input type="checkbox"/>
--	--	--	---	---

Q2. It's 3 'o clock on watch. Which angle is made by hands of the watch?

گھڑی پر 2 بجے ہیں۔ گھڑی کی سویوں کے درمیان کیا زاویہ ہے؟

A. Right <input type="checkbox"/>	B. Reflex <input type="checkbox"/>	C. Straight <input type="checkbox"/>	D. Acute <input type="checkbox"/>	E. Don't Know <input type="checkbox"/>
--------------------------------------	---------------------------------------	---	--------------------------------------	---

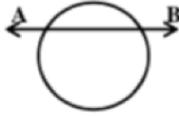
Q3. A triangle has two angles measuring  $41^\circ$  and one angle measuring  $98^\circ$ . What type of triangle is this?

ایک مثلث کے دو زاویوں کی پیمائش  $41^\circ$  ڈگری ہیں اور تیسرے زاویے کی پیمائش  $98^\circ$  ہے۔ یہ کس قسم کی مثلث ہے

A. equilateral triangle مساوی الاضلاع <input type="checkbox"/>	B. isosceles triangle مساوی الساقین مثلث <input type="checkbox"/>	C. Obtuse angled triangle منفرجہ زاویہ مثلث <input type="checkbox"/>	D. scalene triangle مختلف زاویہ مثلث <input type="checkbox"/>	E. Don't Know <input type="checkbox"/>
--	---	--	---	---

**Topic 3: Geometry (Angles, polygons, circle, square)**

Q1. In the given figure,  $\overleftrightarrow{AB}$  is:



$\overleftrightarrow{AB}$  شکل میں ہے

A. Chord وتر <input type="checkbox"/>	B. Segment قطع خط <input type="checkbox"/>	C. Tangent ماس <input type="checkbox"/>	D. Secant خط قاطع <input type="checkbox"/>	E. Don't Know <input type="checkbox"/>
---	--	---	--	---

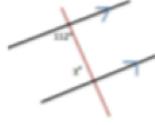
Q2. The measure of each interior angle of a regular polygon is  $135^\circ$ , then number of sides in the regular polygon is:

ایک باقاعدہ کثیرالاضلاع کا ہر زاویہ  $135^\circ$  کا ہو تو اس کے اضلاع کی تعداد ہو گی،

A. 5 <input type="checkbox"/>	B. 6 <input type="checkbox"/>	C. 7 <input type="checkbox"/>	D. 8 <input type="checkbox"/>	E. Don't Know <input type="checkbox"/>
----------------------------------	----------------------------------	----------------------------------	----------------------------------	---

Q3. The measure of angle  $y$  in the given figure is:

دی گئی شکل میں زاویہ  $y$  کی پیمائش ہے



A. $112^\circ$ <input type="checkbox"/>	B. $58^\circ$ <input type="checkbox"/>	C. $68^\circ$ <input type="checkbox"/>	D. $75^\circ$ <input type="checkbox"/>	E. Don't Know <input type="checkbox"/>
---	--	--	--	--

### SECTION - B

#### Topic 1: Unitary Method (Concept of unitary method, direct and inverse proportion)

Q 1. Amina reads 40 pages of a book in 4 days. If the book has 60 pages, then in how many days Amina will read the complete book?

آمینہ ایک کتاب کے 4 دنوں میں 40 صفحات پڑھتی ہے۔ اگر کتاب کے 60 صفحات ہیں تو پھر آمینہ کتنے دنوں میں پوری کتاب پڑھے گی؟

A. 2 <input type="checkbox"/>	B. 6 <input type="checkbox"/>	C. 10 <input type="checkbox"/>	D. 15 <input type="checkbox"/>	E. Don't Know <input type="checkbox"/>
-------------------------------	-------------------------------	--------------------------------	--------------------------------	--

#### Topic 2: Angles, Triangles, Quadrilateral

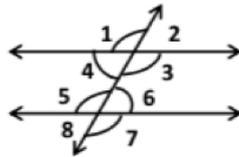
Q2. The pair of complementary angles is:

کمپلیمنٹری زاویوں کا جوڑا ہے:

A. $35^\circ, 55^\circ$ <input type="checkbox"/>	B. $130^\circ, 50^\circ$ <input type="checkbox"/>	C. $110^\circ, 70^\circ$ <input type="checkbox"/>	D. $130^\circ, 40^\circ$ <input type="checkbox"/>	E. Don't Know <input type="checkbox"/>
--	---	---	---	--

#### Topic 3: Geometry (Angles, polygons, circle, square)

Q3. In the figure, a pair of complementary angles is:



دی گئی شکل میں complementary زاویوں کا جوڑا ہے

A. $\angle 3$ & $\angle 7$ <input type="checkbox"/>	B. $\angle 3$ & $\angle 6$ <input type="checkbox"/>	C. $\angle 4$ & $\angle 6$ <input type="checkbox"/>	D. $\angle 4$ & $\angle 8$ <input type="checkbox"/>	E. Don't Know <input type="checkbox"/>
---	---	---	---	--

End time:  :  :

## Appendix D: Alternative Explanations

We also explore whether our (tournament-like) recognition schemes threaten to challenge individual beliefs about ability, resulting in lower effort on the part of the individual (as observed empirically in Ashraf et al 2013). This is consistent with the literature on belief utility and information avoidance and is predicted to be more likely at play for individuals with low ability (Bénabou and Tirole 2003, Köszegi 2002). In our particular setting, we are interested in exploring whether the negative effects for extrinsic individuals are primarily coming from low ability individuals.

To explore this hypothesis, we construct quartiles of teacher ability using the pretest score as a proxy for ability, and study the treatment response across the distribution of ability.

Table A.7 shows that we do not find consistent evidence of a negative treatment response at the bottom of the ability distribution. For the full sample, column 2 shows that the treatment coefficient is positive at the bottom quartile of the distribution (although not significant), negative in the second (and significant at the 1% level), but becomes positive again for the third quartile. Column 4 shows that low ability extrinsic individuals do not particularly respond negatively to the treatment, but instead the negative treatment response comes from across the distribution. Similarly, the intrinsically motivated individuals seem to respond positively across the distribution (see columns 5–6).

These results do not show any evidence for the belief utility and information avoidance hypothesis.

# 2

## The Double-Edged Sword of Non-Financial Incentives (Part II): Evidence from the Education Sector in Pakistan

# The Double-Edged Sword of Non-Financial Incentives (Main Experiment- Part II): Evidence from the Education Sector in Pakistan

Zahra Mansoor \*

October 2020

## Abstract

I present experimental evidence on the impact of employer recognition, a type of non-financial incentive, on the performance of head teachers in an in-service training programme conducted by the Teacher Training Academy in Punjab Pakistan. I randomize 3,394 head teachers attending this mandatory government training into 4 different designs of recognition incentives that are tied to training performance, and a control group. The first recognition incentive makes peer/collegial approval and peer-esteem salient; the second makes potential career benefits of recognition salient, and the third and fourth cross the first two treatments with a self-efficacy enhancing frame. I find that employer recognition can improve teacher training performance if it is linked to tangible career benefits in the future but the positive effects can also backfire depending on how these incentives are framed. When a self-efficacy frame is added to either of the first two treatments, it improves teacher self-efficacy but makes teachers overconfident, reducing their performance in the training. These findings have implications for how to design more powerful non-financial incentives that are effective for eliciting higher teacher effort, specifically in trainings and more broadly for teacher tasks at the school level. At the same time, they point towards the sensitivity of such incentives to framing effects and suggest caution in how these are framed.

**Acknowledgements:** This research was made possible through numerous insights from stakeholders at QAED. In particular, Mr. Waseem Shirazi, Mr. Nadeem Hussain, and Mr. Adnan Bashier. I gratefully acknowledge funding from Economic Development Institutions (EDI), Center for the Study of African Economis (CSAE), and the Blavatnik School of Government (BSG). Usman Zahid offered excellent research support.

\* Dphil Candidate, Blavatnik school of Government

# 1 Introduction

A motivated and high-performing bureaucracy is central to improving public service delivery, economic growth, and development. This is especially important in the education sector where developing countries continue to report low levels of student learning outcomes (Glewwe and Muralidharan, 2016). Public school teachers, who are the frontline agents in the education sector, are central to improving student learning outcomes (Das et al., 2012; Chetty et al., 2014). However, across several developing country contexts, teacher quality and effort remain a critical concern. For example, a study across six countries in Sub-Saharan Africa shows that nearly 40 percent of primary school teachers are not as knowledgeable as their students should be. In addition to these gaps in quality, teacher absenteeism and low time on task leads to loss of instructional time with significant implications for student learning losses (Bold et al., 2017).

To address these challenges around teacher quality and effort, the education literature to date has largely focused on teacher recruitment, teacher characteristics that can best predict increases in student test scores, and teacher compensation and incentives (see Rockoff, 2004; Rivkin et al., 2005; Muralidharan and Sundararaman, 2011; Duflo et al., 2012; de Ree et al., 2018; Bau and Das, 2020). However, two key gaps remain in this literature. First, there is limited evidence on how public school teachers can learn more effectively on the job, the main vehicle for which is teacher in-service trainings. This is evident in the dearth of studies on the effectiveness of teacher in-service trainings or on how to motivate teachers to learn more effectively in such trainings (Popova et al., 2016). Second, the literature on incentives largely focuses on financial incentives which, though effective when designed well, can often be distortionary via creating perverse incentives or crowding out intrinsic motivation (Holmstrom and Milgrom, 1991; Glewwe et al., 2010; Bénabou and Tirole, 2003). Non-financial incentives, on the other hand, could be effective in pro-social settings like education where teachers may put a lower weight on financial incentives (Besley and Ghatak, 2005 ) but evidence on their effectiveness or the channels through which they operate remains limited in the public sector with the exception of a few studies (see

Ashraf, Bandiera, and Lee, 2014; Ashraf, Bandiera, and Jack, 2014; Gauri et al., 2018).<sup>1</sup>

We address these gaps in the literature by studying the impact of non-financial incentives on teacher performance in in-service trainings. We specifically study the impact of employer recognition and ask the following two questions. *First*, could non-financial incentives such as employer recognition (tied to training test scores) improve teacher effort during training, and result in improved teacher knowledge as measured at the end of training? If yes, what is the impact of employer recognition that makes peer/collegial approval salient versus recognition that makes future tangible career benefits salient?<sup>2</sup> While existing evidence documents the impact of public versus private forms of employer recognition (as in Ashraf, Bandiera, and Lee, 2014) or the impact of employer versus community recognition (as in Gauri et al., 2018), no existing study looks at these two main channels of employer recognition and how the strength of each varies from the other. *Second*, does framing these recognition incentives with a self-efficacy enhancing frame improve their effectiveness? Existing evidence highlights that incentives (particularly those that are tournament-based) can often create negative effects by damaging employee morale, self-efficacy, and self-esteem (see Lazear and Rosen, 1981; Bénabou and Tirole, 2003; Ashraf, Bandiera, and Lee, 2014; Mansoor, 2019). To the best of our knowledge, our study is the first that tries to exogenously vary teacher self-efficacy through self-efficacy enhancing frames and offers this combination with incentives to improve the design of incentive schemes.

To shed light on these questions, I design and implement a field experiment in collaboration with the Teacher Training Academy in Punjab, Pakistan called the Quaid-e-Azam Academy for Educational Development (QAED). Punjab is Pakistan’s most populous province. It employs a workforce of approximately 400,000 teachers spread across 52,000 schools.<sup>3</sup> QAED holds the mandate to offer in-service trainings to all public school teachers in Punjab to improve their

---

<sup>1</sup>A significant literature has looked at the impact of non-financial incentives in the private sector (see Frey and Neckermann, 2008; Markham et al., 2002; Luthans and Stajkovic, 1999; Kosfeld and Neckermann, 2011; Kosfeld, Neckermann, and Yang, 2014).

<sup>2</sup>See Besley and Ghatak (2008) and Frey (1997) for a discussion on why agents may value peer/collegial approval and social distinction. See Dewatripont et al. (1999) and Ashraf, Bandiera, and Lee (2014) for how agents may use recognition to signal performance to supervisors as a way to gain potential career benefits.

<sup>3</sup>Annual School Census Data 2017.

skills and quality on the job. However, teacher quality remains a challenge and is widely seen as one of the main explanations for low student learning outcomes in the province. Our baseline sample from Mansoor (2019) shows that more than 40% of teachers score less than 50% on grade 5 math problems, which highlights the scope for improving teacher knowledge (and quality) through improving teacher engagement in trainings. However, despite yearly investments in various professional development trainings, QAED does not employ any systematic strategies or tools to improve teacher engagement in these trainings. Given the scale of in-service trainings offered by QAED, understanding the effectiveness of non-financial incentives in this setting offers significant potential.

As part of this experiment, QAED randomly allocated training sessions for an in-service training on school management and leadership offered to 3,394 head teachers to four different employer recognition schemes and a control group. Treatment 1 (Peer arm) makes peer/collegial approval and peer-esteem salient – trainees are told that those with the top score in the training post-test and the most improved score (over the pre-test) will be provided certificates in a district level ceremony which will be attended by their peers and colleagues in their district; Treatment 2 (Career arm) offers the same certificate as Treatment 1 but makes career benefits salient instead. Trainees are told that those who qualify will receive the certificates privately but at the same time their name will be added to an ‘excellent teacher list’ (which will be shared with the School Education Department leadership) that can help them qualify for future career opportunities in their districts or the department. While formal career incentives such as promotions are purely based on seniority in our setting, this treatment leverages the availability of informal career incentives in the system such as postings to preferred schools, transfers to other lateral postings, or getting selected for promotion earlier once eligible; Treatments 3 and 4 (Public PLUS and Career PLUS) cross the first two treatments with a self-efficacy enhancing frame that aims to bolster teacher perceptions in their ability to do well in the training and their jobs more broadly.

There are three key advantages to our setting. First, we are able to incentivize training test scores which are one-dimensional in effort and easier to measure as opposed to measuring teacher performance in the classroom. This allows us to avoid the standard issues of multi-tasking (as

raised by Holmstrom and Milgrom, 1991). Second, the training is offered to head teachers on school management and leadership which creates a direct link between improving head teacher performance in the training and their performance in their schools. This allows us to study the downstream effects of our incentives in the training on school-level outcomes. Third, our setting allows us to observe teacher preferences for these rewards in a cost-effective and contained, but real-world setting. This allows us to draw implications of such recognition incentives, although with limitations of external validity, to teacher performance more broadly.

We report five main sets of results based on the experiment. First, we find that employer recognition can work when it is linked to potential career benefits in the future. We find indicative evidence that the Career arm leads to a  $0.26\sigma$  increase in training test scores as compared to the control group. In comparison, the Peer arm has a coefficient of 0.03 and is insignificant. Quantile treatment effects show that the Career arm has a positive coefficient in the range of 0.27-0.35 across the distribution. In the upper tail of the distribution, these coefficients are significant and also significantly different from the Peer arm. The latter is also confirmed by the Kolmogorov–Smirnov test of equality of distribution between the Career and Peer arm, showing that the distribution of post training test scores for these two arms is significantly different at the 1% level of significance.

Second, we find that the positive effects of such incentives can backfire depending on how they are framed. The net impact of adding the self-efficacy frame in Career PLUS is negative and significant – a reduction of  $0.28\sigma$  in training test scores. The net impact of the frame in Peer PLUS is also negative but insignificant. Kolmogorov–Smirnov tests of equality of distribution between the Peer and Career arm and their PLUS counterparts show that the distributions are significantly different from each other at the 1% level of significance. Overall, the net impact of adding the self-efficacy frame to both the Peer and Career arms is to lower training test scores by  $0.21\sigma$ .

Third, when the Career arm works, it works through individuals who are due for their next promotion sooner or have higher visibility to their leadership such as Secretary Education and District Education Officers. While an upcoming promotion would create direct incentives to get

promoted to a position of choice, higher visibility to senior leadership would increase opportunities to benefit from various types of informal career incentives in the system. Given the Career arm makes future tangible career benefits salient, our results lend support to the career-benefits channel of recognition. We also find meaningful heterogeneity along several other dimensions - the treatment effect on the Career arm is significant for women which highlights that females have a demand for opportunities that allow them to access informal career incentives. This is consistent with existing work which shows that males in the Pakistani bureaucracy are more embedded in informal networks that provide access to informal career benefits as compared to women (Tanwir, 2014). We also find that the treatment effect on the Career arm is stronger for teachers who score higher on the Big Five Personality and Perry (1996) pro-social motivation (PSM) index which is inline with earlier work that highlights the importance of such non-cognitive traits in how individuals respond to incentives (Almlund et al., 2011; Callen et al., 2016; Lee, 2018).

Fourth, treatment effects on our main secondary outcome - teacher self-efficacy - are insignificant for the Peer and Career arms but positive and significant for the Peer PLUS and Career PLUS arms, with an increase of  $0.11\sigma$  and  $0.14\sigma$  respectively. This highlights that despite negative effects on training test scores, the PLUS treatments do bolster teacher self-efficacy as expected. Our main hypothesis for explaining the negative effects of the PLUS treatments on training test scores explores if the frames made teachers' over confident which resulted in lower effort (as argued by Baumeister, 1999; and Swann, 1996). We construct a direct measure of overconfidence as the difference between what teachers' believed they scored on the post-test at endline and what they actually scored. We find that the net impact of the frames is to increase teacher overconfidence by 5-6 percentage points for the Peer PLUS and Career PLUS arms. Mediation analysis (following guidelines as per Acharya et al., 2016) shows that overconfidence can explain up to 50% of the observed negative effects of the PLUS treatments on training test scores.

Finally, we investigate the impact of our treatments on a set of downstream outcomes at the school level. Since the implementation of training knowledge to actual practices can be challenging, especially without top-down monitoring or support structures (see Banerjee et al., 2016), we

are interested in exploring whether incentives in training can support the implementation and mainstreaming of trainings into actual practices. Using government administrative data from monthly school monitoring visits, we observe treatment effects on school-level outcomes that are two steps down the causal chain of how these trainings are intended to effect school performance. We find no significant treatment effects of our incentives on % teachers absent, % classrooms used, % student attendance, number of school council meetings held, and the number of additional coaches hired for teaching. Given we do not have access to relevant data for outcomes at the first step of the causal chain (such as direct implementation of training lessons by the head teachers to their own management practices in schools), clear conclusions on the effectiveness of our incentives on downstream outcomes are hard to draw. However, these results (although null) reiterate the need for further research that collects process-based information and data at various steps of the causal chain from the point when trainings are delivered to school-level outcomes. This would enable answering the question of how further along the causal chain can incentives support the implementation (and mainstreaming) of training lessons to actual practices.

Our results have direct relevance for policy discussions on how to bolster teacher engagement, effort, and subsequently the acquisition of knowledge and skills in in-service trainings. Existing evidence, though limited, indicates that targeted instruction, provision of learning material alongside training, and linking teacher participation in training to incentives such as promotion or salary implications could be effective ways of improving the impact of in-service trainings (Popova et al., 2016). Our results indicate that non-financial incentives that are linked to tangible career benefits in the future could also be a cost-effective way for improving the impact of in-service trainings, where some of these career benefits such as postings to preferred schools or transfers to other lateral positions may be less formal and easier to exercise in the system as compared to promotions. At the same time, our results indicate caution in how such incentives are framed and designed for use in teacher trainings. While a considerable literature in education research demonstrates the positive impacts of teacher self-efficacy on teacher psychological well-being and practices (see synthesis by Zee and Koomen, 2016), these studies do not exogenously vary teacher self-efficacy. Our study is the first to exogenously vary teacher self-efficacy and show the pitfalls of such interventions via over-correcting teacher beliefs about ability. More

broadly, this points to the need to understand the degree to which teachers may have unrealistic beliefs about their own ability and its implications for teacher performance and student learning outcomes.

Our results in this experiment and the pilot highlight important differences which have implications for the use of non-financial incentives in teacher trainings in particular, and for teacher performance more broadly. The pilot was conducted with primary and secondary school teachers while the main experiment was conducted with head teachers. While the peer/collegial approval channel does not work for head teachers in this experiment, it worked for primary and secondary school teachers in the pilot under some conditions (Mansoor, 2019). Our qualitative discussions highlight that head teachers may not value social approval by peers/colleagues because they have already risen through the ranks and established peer-esteem. However, this may not be true for primary and secondary school teachers who are more junior. This indicates that different types of non-financial incentives may work differently across in-service training programmes targeted at different cadres. Overall, these results highlight the highly contextual nature of these type of incentives (as in Gauri et al., 2018) and the sensitivity of such incentives to group-type and framing effects.

More broadly, our results have implications for types of non-financial incentives that may be effective in addressing a slack in teacher effort in the classroom or the school. While it is true that formal career incentives in public bureaucracies are negligible given promotions are determined by seniority, our results indicate that head teachers have career concerns through other informal mechanisms. Recognition incentives that are linked to such concerns can create an opportunity to reap these benefits (when and if the time comes). While formal incentive-based reforms can often be hard to implement, understanding the sources of different teacher motivations, informal career concerns in the system being one such source, and designing “soft” non-financial incentives around them could be one way to address part of the weakness in incentive structures.

This paper is organized as follows. Section 2 presents the theory and key hypotheses which the experimental design aims to test. Section 3 describes the context of the experimental setting. Section 4 describes the experimental design, randomization, and key data sources. Sections 5

presents the empirical strategy and main results. Section 6 concludes.

## 2 Theory and Related Literature

In this section, I present theory underlying the experimental design of the recognition schemes and the key hypotheses that are put to test.

Agents can have several motivations for employer recognition. Two of such key motivations are as follows. First, the motivation for social approval. Individuals have an innate desire to make themselves socially distinct, and this can be especially high when there are no other means of receiving distinction (Besley and Ghatak, 2008; Frey and Neckermann, 2008). Such awards can also lead to respect, reputation, and peer-esteem (Frank, 1985). Within the context of the public sector in developing countries, frontline workers such as teachers or health workers rarely have opportunities for receiving recognition socially in front of their peers and colleagues. Second, agents may want to use recognition rewards as a way to signal performance to supervisors which can in-turn result in tangible career benefits in the future (Dewatripont et al., 1999; Besley and Ghatak, 2008; Ashraf, Bandiera, and Lee, 2014). Within the context of the public sector, while promotions are primarily based on experience and seniority, bureaucrats may have informal career concerns based on opportunities such as postings to preferred locations, selection for coveted trainings, or lateral postings to influential donor programmes.

While recognition rewards can bolster performance, both economists and psychologists have argued that incentives (whether monetary or non-monetary) can often create negative effects by damaging employee morale, self-efficacy, and perceived competence. For example, Lazear and Rosen (1981) argue that incentives that work like ‘tournaments’ (which is how recognition incentives often operate) can at times decrease employee morale by creating excessive competition. Bénabou and Tirole (2003) arrive at a similar conclusion arguing that the use of incentives can convey a pessimistic signal to agents about their ability and damage self-confidence. In the psychology literature, Deci et al. (1999) argue how feelings of competence and self-determination can

be effected negatively through rewards. In relation to employee performance, Bandura (1986) highlights the importance of individual self-efficacy - i.e. beliefs about ability to achieve a given goal - as playing a central role in performance. These negative effects on individual self-esteem and self-efficacy are largely perceived by psychologists (and more recently economists) as situations when rewards/incentives crowd-out intrinsic motivation instead of bolstering it.

Our design of the recognition incentives presented in Section 4 below test the impact of the two key channels of employer recognition presented above – peer/collegial (social) approval and potential career benefits in the future – and the impact of framing these incentives with self-efficacy enhancing frames to mitigate the potential negative effects of the incentives alone.

We assume that the principal cares about maximizing the value from the training (i.e. learning gains from training) and can use two policy levers to achieve his objective - 1) recognition-based incentives that are tied to training performance; and 2) shaping agent beliefs about ability (i.e. agent self-efficacy) that can affect agent intrinsic motivation for exerting effort in the training. Agents, on the other hand, care about maximizing expected utility over learning gains from the training, value from recognition-based rewards (shaped by policy lever 1), and intrinsic value from exerting effort in the training (shaped by policy lever 2).<sup>4</sup> Our hypotheses outlined below test the following:

Hypothesis I: Employer recognition tools that leverage the peer/collegial approval channel improve training performance of teachers relative to the control group because agents care about social approval from their peers and colleagues.

Hypothesis II: Employer recognition tools that leverage the career-benefits channel improve training performance of teachers relative to the control group because agents care about signaling their performance to their supervisors for potential career benefits in the future.

---

<sup>4</sup>of course intrinsic utility derived from the training may also be non-effort dependent such as warm glow ( as argued by Andreoni, 1990). We abstract away from this as this does not determine the agent’s decision to exert effort. We restrict our attention to effort dependent intrinsic utility which may be derived as a result of pure joy of learning by exerting effort or a feeling of high self-esteem and competence by exerting effort in the training.

Hypothesis III: Employer recognition tools that leverage the peer/collegial approval channel and offer a self-efficacy enhancing frame improve training performance more than the peer channel only.

Hypothesis IV: Employer recognition tools that leverage the career benefits channel and offer a self-efficacy enhancing frame improve training performance more than the career benefits channel only.

Hypotheses I and II aim to test why agents may value recognition. We do not take a stance on which channel is more effective. While both I and II hypothesize that the treatment will lead to performance that is higher than the control condition, each theorizes a different underlying channel for why agents care about recognition. The experimental results will shed light on whether both channels are effective or if one is stronger than the other. Hypotheses III and IV test whether the self-efficacy frames and incentives work together as complements, with the frames mitigating the potential negative effects of the incentives on agent self-efficacy beliefs.

## 3 Experimental Context

### 3.1 Punjab Education Sector

Punjab, the context for this study, is Pakistan’s largest province with 36 districts and a population of 110 million.<sup>5</sup> The public education system employs a workforce of approximately 400,000 teachers responsible for educating nearly 11 million children spread across 52,000 schools.<sup>6</sup> The School Education Department (SED) is the public body which holds the mandate for all policy and implementation pertaining to primary and secondary education.<sup>7</sup>

---

<sup>5</sup>Pakistan Population Census, 2017 (Pakistan Bureau of Statistics).

<sup>6</sup>Annual School Census Data 2017.

<sup>7</sup>Schools are further divided into primary (grades 1-5), elementary (grades 6-8), secondary (grades 9-10), and higher secondary (grades 11-12) schools.

Improving education outcomes has been one of the top priorities of the Government of Punjab over the last decade. The provincial education budget has doubled in the last 7 years and a range of reforms have been implemented under the umbrella of the Punjab Education Sector Reform Programme (PESRP).<sup>8</sup> Amongst many other reforms, these included setting up a School Education ‘Reforms Roadmap’, an extensive monthly school monitoring programme (including yearly and monthly audits), and a merit-based teacher recruitment strategy (Javed and Naveed, 2019). Despite these efforts, progress in learning outcomes has remained low. The ASER (2019) report shows that nearly 40% of children in grade 5 have not reached grade 2 levels of learning in literacy and numeracy (this includes English, Math and the national language Urdu).

Low levels of teacher quality and effort is widely perceived as one of the main reasons for low levels of student learning. Anecdotal evidence indicates that while the extensive school monitoring system may have addressed part of the agency problem and reduced teacher absenteeism, it also led to an unbalanced incentive system which relied on too much monitoring and very little rewards. A pilot performance-based pay programme for teachers was launched from 2010 to 2013 but its impact evaluation showed null effects on student test scores (see Barrera-Osorio and Raju, 2017). To address questions around how to improve teacher motivation and effort, the government conducted a Teacher Motivation survey in March 2017 across 8,400 teachers in 3,100 randomly selected schools in Punjab. Descriptive statistics from the survey highlighted that teachers value non-financial drivers such as employer recognition, community recognition, and professional development opportunities.<sup>9</sup> This contextual setting creates direct relevance and demand to explore questions around the effectiveness of non-financial incentives more deeply.

### **3.2 Quaid-e-Azam Academy for Educational Development (QAED)**

The experiment is set-up in collaboration with the Quaid-e-Azam Academy for Educational Development (QAED), an attached department of the Punjab School Education Department,

---

<sup>8</sup>See I-SAP (2017).

<sup>9</sup>Based on author’s own analysis of the Teacher Motivation survey data.

which holds the mandate to provide professional development to all pre-service and in-service public school teachers in Punjab. The academy offers a range of in-service professional development opportunities such as trainings in subject specific content, pedagogy, and leadership and management to name a few.

Despite yearly investments in various professional development trainings, QAED is yet to establish a process for collecting rigorous evidence on whether teachers engage and learn in these trainings and how the knowledge acquired through trainings is translated to their practices in the classroom and the school. In addition, while different master trainers employ various strategies to improve teacher engagement and effort in trainings, there are no institutionalised financial or non-financial incentives linked to high performance in trainings.<sup>10</sup>

I partnered with QAED on a specific training called the ‘Student Leadership Development Programme’ (SLDP) which was targeted at 15,000 head teachers of elementary, secondary, and higher secondary schools across Punjab (grades 6 to 12). The training was a specialized programme for providing skills in coaching, leadership, and school management over four days.<sup>11</sup> The curriculum was designed by the British Council, following which trainings were provided to a selected pool of 634 master trainers. After the initial training, 500 master trainers were validated by the British Council for cascading the trainings further down to the head teachers.

The training was organized and implemented at the district level at the relevant district training center. Given the high number of head teachers in each district, multiple training sessions were formed to receive the 4-day training with a cap of 30 teachers per session. Depending on the size of the district, each district had a total of 12-24 sessions with the total number of teachers at under 30 in each session. The process of assigning trainees to these sessions was done by the QAED head quarters such that each session had equal representation of rural and urban school

---

<sup>10</sup>This is barring promotion-linked trainings which are offered to promotion eligible teachers. All participants are required to score above a certain grade to remain eligible for promotions. Although these trainings do have a clear incentive, these are only offered to promotion eligible teachers and form a small proportion of the overall portfolio of trainings offered to all teachers.

<sup>11</sup>Specific modules included the following: 1) The power of coaching, 2) Co-curricular activities , 3) Protecting children, 4) Student leadership, 5) Staff and distributed leadership, 6) Leave rules, and 7) Pupil voice.

head teachers. Given capacity constraints at the district training center, the training was spread over 3 sequential rounds to accommodate all the sessions.<sup>12</sup>

Each training session also included a training pre-test and post-test to measure learning gains from the training. These were designed by the project director of the SLDP at QAED and were subsequently validated by their British Council counterparts.<sup>13</sup>

## 4 Experimental Design

### 4.1 Treatment Arms

Our four different treatment groups are different variants of employer recognition. Two of the recognition incentives either make peer/collegial approval or career-benefits of recognition salient. The other two treatments cross a self-efficacy enhancing frame with the first two recognition incentive schemes. The recognition incentive is a standard tournament-based employer recognition reward. Within a training session, teachers who score the highest in the training post-test score or show the maximum improvement (over the pre-test score) qualify for a prestigious certificate that is authenticated by the QAED head quarters. This encourages teachers across the entire distribution of the classroom instead of only high ability teachers (as in Ashraf, Bandiera, and Lee, 2014). The experimental design aims to test the four hypotheses listed in Section 2.

The sequencing of activities in the training is as follows. On the first day, teachers fill out a self-administered baseline survey and take the training pre-test which is managed by our team of enumerators. Teachers do not learn about their pre-test score following the test. After the pre-test, enumerators administer the relevant recognition incentive following a specific predetermined script.<sup>14</sup> This is followed by the scheduled training over the next four days. On the fourth and

---

<sup>12</sup>For example, if a district had a total of 12 training sessions, these were allocated across 3 sequential rounds such that each round had 4 training sessions operating simultaneously.

<sup>13</sup>we return to the discussion on the design of the tests and nature of question in Section 4.3.1.

<sup>14</sup>To ensure quality and uniformity in the administration of the recognition incentives across all treatments,

final training day, teachers take a training post-test at the end of the training followed by an endline survey. Details of each treatment arm are given below:

*Control group:* Teachers in this group are administered a neutral script by the enumerator which highlights the broad goals of the SLDP training. All other activities such as the baseline survey, pre-test, post-test, and endline survey operate as in all the other groups.

*Peer Recognition (T1):* Teachers in this treatment group are informed that if they meet the required qualification conditions, they would be eligible for receiving a prestigious recognition certificate in a district ceremony which would be attended by their peers and district staff. This treatment leverages the motivation for peer/collegial esteem and social distinction. The script for T1 is exactly the same as the control group except for the additional information about the recognition incentive. All other activities such as the baseline survey, pre-test, post-test, and endline survey operate as in all the other groups.

Within our context, head teachers rarely have opportunities for receiving recognition socially in front of their peers and colleagues.

*Career-based Recognition (T2):* Teachers in this treatment group are informed that if they meet the required qualification conditions, they would be eligible for receiving a prestigious recognition certificate which would be given to them privately. In addition, they are also told that the names of the winning employees would be included in an ‘excellent teacher list’. This list would be shared with the leadership of the School Education Department and their district’s leadership which could make them eligible for future career opportunities within the department. The script for T2 is exactly the same as T1 except for the difference in the way the certificate is distributed and its related benefits are made salient . All other activities such as the baseline survey, pre-test, post-test, and endline survey operate as in all the other groups.

While formal career incentives for teachers are limited since promotions are linked to seniority,

---

standardized delivery of the script across enumerators was essential. To do this, a master version of each script was pre-recorded and shared with the enumerators along with guidelines on necessary pauses and momentum. Each enumerator was given targeted feedback on their delivery prior to being approved for the job.

three types of informal career incentives might be relevant for how head teachers can use the recognition reward to their advantage in our setting. First, head teachers may want to be posted to better performing schools as opposed to poor performing schools. Second, once head teachers become eligible for promotion they may want to be selected for promotion before other competing colleagues.<sup>15</sup> Third, head teachers may have preferences to be posted laterally to other positions within the department (such as additional charges for managing donor-funded programmes or any other vacant positions at the same grade level).

*Peer PLUS (T3)*: Teachers in this group are administered the same script as T1. However, prior to the T1 script, they are administered a PLUS script which uses a self-efficacy enhancing frame to bolster individual self-efficacy to do well in the training and their job more broadly. More details on the frame are provided below.

*Career PLUS (T4)*: Teachers in this group are administered the same script as T2. Prior to that, teachers are administered the PLUS script which is the same self-efficacy enhancing frame as in T3.

*Self-efficacy frame*: Bandura (1986) defines the concept of self-efficacy as the “perception of one’s capability to accomplish a given level of performance”. There is a distinction between generalized self-efficacy and domain-specific self-efficacy which is important since self-efficacy of individuals can vary across different domains (Bandura, 1986). Teacher self-efficacy, for example, measures self-efficacy within the specific domain of the teaching profession.

Our self-efficacy frame aims to enhance teachers’ domain specific self-efficacy (which includes teacher perceptions in their ability to do well in the training and their job). Our intervention is centered around a one-pager with three inspirational stories of head teachers from Punjab that trainees are asked to read and reflect upon. The stories are meant to serve as role models to bolster existing levels of self-efficacy (as in Beaman et al., 2012 and Tanguy et al., 2014 for example).<sup>16</sup>

---

<sup>15</sup>Employees who are eligible for promotion have to wait for their turn to get their promotion approved.

<sup>16</sup>We select these stories from a report on star teachers compiled by the Punjab School Education Department

To create moments of reflection that are relevant to teacher performance in the training and their jobs, the story-reading activity is sandwiched between two sets of activities. Prior to sharing the stories, trainees are asked to reflect on one key achievement and one key limitation that affects their performance in trainings and their jobs. This is followed by the story-reading activity. After reading the stories, trainees are asked to reflect on how they can address their own limitations. We repeat a refresher version of this activity on the third day of the training as well. Appendix E.2 shows the self-efficacy one-pager which was shared with teachers.

	Experimental Treatments			
	Neutral script:	Script for incentive:		Self-Efficacy script:
		Peer-Recognition	Career-Based	
Control	+			
Peer Recognition	+	+		
Peer PLUS	+	+	+	
Career-Based Recognition	+		+	
Career PLUS	+		+	

Figure 1: Summary of Treatments

## 4.2 Randomization

While the SLDP training was spread across all 36 districts of Punjab, our experiment focuses on 7 districts spread across the north, south, and central regions of the province (See Figure 2). Training sessions in each district were assigned a *session number*. Stratifying by district, we randomly allocate a total of 131 training sessions to four different treatments and the control group. This yields a sample of 3,394 head teachers across 131 training sessions in 7 districts of

---

in 2017 to identify and record high performing teachers.

Punjab. Descriptive statistics in Table A.1 show that our sample is 57% female, has an average teacher age of 46 years, and an average of around 20 years of experience in the service.

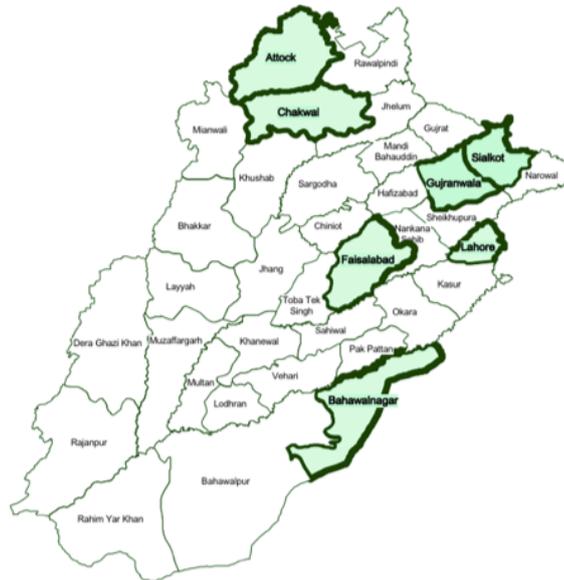


Figure 2: Districts included in the Sample

## 4.3 Data and Balance Checks

### 4.3.1 Data

*Teacher Training Performance Data.* Our primary outcome of interest is teacher training test scores. Both the pre and post-tests were developed by the SLDP staff at the QAED headquarters. The tests included a total set of 15 MCQ questions that were directly related to the taught content. For quality control, the items were validated by the SLDP coordinator followed by the British Council. Given the training in each district had multiple rounds, the pre and post-test questions were different across rounds (although tested the same learning objectives). Within each round, the pre and post-tests included the same set of questions with the only difference being in the ordering of the questions in the two tests (and the ordering of options within the questions). To reduce chances of gaming, the tests in the experimental districts (and other districts) were not shared with the master trainers ahead of time but were instead shared by the research team on the day of pre-test. Our baseline pre-test score presented in Table A.1 shows that head teachers scored 34% on average with very few teachers subject to ceiling or floor effects.<sup>17</sup>

We are also interested in understanding the extent to which teacher effort on the training test is driven purely by the incentives. To capture this, we add an additional dimension to the design of the tests. Both the training pre and post-tests include two sections – an *incentivized* and a *non-incentivized* section. When trainees are administered the incentives, they are explicitly told that they will qualify for the recognition certificate based on their performance on the incentivized section only. Since effort could be diminishing in the length of the test, we also randomize the order of the two sections in the tests. This provides us with additional data to observe differences in treatment effects between the incentivized and the non-incentivized dimension. In addition, it allows us to compare how our PLUS treatments perform on the incentivized dimension where the incentives interact directly with the self-efficacy frame versus the non-incentivised dimension where the interaction effects are muted in comparison.

---

<sup>17</sup>See Appendix D for a sample of the training tests.

*Teacher Surveys.* We collect a set of key secondary outcomes through teacher surveys at endline. These include indicators such as teacher self-efficacy, beliefs about performance on the post-test, intrinsic motivation, and perceived credibility of the department. We measure self-efficacy using a short 4-question tested scale that measures domain specific self-efficacy of teachers as in Fackler and Malmberg (2016). Extrinsic and intrinsic motivational orientations are captured using a tested battery of questions as in Amabile et al. (1995). Departmental credibility is captured through a set of questions that measure teacher perception of the likelihood that the department would disburse any hypothetical promised recognition rewards or bonuses.

We also capture a range of variables in our baseline survey to study heterogeneous treatment effects. These include basic teacher characteristics such as age, gender, salary, and years of experience; job characteristics such as civil service grade, time till next expected promotion; and non-cognitive traits such as personality type, intrinsic/extrinsic motivational orientation, pro-social motivation, self-efficacy, and locus of control. We measure personality through the short Big Five Inventory (Rammstedt and John, 2007; Soto and John, 2017) and pro-social motivation through the Perry PSM index (Perry, 1996).<sup>18</sup>

*Master Trainer and Enumerator Data.* We also collect data on enumerator characteristics such as age, years of experience, years of education, and communication skills to be able to control for enumerator effects in our estimation.

In addition, we also collected on a range of master trainer characteristics such as age, years of experience, and number of trainings attended (as a proxy for quality). These are used as controls in the analysis.

*School Performance Data.* Although our primary outcome of interest is training test scores, we are also interested in capturing the impact of our treatments on downstream outcomes at the school level. For these outcomes, we acquire access to administrative data on monthly school monitoring visits from the Education Department. This data includes a range of indicators such

---

<sup>18</sup>For personality, we measure each trait separately and then convert them into z scores. These are then averaged to form one Big Five Index as in Callen et al. (2016). For the PSM index, we calculate the index as an average of all the scale items and then normalize the index.

as status of school facilities, teacher attendance, student attendance, number of school council meetings held, and whether the school report card is displayed in the school or not. We collate this data for the entire period of January to November 2019. Since the experiment was rolled out in June 2019, the pre June data is used as the baseline in our estimations and the post June data is used as outcome data.

*Qualitative Data.* After completing our analysis, we also held several individual discussions with key stakeholders at QAED to understand our results. This included the SLDP coordinator, the assistant to the SLDP coordinator, the QAED Training and Planning Coordinator, and five different QAED master trainers. The discussions included a presentation of the key results from our analysis to each stakeholder followed by comments and observations from the relevant stakeholder.

### 4.3.2 Balance Tests and Implementation

Table A.2 shows balance across treatment arms for four different categories of variables: basic teacher characteristics, job characteristics, training baseline test score, and teacher non-cognitive traits. We conduct tests of equality for each variable across all treatment groups. Our training baseline score is balanced at the 5% level of significance. Out of a total of 104 tests, 7 are different from zero at the 5% level. We control for these variables in our analysis.<sup>19</sup>

Attrition was not a serious concern in our study given the trainings were mandatory for head teachers to attend. However, there is small attrition in our sample (3%) due to teachers being missing on the fourth day of the training when the post-test took place.<sup>20</sup> Table A.3 shows that attrition is not related to any of our treatment groups and Table A.4 shows that the attrited and main sample are balanced across a range of teacher characteristics at baseline.

Where spillovers are concerned, these are unlikely in our setting. The treatment is at the training

---

<sup>19</sup>We also conduct joint F tests across all groups. All p-values for the joint test are greater than or equal to 0.10.

<sup>20</sup>This occurred either due to personal emergencies or teachers being absent without any officially sought leave.

session level and there is minimal interaction between sessions during the day as trainings are conducted from 8:00 am to 2:00 pm every day within specific training classrooms after which trainees head home. In addition, we only have one teacher from each school which further minimizes the chances of spillovers after the training is over each day. Spillovers across rounds of trainings are also unlikely given there is no time span across rounds to enable interaction between teachers across schools.

## 5 Empirical Strategy and Results

### 5.1 Empirical Strategy

To identify the causal effect of our interventions, we estimate the following:

$$y_{isd}^{Post} = \beta_0 + \rho \cdot y_{isd}^{Pre} + \sum_j \beta_j T_{isd}^j + \gamma X_{isd} + \mu_d + \alpha_r + \epsilon_{isd} \quad (1)$$

Where  $y_{isd}^{Post}$  is the post-test score for teacher  $i$  in session  $s$ , and district  $d$ ;  $y_{isd}^{Pre}$  is the pre-test score that serves as our baseline measure for the ANCOVA estimation. The post-test and pre-test scores are normalized by the mean and standard deviation of the control group. Hence, the treatment effects are observed in standard deviations units.  $X_{isd}$  is a vector of teacher, enumerator, and master trainer controls that we include in our estimation for power. These are selected through the LASSO post double selection procedure following Belloni et al. (2014). Since our randomization is stratified by district, we include district fixed effects (as captured by  $\mu_d$ ) to increase the efficiency of our estimate. We also control for training round effects,  $\alpha_r$ , by adding round dummies. Finally, errors are clustered at the training session level which is our unit of randomization (as in Abadie et al., 2017).  $\beta_j$  are the coefficients of interest.

We use Intention to Treat (ITT) to estimate our treatment effects. A small proportion of teachers

(6%) refused to participate in the employer recognition scheme.<sup>21</sup> Table A.5 shows that non-consent is not significantly related to any of the treatment groups.<sup>22</sup> However, non-consenting teachers are significantly older, have more years of experience, and also do worse on the baseline pre-test as compared to the teachers who consent (Table A.6).

## 5.2 Treatment Effects on Training Test Scores

*Treatment effects of Peer (T1) and Career (T2) Arms.* Our main treatment effects on training test scores can be seen in Table 1, both with and without controls. Since we had imbalance on some teachers characteristics, we discuss the results where we include controls in column 2. We find that teachers in the Career arm score  $0.25\sigma$  higher post-test scores compared to the control group (significant at the 10% level). In comparison, the Peer arm has a coefficient of 0.03 and is insignificant.

We investigate whether the treatment effect on the Career arm is driven by the lower or upper tail of the distribution of training test scores. Quantile treatment effects in Table 2 show that the Career arm has a positive coefficient in the range of 0.20 - 0.37 across the entire distribution and the coefficient is always higher than the Peer arm (see Figure 3). The coefficients on the Career arm are also significant at the 5% level in the upper tail of the distribution, where they are also significantly different from the Peer arm. We also conduct Kolmogorov–Smirnov tests of equality of distribution between the Peer and Career arm which confirm that the two distributions are significantly different at the 1% level (p-value<0.01).<sup>23</sup>

These results provide indicative evidence that the motivation to receive some form of future tangible career benefits through recognition is a stronger motivator as compared to receiving peer/collegial approval. Given the Career arm made potential career benefits salient, these results also point towards the value of informal career benefits that the teachers could have

---

<sup>21</sup>This included 207 teachers which is roughly 6% of the sample.

<sup>22</sup>Our main treatment estimates remain the same with TOT estimation.

<sup>23</sup>The distributions are presented in Appendix B.3.

accessed through the certificate (such as getting transfers to preferred schools, getting selected for promotions faster once eligible, or getting appointed to higher grade positions on the same salary scale if positions become vacant).

Qualitative discussions with selected trainees and our main counterparts at QAED help explain why we observe indicative effects on the Career arm but no effects on the Peer arm. They suggest that the strength of the peer/collegial approval channel may be weak for head teachers given they have already risen through the ranks and established respect, reputation, and esteem amongst their peers and colleagues. Hence, such a channel may be more effective for primary and secondary school teachers who are younger and looking to establish their reputation amongst their peers (as we observe in Mansoor, 2019). However, where the career benefits channel is concerned, head teachers have several informal career incentives in the system such as postings to their choice of school or other influential lateral appointments. We test the theory behind the career benefits channel further by observing heterogeneous treatment effects by variables that can make such informal career incentives more salient in Section 5.3.

## Quantile Regressions

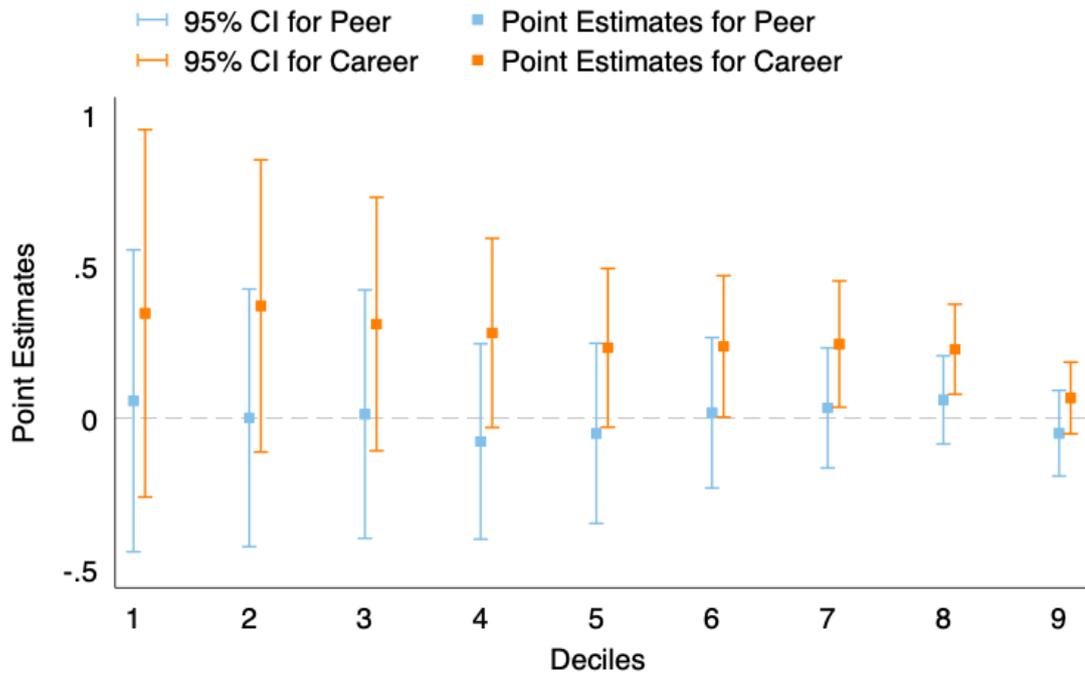


Figure 3: Quantile Regressions for Training Test Scores

*Treatment effects of Peer PLUS (T3) and Career PLUS (T4) Arms.* Column 2 in Table 1 shows that the net impact of adding the self-efficacy frame in Peer PLUS and Career PLUS is negative – a reduction of  $0.15\sigma$  for Peer PLUS (not significant) and a reduction of  $0.28\sigma$  for Career PLUS (significant at the 5% level). We also conduct Kolmogorov–Smirnov tests of equality of distribution between the Peer and Career arm and their PLUS counterparts and find that the distributions are significantly different from each other (p-value  $<0.01$  for both tests).<sup>24</sup> Since the net impact of adding the frame moves in a negative direction for both arms, we pool the PLUS treatments (i.e. those receiving the frame) and the non-PLUS treatments (i.e. those not receiving the frame) in Columns 3 and 4. Column 4 shows that the net impact of adding the self-efficacy frame to either of the arms is to lower training test scores by  $0.24\sigma$  (significant at the 5% level). This indicates that adding the self-efficacy enhancing frames to our recognition incentives resulted in teachers reducing effort on the test.

We further investigate if the negative effects of the self-efficacy frame specifically come from the lower or upper tail of the distribution of training test scores. Quantile treatment effects in Table 2 show that the observed negative effects are consistent across the distribution and do not particularly come from the lower or upper tail of the distribution.

The net negative impact of the frames implies that they operate like substitutes instead of complements to our recognition incentives, which runs counter to our hypotheses in Section 2.<sup>25</sup> While our findings are inconsistent with the positive effects of self-efficacy interventions on a range of other outcomes such as job search and health-seeking behaviours (see Eden and Aviram, 1993; Haushofer, John, and Orkin, 2019), they are in line with the arguments of skeptics who suggest that creating “positive illusions” of oneself can often be misleading and result in

---

<sup>24</sup>The distributions are presented in Appendix B.4.

<sup>25</sup>Our analysis of the net impact of the frames assumes that the PLUS treatments are a linear combination of the incentive scheme and the self-efficacy frame. However, if the two treatments bundle together in non-linear ways our point estimates of the net impact of the frame would not be accurate. Irrespective of this assumption, our results do show that the bundled treatment with the frames does worse than the incentive alone. This implies that while this negative effect may not represent the ‘true’ net impact of the frames, it does capture the substitution effect of including the frames on the effect of the incentives. This interpretation is equally relevant for the implications of these results.

negative results (Baumeister, 1999; Swann, 1996).<sup>26</sup> In Section 5.4.1, we present supplementary evidence that explores mechanisms for understanding why teachers reduce effort in response to the frames.

*Treatment Effects on the Non-Incentivized Dimension.* Our design has a built-in incentivized and a non-incentivized dimension of the test. We leverage this aspect of our design to study whether the response to the incentives is purely strategic (as captured through the incentivized dimension) or whether the incentives also motivate effort on the non-incentivized dimension. To estimate these effects, we use the same specification as (1), but with the non-incentivized test scores as our outcome variable.

Column 6 in Table 1 presents the results. We find that the treatment coefficient on the Career arm is insignificant and much smaller as compared to the incentivized dimension (0.025 versus 0.255). This indicates that the positive effects observed on the Career arm are strategic, i.e. teachers respond strategically by simply exerting more effort on the incentivized dimension to acquire the incentive.<sup>27</sup>

The net negative impact of the PLUS treatments on the incentivized dimension (that we present earlier in columns 2 and 4) implied that teachers reduced effort on the test in response to the frames. However, this could either be due to interaction effects between the frame and the incentives or irrespective of such interaction effects. Column 6 shows that the net impact of the PLUS arms on the non-incentivized dimension remains negative as on the incentivized dimension, with the results particularly negative and significant for the Career PLUS arm. These results imply that teachers who received the frames did not simply reduce effort due to interaction effects

---

<sup>26</sup>Bénabou and Tirole (2002) explain this further through a theoretical model which shows how rational individuals process outside information to balance the benefits and risks of positive perceptions, and the conditions under which positive perceptions can have negative effects.

<sup>27</sup>It is possible that this strategic effort is exerted to cheat/game the test instead of exerting more effort on the test. In terms of gaming the test, trainees could have received test questions ahead of the training test or tried to cheat during the test. Our implementation ensured that neither was possible. It might still be possible that trainees who wanted to score better tried to recall questions from the pre-test and memorized those responses ahead of the test. However, we see the latter not as an indication of gaming but as evidence of wanting to exert more effort on the test as a result of the incentive.

between the incentives and the frames but actually reduced overall effort in the training.

### 5.3 Heterogeneity

*Heterogeneity by Key Moderators (of the peer recognition and career-benefits channel of recognition).* We investigate whether our two main recognition incentives – the Peer and Career arms – work through the channels hypothesized in Section 2. To test the theory for the peer/collegial approval channel, we explore whether treatment effects vary by number of peers each trainee knows in their group and the average proportion of peers known to each other in a training session. We expect that the Peer arm should work better if trainees know their peers or if the proportion of peers that know each other in a session is higher. This rests on the assumption that the peer-esteem from the Peer Recognition arm would be stronger if the teacher knows his/her peers well. To test the theory for the career-benefits channel, we explore whether treatment effects vary by expected time till next promotion and frequency of visits by senior leadership (such as Secretary and District Education Officers) to the trainees' districts. While an upcoming promotion would create direct incentives to get promoted to a position of choice, higher visibility to senior leadership would increase opportunities to benefit from various types of informal career incentives in the system. We expect that the career-benefits channel should be stronger for teachers whose promotions are due sooner and those who have more visibility to the senior leadership as both of these would allow teachers to use the recognition certificate in the Career arm to their advantage.

Columns 1 and 2 in Table 3 show that when the Career arm works, the treatment effect comes from individuals who are due for their next promotion sooner or have a higher frequency of visits by the Secretary. This lends support to the theory behind the career-benefits channel. Columns 3 and 4, on the other hand, show no support for the Peer arm being stronger for trainees who individually know more peers or those who are in sessions where more peers are known to each other. This links back to our qualitative discussions which highlight how the peer approval channel may not be strong for head teachers in the first place.

*Heterogeneity by Teacher Characteristics.* We explore heterogeneous treatment effects by gender, years of experience, ability (proxied by baseline test scores), personality type, and pro-social motivation. For each of our continuous variables, we construct above and below median categories to conduct our analysis.

Columns 1 and 3 in Table 4 show that the treatment effect on the Career arm is significant for females and teachers with more years of experience. The stronger treatment effect for females is intuitive given our experiment is based in a contextual setting where women have limited networking opportunities as compared to men, which is an important mechanism for advancing careers (Tanwir, 2014). The Career arm may have therefore created new opportunities for women to access informal career incentives. The stronger treatment effect for more experienced teachers could be due to the fact that experienced teachers have more information about the system and how to use the career-based recognition reward to their advantage. Columns 4, 5, and 6 highlight a consistent trend that the treatment effect for the Career arm is significant for individuals who score higher on the Big Five Personality (John et al., 2008) and the Perry PSM index (Perry, 1996), and donate more in a hypothetical dictator game. This is inline with Callen et al. (2016) who show how health monitors who score higher on the BFI index respond more to a monitoring intervention in Punjab Pakistan.

Table 5 pools the PLUS treatments to investigate heterogeneity along these dimensions to understand the negative impact of the PLUS treatments. Column 1 shows that the decrease in test scores in PLUS treatments is more salient for men versus women. We also find that more experienced teachers show a relatively greater decrease in test scores (Column 3). Further subgroup analysis suggests that the negative effects of the frames are particularly salient for men with more years of experience.

## 5.4 Treatment Effects on Self-Efficacy

The results presented so far look at our primary outcome of interest - training test scores. In this section, we present treatment effects on endline self-efficacy which is our key secondary outcome. The main assumption behind the design of the PLUS treatments is that recognition incentives may have negative effects on individual self-efficacy and subsequently on training test scores, and hence introducing self-efficacy enhancing frames can address such negative effects. We investigate two questions in relation to this assumption. First, whether the Peer and Career recognition arms have any negative effects on endline self-efficacy as expected? Second, whether the PLUS treatments bolster individual self-efficacy despite the negative effects on test training test scores?

To estimate our treatment effects, we run the same specification as (1) with teacher self-efficacy as the outcome measure. We normalize the self-efficacy variable at baseline and endline by the mean and standard deviation of the control group and report our treatment effects in standard deviation units.

Column 2 in Table 6 shows that the Peer and Career arms have no significant effects on endline self-efficacy. However, Peer PLUS and Career PLUS both have positive and significant effects on endline self-efficacy, with an increase of  $0.11\sigma$  and  $0.14\sigma$  respectively. The net impact of the self-efficacy frame is also positive and significant at the 10% level for the Career PLUS arm. Column 5 pools the PLUS and the non-PLUS arms and shows that the PLUS arms lead to a positive and significant increase of  $0.12\sigma$  in endline self-efficacy, with the net impact of the self-efficacy frames at  $0.11\sigma$ . These results confirm that while the frames reduce training test scores, they do boost self-efficacy.

While teacher self-efficacy improves in the PLUS treatments, the content of the frames could have still affected teacher motivation or their perceptions about their jobs negatively resulting in lower effort on the test. To investigate this, we test for any treatment effects on three other secondary outcomes that we collect during our endline survey – teacher intrinsic motivation,

locus of control, and credibility of the department. Table A.7 in the appendix shows no negative effects of the frames on these outcome variables. In fact, we find indicative evidence of higher intrinsic motivation and internal locus of control in the Career PLUS arm.

#### 5.4.1 Mechanisms for Understanding Negative Effects of the Self Efficacy Frames

In this section we investigate the mechanisms underlying the negative net impact of the self-efficacy frames on training test scores. We hypothesize two potential explanations. First, while the frames improved teacher self-efficacy they could have simultaneously made teachers overconfident in their ability to do well in the training which could have led to a reduction in teacher effort (and ultimately training test scores). This explanation is consistent with skeptics who argue that interventions that aim to improve individual self-esteem or efficacy can at times *overcorrect* beliefs about ability leading to dangers of overconfidence (Swann, 1996; Baumeister, 1999; Bénabou and Tirole, 2002). Second, while the frames improved teacher self-efficacy, they could have simultaneously compromised their cognitive bandwidth (Mullainathan and Shafir, 2013).<sup>28</sup> In our experiment, the frames provided teachers with additional information which may have caused an information overload that mentally taxed teachers or diverted their attention away from the training.

*Overconfidence.* In our endline survey, we ask teachers to report how much they believe they scored on the training post-test. This allows us to construct a direct measure of teacher overconfidence as the difference between beliefs about performance and actual performance on the post-test in percentage terms. To investigate treatment effects on overconfidence, we run the same specification as (1) but use teacher overconfidence as our outcome measure.

Columns 1 and 2 in Table 7 show that the net impact of the self-efficacy frame on overconfidence is positive and significant, making teachers 6% points more overconfident in Peer PLUS and 5% points more overconfident in Career PLUS. Columns 3 and 4 pool the PLUS and the non-PLUS

---

<sup>28</sup>The authors explain that the human brain processes for tasks that require effort can at times be compromised when individuals are mentally taxed. They refer to this as “bandwidth”.

treatments and show that the net impact of adding the self-efficacy frame to either of the PLUS arms is to make teachers 5.6% points more overconfident. Since our measure of overconfidence is a continuous variable where outliers might drive results, we re-define the measure as above and below median overconfidence and repeat the estimation of our treatment effects in Columns 5-8. Column 8 shows that that the net impact of the PLUS arms on overconfidence remains positive and significant, with the frames increasing the proportion of above-median overconfident trainees by 12% points on average.

*Compromised bandwidth.* The other competing theory that we hypothesize is that the frames could have compromised teachers' cognitive bandwidth by giving them too much information which could have resulted in added stress or distraction from the training leading to reduced effort.

We are limited by the lack of direct observational or survey data that measures distraction, stress or other aspects of cognitive bandwidth that can allow us to completely rule out this channel. However, if compromised bandwidth was indeed the leading explanation for our negative results, it would not help explain our strong and robust effects on overconfidence. This implies that the negative results are most likely coming from the overconfidence channel.

*Mediation Analysis for Overconfidence.* We use mediation analysis to quantify the strength of the overconfidence channel in explaining the negative effects of the self-efficacy frames.

We use the procedure of sequential g-estimation as laid out in Acharya et al. (2016) to identify the Average Controlled Direct Effect (ACDE) of the PLUS treatments after accounting for the effects of overconfidence. While ACDE is often calculated by including the post-treatment mediator in the original estimation, this leads to inconsistent estimates due to selection bias. The sequential g-estimation procedure of estimating ACDE, on the other hand, excludes the effect of the mediator (in this case overconfidence) from the observed treatment effect by fixing it at the same level for all units which helps avoid issues of selection bias.<sup>29</sup> The identification of the

---

<sup>29</sup>This includes two stages. Stage 1 includes regressing the main outcome on treatment, pre-treatment controls, the mediator, interaction between the mediator and treatment, and interaction between the mediator and all other

estimates rests on one central assumption - sequential unconfoundedness - which incorporates two further assumptions: a) there is no omitted variable that is correlated with the error term and the outcome variable; and b) there is no omitted variable that confounds the effect of the mediator on the treatment post controlling for pre-treatment variables and other post-treatment controls. In our setting where treatments are randomly assigned, a) is not violated by design. We assume b) is not violated in our particular setting.<sup>30</sup>

Since overconfidence is measured as the difference between beliefs about training post-test score and actual post-test score, including this measure of overconfidence to estimate the de-mediated outcome poses endogeneity concerns due to the high mechanical correlation between overconfidence and the outcome variable – post-test scores. To address these concerns, we predict overconfidence in our sample using baseline variables selected by LASSO as the best predictors of overconfidence. Table A.8 in the appendix shows that the correlation between predicted overconfidence and actual overconfidence is around 0.31. We also estimate treatment effects on our predicted measure of overconfidence and find that the impact of the PLUS treatments on overconfidence remains positive and significant, though the effects are smaller (see Table A.9). This gives us confidence in using our predicted measure of overconfidence for the mediation analysis.

We use predicted overconfidence to estimate our de-mediated outcome (i.e. training test scores) and then re-estimate our treatment effects (see original estimation and a revised estimation based on the de-mediated outcome in Table A.10). Figure 4 below shows that the treatment coefficient on the de-mediated outcome reduces by almost 50% suggesting that overconfidence approximately explains up to 50% of the observed negative treatment effects of the PLUS arms.

---

pre-treatment variables. Following this, we calculate the de-mediated outcome which is the predicted outcome excluding all coefficients that include the mediator fixed at a specific value. Stage 2 includes regressing the de-mediated outcome on the treatment. The coefficient on the treatments in the second stage is the ACDE.

<sup>30</sup>Assumption b) is unlikely to be violated in our setting since individual beliefs of overconfidence are unlikely to have many other confounders (other than a key set of variable such as individual self-efficacy and locus of control) that lead to a reduction in teacher effort and test scores. We control for such potential post-treatment confounders such as self-efficacy and locus of control.

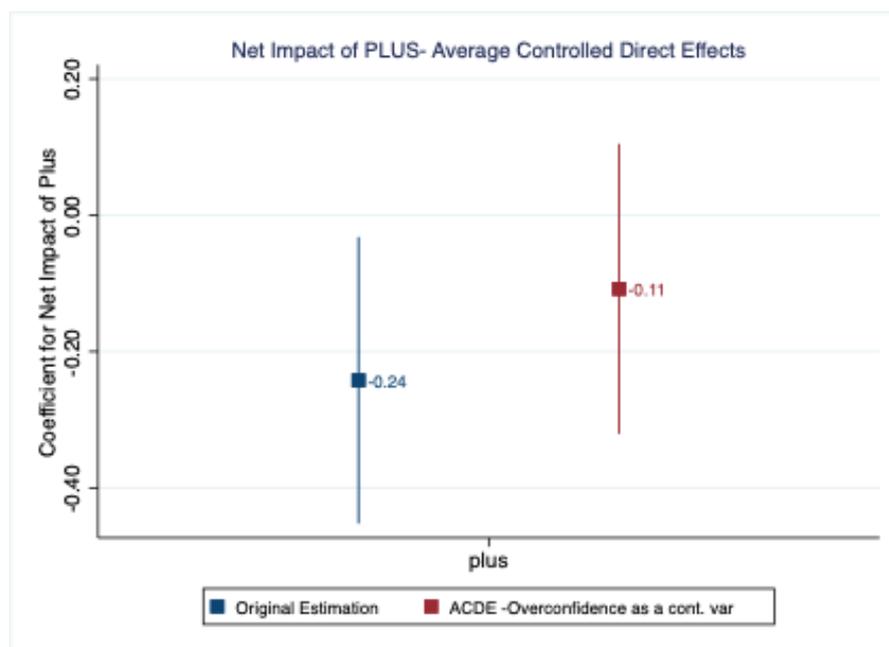


Figure 4: Average Controlled Direct Effect (De-mediated Test Scores)

## 5.5 Treatment Effects on School-Level Outcomes

While the immediate aim of the SLDP training was to improve head teacher knowledge in school management and leadership that we capture through endline training test scores, ultimately the education academy’s goal was to create improvements in three levels of downstream outcomes as a result of the training. First, through improved knowledge from the training, an implementation of the lessons learnt from the training by the head teachers to their own management practices and attitudes in their schools; second, as a result of improved head teacher management practices, an improvement in a range of school-level outcomes and teacher practices; and third, as a result of improved school-level outcomes and teacher practices, an improvement in student learning outcomes. We are interested in investigating the extent to which our recognition incentives in the training are effective in improving these downstream outcomes.

While we do not have access to relevant data to measure treatment effects on the first and third set

of outcomes, we have access to administrative data from monthly school-level monitoring visits that are conducted by the Education Department. This provides us with a range of school-level performance indicators such as teacher absenteeism, student attendance, number of classrooms used, and number of school council meetings held that are two steps down the causal chain.

We investigate treatment effects on these school-level indicators through the following specification:

$$S_{isd}^{Post} = \beta_0 + \rho \cdot S_{isd}^{Pre} + \sum_j \beta_j T_{ij} + \gamma X_{isd} + \mu_d + \epsilon_{isd} \quad (2)$$

Where  $S_{isd}^{Post}$  represents the relevant school-level outcome for teacher  $i$  in session  $s$  and district  $d$ . Each outcome measure is calculated by averaging out performance over two months after the training period.  $S_{isd}^{Pre}$  represents the baseline measure for these outcome variables, averaged over two months before the treatment was administered. As in other specifications, we use district fixed effects and cluster errors at the session level.

Table 8 shows treatment effects on a range of school level outcomes - % teachers absent, % classrooms used, % student attendance, number of school council meetings, whether school is clean, whether report cards are displayed in the school, and number of additional coaches hired for teaching. We do not find significant treatment effects on any of these outcomes for the Career arm or the PLUS arms which showed impact in the training. While Peer PLUS has a significant impact on student attendance, the treatment coefficient is almost zero.<sup>31</sup>

Clear conclusions on whether or not recognition incentives can encourage the implementation of training knowledge to downstream attitudes and practices are hard to draw despite these null results. This is due to the fact that our analysis is limited by the lack of information on outcomes that are only one step down the causal chain such as the head teacher's own practices and attitudes which may have been more likely to be affected. In addition, we do not have

---

<sup>31</sup>We also find a significant impact of the Peer arm on the number of coaches but since the Peer arm has no impact on training scores, we see this as a chance treatment effect on a single outcome.

access to other indicators at the school and teacher level such as school-level budget utilization or teacher practices in the classroom which may have been more likely to be affected by the training. However, these results do indicate the need for further research that collects process-based information and data at various steps of the causal chain from the point of delivery of the training to school-level outcomes to understand how further along the chain can incentives support the implementation (and mainstreaming) of trainings to actual practices.

## 6 Differences and Commonalities with the Pilot Study

In this section, we outline differences and commonalities between the pilot and the main experiment presented in this paper to draw out several key implications.

The pilot and this experiment have several differences in their results which highlights the highly contextual nature of the kind of recognition incentives that we implement in these two experiments. First, the pilot was conducted with primary and secondary school teachers whereas this experiment was conducted with head teachers. Both cadres of teachers have important differences between them in terms of their rank, years of experience, and career aspirations. While the peer approval channel of recognition appears to work for primary/secondary school teachers under some conditions in the pilot, it does not work for head teachers in this paper. Our qualitative discussions indicate that this may be so because head teachers have already risen through the ranks and have established peer-esteem and respect which younger teachers (such as primary and secondary school teachers) have not. This indicates that the motivation for different types of recognition incentives can vary by cadre of teachers. Second, when recognition incentives work in the pilot, they do so via a self-efficacy channel, indicating that effort and individual self-efficacy are complements. We do not find similar evidence in the main experiment. The treatments do not work via a self-efficacy channel. In fact, introducing self-efficacy frames makes head teachers overconfident with negative effects on training test scores. This implies that the way individual non-cognitive traits are affected by such incentives can vary significantly by group-type.

Several important commonalities between the two studies also stand out. First, the career-based channel of recognition appears to work both for the pilot (under some conditions) and the main experiment presented in this paper. This indicates that teachers value informal career incentives in the system such as postings to their preferred schools, lateral postings to other donor-funded projects, or additional charges for any other positions. Second, when recognition incentives work in either of the studies, the treatment effects tend to be significant for individuals who score higher on the Big Five Personality Index (BFI) or report higher levels of intrinsic motivation at baseline.<sup>32</sup>

## 7 Conclusion

We present experimental evidence on the impact of employer recognition on teacher training performance (as measured via training test scores) in in-service trainings held by the Teacher Training Academy of Punjab Pakistan. The study shows that employer recognition can improve teacher performance in trainings if it is linked to tangible career benefits in the future. Despite these positive results, we find that these effects can backfire depending on how such incentives are framed. In particular, we find that adding a self-efficacy enhancing frame to our recognition treatments “over corrects” teacher beliefs about ability to do well in the training leading to overconfidence and reduced effort.

Our results have two key policy implications. First, they open up a discussion on how the public sector can design more effective non-financial incentives for in-service trainings, and for eliciting higher teacher effort more broadly. The career-linked recognition incentive used in this experiment was fairly light touch, yet we find encouraging results which indicate towards the value of informal career benefits in the system. In our particular context, there are several

---

<sup>32</sup>One important difference in this result between the pilot and the main experiment is that individual responses vary by type of intrinsic motivation in the pilot. For example, in the pilot while intrinsic motivation of teachers due to interest/enjoyment in the job matters for how they respond to the treatment, pro-social motivation does not. On the other hand, in the main experiment, pro-social motivation predicts response to treatments.

informal career incentives for teachers such as getting a transfer to a school of liking, getting laterally appointed to an influential position such as Project Director of a large donor-funded program which may be associated with additional pay, or getting appointed to a higher grade position (with the same pay and civil service grade) if a vacancy arises. In the public sector where formal incentive-based reforms are often hard to implement and formal career incentives such as promotions are primarily linked to seniority, designing “soft” non-financial incentives that can leverage informal career incentives can address part of the inefficiency in incentive systems. Second, the sensitivity of our results to framing effects and group-type points towards the external validity of studies that evaluate the impact of such incentives. Our results across the pilot and the main experiment indicate that recognition incentives are highly contextual where the effects can vary a lot across different populations. This requires caution in how such incentives are designed across different contexts and also the importance of piloting before scaling up across the board.

Several additional questions remain open to inquiry. Our experiment was only able to offer the incentive for a single time. Future work could look at the decay rate in the impact of such incentives in trainings, and circumstances under which the effects are sustained. Recognition has been often modelled in standard principal-agent utility maximizing frameworks, but clarity around the weight placed on such incentives in comparison to financial incentives would be useful in calibrating their value and assessing the cost effectiveness of such incentives more explicitly. Our null results on school-level outcomes raise important questions regarding the extent to which incentives in trainings can improve the intended downstream impacts of in-service trainings at the school and classroom level. Given implementation of high quality trainings and achieving their intended downstream effects is generally challenging, this calls for academic inquiry into the extent to which incentives in trainings can encourage such downstream implementation, and whether certain types of incentives are more effective than others in achieving this. Finally, our experiment showed that creating exogenous variation in self-efficacy beliefs of public sector bureaucrats is possible. This opens up the possibility of additional research on how to create exogenous variation in intrinsic motivation and/or other non-cognitive traits through the platform of in-service trainings, and its impact on training performance, the implementation of training

lessons to actual work practices, and in shaping identities, norms, and culture at the workplace.

Table 1: Main Treatment Effects

	Incentivised				Non Incentivised			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Peer	0.014 (0.150)	0.028 (0.155)			-0.095 (0.106)	-0.111 (0.108)		
Career	0.234 (0.151)	0.255* (0.148)			0.045 (0.082)	0.025 (0.091)		
Peer PLUS : Net Impact of Frame	-0.164 (0.144)	-0.145 (0.142)			-0.017 (0.125)	-0.046 (0.122)		
Career PLUS : Net Impact of Frame	-0.246* (0.135)	-0.279** (0.129)			-0.171* (0.095)	-0.191** (0.096)		
Peer and Career			0.125 (0.133)	0.183 (0.137)			-0.023 (0.082)	-0.042 (0.089)
PLUS: Net Impact of Frame			-0.201* (0.103)	-0.242** (0.106)			-0.095 (0.080)	-0.120 (0.080)
Peer PLUS*	-0.15	0.11			-0.11	-0.16		
Career PLUS*	-0.01	-0.03			-0.12	-0.17		
PLUS*			-0.07	-0.06			-0.11	-0.16
Observations	3394	3392	3394	3392	3394	3383	3394	3383
Adjusted R <sup>2</sup>	0.101	0.157	0.094	0.141	0.256	0.258	0.254	0.256
PDS LASSO controls	No	Yes	No	Yes	No	Yes	No	Yes
District Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Errors clustered at the training session level which is the unit of randomization. All regressions are an ANCOVA estimation with baseline values of the dependent variable and with district FE. Controls include trainee-level teacher controls, master trainer controls, and enumerator controls. Training post test and pre test scores are normalized by the mean and standard deviation of the control group. The PLUS treatments with the asterisks present the overall impact of the treatments (Incentive + the frame). Estimates are significant at the \*10%, \*\*5%, and \*\*\*1% level

Table 2: Quantile Treatment Effects

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	10th	20th	30th	40th	50th	60th	70th	80th	90th
Peer	0.057 (0.254)	0.001 (0.217)	0.013 (0.209)	-0.077 (0.165)	-0.050 (0.152)	0.018 (0.127)	0.034 (0.101)	0.060 (0.074)	-0.050 (0.072)
Career	0.346 (0.310)	0.371 (0.246)	0.311 (0.214)	0.281* (0.159)	0.232* (0.134)	0.237** (0.119)	0.245** (0.106)	0.227*** (0.076)	0.067 (0.060)
Peer PLUS : Net Impact of Frame	-0.151 (0.209)	-0.133 (0.207)	-0.134 (0.194)	-0.065 (0.161)	-0.059 (0.157)	-0.088 (0.116)	-0.093 (0.088)	-0.114 (0.074)	-0.115 (0.074)
Career PLUS : Net Impact of Frame	-0.333 (0.273)	-0.287 (0.232)	-0.291 (0.181)	-0.289** (0.137)	-0.255** (0.111)	-0.265*** (0.096)	-0.225** (0.090)	-0.174** (0.072)	-0.141*** (0.043)
Observations	3392	3392	3392	3392	3392	3392	3392	3392	3392
R - Squared	0.214	0.247	0.249	0.247	0.245	0.241	0.233	0.209	0.154
PDS LASSO controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
District Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The regressions report quantile treatment effects. Errors clustered at the training session level which is the unit of randomization. All regressions are an ANCOVA estimation with baseline values of the dependent variable and with district FE. Controls include trainee-level teacher controls, master trainer controls, and enumerator controls. Training post test and pre test scores are normalized by the mean and standard deviation of the control group. Estimates are significant at the \*10%, \*\*5%, and \*\*\*1% level

Table 3: Heterogeneous Treatment Effects - by Moderators

	Post Test Scores			
	(1)	(2)	(3)	(4)
<b>Hetero effects by:</b>	Time till next promotion	Secretary visibility	Peer known in class	Proportion peers known
Below Median x Peer	0.034 (0.166)	-0.029 (0.152)	0.099 (0.154)	0.162 (0.201)
Above Median x Peer	-0.041 (0.160)	0.243 (0.216)	-0.084 (0.168)	-0.197 (0.209)
Below Median x Career	0.328** (0.162)	0.206 (0.157)	0.214 (0.160)	0.099 (0.216)
Above Median x Career	0.126 (0.175)	0.391** (0.184)	0.256 (0.159)	0.241 (0.190)
Below Median x Peer PLUS	-0.225 (0.187)	-0.218 (0.166)	-0.215 (0.172)	-0.015 (0.237)
Above Median x Peer PLUS	-0.105 (0.165)	0.230 (0.214)	-0.092 (0.185)	-0.338 (0.215)
Below Median x Career PLUS	0.030 (0.161)	-0.016 (0.143)	0.019 (0.148)	0.180 (0.166)
Above Median x Career PLUS	-0.051 (0.156)	0.023 (0.207)	-0.063 (0.160)	-0.300 (0.237)
Observations	2181	3394	3394	3394
Adjusted R <sup>2</sup>	0.106	0.104	0.103	0.120
PDS LASSO controls	No	No	No	No
District Fixed Effects	Yes	Yes	Yes	Yes

Notes: Errors clustered at the training session level which is the unit of randomization. The dependent variable is post test scores which is normalized by the mean and standard deviation of the control group. All regressions are an ANCOVA estimation with baseline test scores and district FE. Each column represents heterogeneous treatment effects by a different moderator. Estimates are significant at the \*10%, \*\*5%, and \*\*\*1% level

Table 4: Heterogeneous Treatment Effects - by Teacher Characteristics- I

	Post Test Scores					
	(1)	(2)	(3)	(4)	(5)	(6)
<b>Hetero effects by:</b>	Gender	Ability	Yrs Exp	Personality	Pro-sociality (Perry Index)	Pro-sociality (Hypothetical Dictator)
Below Median (Male) x Peer	-0.023 (0.175)	0.126 (0.193)	0.004 (0.159)	0.015 (0.165)	0.058 (0.158)	-0.104 (0.157)
Above Median (Female) x Peer	0.105 (0.163)	-0.127 (0.134)	0.059 (0.162)	0.023 (0.151)	-0.018 (0.157)	0.141 (0.159)
Below Median (Male) x Career	0.147 (0.181)	0.303 (0.191)	0.138 (0.165)	0.177 (0.158)	0.214 (0.154)	0.147 (0.170)
Above Median (Female) x Career	0.310** (0.150)	0.152 (0.133)	0.353** (0.156)	0.321** (0.148)	0.264* (0.156)	0.317** (0.148)
Below Median (Male) x Peer PLUS	-0.349* (0.206)	-0.091 (0.199)	-0.165 (0.155)	-0.229 (0.175)	-0.132 (0.176)	-0.179 (0.175)
Above Median (Female) x Peer PLUS	0.102 (0.150)	-0.224 (0.149)	-0.095 (0.190)	-0.075 (0.167)	-0.161 (0.169)	-0.142 (0.168)
Below Median (Male) x Career PLUS	-0.114 (0.185)	0.059 (0.181)	-0.041 (0.137)	-0.025 (0.156)	0.029 (0.154)	-0.090 (0.159)
Above Median (Female) x Career PLUS	0.050 (0.149)	-0.11 (0.124)	0.027 (0.169)	0.013 (0.149)	-0.044 (0.149)	0.058 (0.143)
Observations	3394	3394	3394	3382	3393	3335
Adjusted R <sup>2</sup>	0.144	0.102	0.115	0.108	0.103	0.102
PDS LASSO controls	No	No	No	No	No	No
District Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Errors clustered at the training level which is the unit of randomization. The dependent variable is post test scores which is normalized by the mean and standard deviation of the control group. All regressions are an ANCOVA estimation with baseline test scores and district FE. Each column represents heterogeneous treatment effects by a different moderator. Estimates are significant at the \*10%, \*\*5%, and \*\*\*1% level

Table 5: Heterogeneous Treatment Effects - by Teacher Characteristics -II

	(1)	(2)	(3)	(4)
Male x Peer + Career	0.056 (0.154)			
Female x Peer + Career	0.213 (0.140)			
Male x Net PLUS	-0.302** (0.139)			
Female x Net PLUS	-0.141 (0.099)			
Below Med. Ability x Peer + Career		0.215 (0.173)		
Above Med. Ability x Peer + Career		0.013 (0.115)		
Below Med. Ability x Net PLUS		-0.226* (0.116)		
Above Med. Ability x Net PLUS		-0.175 (0.110)		
Below Med. Experience x Peer + Career			0.070 (0.137)	
Above Med. Experience x Peer + Career			0.210 (0.145)	
Below Med. Experience x Net PLUS			-0.164 (0.111)	
Above Med. Experience x Net PLUS			-0.243** (0.112)	
Below Med. Personality x Peer + Career				0.067 (0.138)
Above Med. Personality x Peer + Career				0.198 (0.131)
Below Med. Personality x Net PLUS				-0.216* (0.116)
Above Med. Personality x Net PLUS				-0.198** (0.100)
Observations	3394	3394	3394	3382
Adjusted R <sup>2</sup>	0.138	0.095	0.108	0.099
PDS LASSO controls	No	No	No	No
District Fixed Effects	Yes	Yes	Yes	Yes

Notes: Errors clustered at the training session level which is the unit of randomization. The dependent variable is training test scores. All regressions include district FE. Controls include trainee-level teacher controls, master trainer controls, and enumerator controls. Estimates are significant at the \*10%, \*\*5%, and \*\*\*1% level.

Table 6: Treatment Effects on Self-Efficacy

	(1)	(2)	(3)	(4)
Peer	0.026 (0.047)	0.036 (0.047)		
Career	-0.003 (0.062)	0.001 (0.058)		
Peer PLUS : Net Impact of Frame	0.073 (0.056)	0.078 (0.058)		
Career PLUS : Net Impact of Frame	0.139** (0.066)	0.130* (0.067)		
Peer and Career			0.011 (0.048)	0.018 (0.045)
PLUS: Net Impact of Frame			0.107** (0.045)	0.106** (0.047)
Peer PLUS*	0.10	0.11**		
Career PLUS*	0.14**	0.14**		
PLUS*			0.12**	0.12**
Observations	3366	3364	3366	3364
Adjusted R <sup>2</sup>	0.148	0.150	0.148	0.150
PDS LASSO controls	No	Yes	No	Yes
District Fixed Effects	Yes	Yes	Yes	Yes

Notes: Errors clustered at the training session level which is the unit of randomization. All regressions are an ANCOVA estimation with baseline values of the dependent variable and with district FE. Controls include trainee-level teacher controls, master trainer controls, and enumerator controls. Self-efficacy at baseline and endline is normalised by the mean and standard deviation of the control group. The PLUS treatments with the asterisks present the overall impact of the treatments (Incentive + the frame). Estimates are significant at the \*10%, \*\*5%, and \*\*\*1% level.

Table 7: Treatment Effects on Overconfidence

	Overconfidence - continuous				Overconfidence - above median			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Peer	0.675 (2.824)	0.001 (2.765)			0.032 (0.067)	0.012 (0.062)		
Career	-2.211 (2.898)	-2.565 (2.834)			-0.065 (0.072)	-0.081 (0.066)		
Peer PLUS :								
Net Impact of Frame	6.170** (2.724)	6.526** (2.780)			0.124** (0.058)	0.126** (0.060)		
Career PLUS :								
Net Impact of Frame	5.308** (2.345)	4.624** (2.244)			0.115* (0.060)	0.110* (0.058)		
Peer and Career			-0.748 (2.578)	-1.073 (2.647)			-0.016 (0.063)	-0.034 (0.057)
PLUS: Net Impact of Frame			5.629*** (1.878)	6.120*** (2.034)			0.117*** (0.044)	0.116** (0.045)
Peer PLUS*	6.846**	6.526**			0.157**	0.138**		
Career PLUS*	3.097	2.058			0.050	0.029		
PLUS*			4.881*	5.048*			0.100*	0.082
Observations	3072	3061	3072	3071	3072	3055	3072	3055
Adjusted R <sup>2</sup>	0.042	0.074	0.037	0.064	0.056	0.075	0.047	0.069
PDS LASSO controls	No	Yes	No	Yes	No	Yes	No	Yes
District Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Errors clustered at the training session level which is the unit of randomization. The dependent variable is overconfidence. In the first four columns, its is constructed as a continuous variable that is the difference between teacher beliefs of how well they scored on the test and actual test score at endline. In the last four columns, we construct a dummy variable of above median overconfidence based on the continuous variable. All regressions include district FE. Controls include trainee-level teacher controls, master trainer controls, and enumerator controls. The PLUS treatments with the asterisks present the overall impact of the treatments (Incentive + the frame). Estimates are significant at the \*10%, \*\*5%, and \*\*\*1% level.

Table 8: Treatment Effects on School-Level Outcomes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	% Teacher Absent	% Classrooms Used	% Student Attendance	No. of Sc Meetings	Cleanliness (=1 if clean)	Report Card (=1 if display)	No. Of Coaches
Peer	-0.000* (0.000)	-0.009 (0.007)	0.004* (0.002)	0.034 (0.041)	-0.005 (0.059)	0.007 (0.017)	-0.059** (0.024)
Career	-0.000 (0.000)	-0.009 (0.007)	-0.001 (0.002)	0.025 (0.043)	-0.028 (0.055)	-0.016 (0.021)	-0.007 (0.027)
Peer PLUS : Net Impact of Frame	0.000 (0.000)	-0.007 (0.008)	-0.004** (0.002)	-0.044 (0.033)	-0.057 (0.055)	-0.003 (0.013)	0.051* (0.030)
Career PLUS : Net Impact of Frame	0.000* (0.000)	0.008 (0.008)	0.001 (0.002)	-0.025 (0.042)	-0.016 (0.045)	0.015 (0.019)	0.000 (0.034)
Peer PLUS*	-0.003	-0.016	0.000	-0.010	-0.061	0.005	-0.008
Career PLUS*	0.003	-0.001	0.000	-0.001	-0.0436	-0.001	-0.006
Observations	3322	3322	3322	3322	3322	3316	3322
Adjusted R <sup>2</sup>	0.022	0.485	0.163	0.022	0.142	0.028	0.337
PDS LASSO controls	No	No	No	No	No	No	No
District Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: All regressions are an ANCOVA estimation with baseline values of the dependent variable and with district FE. Errors are clustered at the training session level. The PLUS treatments with the asterisks present the overall impact of the treatments (Incentive + the frame). Estimates are significant at the \*10%, \*\*5%, and \*\*\*1% level.

## References

- Abadie, Alberto, Susan Athey, Guido W Imbens, and Jeffrey Wooldridge. 2017. *When Should You Adjust Standard Errors for Clustering?* [electronic resource]. Working paper series (National Bureau of Economic Research : Online) ; working paper no.24003. Cambridge, Mass.: National Bureau of Economic Research.
- Acharya, Avidit et al. 2016. "Explaining Causal Findings Without Bias: Detecting and Assessing Direct Effects." 110 (3):512–529.
- Almlund, Mathilde et al. 2011. "Personality Psychology and Economics." *NBER Working Paper Series* :16822URL <http://search.proquest.com/docview/1687830854/>.
- Amabile, Teresa M., Karl G. Hil, Beth A. Hennessey, and Elizabeth M. Tighe. 1995. "'The Work Preference Inventory: Assessing intrinsic and extrinsic motivational orientations': Correction." *Journal of Personality and Social Psychology* 68 (4):580–580.
- Andreoni, James. 1990. "Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving." *The Economic Journal* 100 (401):462. URL <http://search.proquest.com/docview/198090143/>.
- ASER. 2019. "Annual Status of Education Report - ASER-PAKISTAN 2019." URL <https://palnetwork.org/wp-content/uploads/2020/02/Annual-Status-of-Education-Report-ASER-PAKISTAN-2019.pdf>.
- Ashraf, Nava, Oriana Bandiera, and B. Kelsey Jack. 2014. "No margin, no mission? A field experiment on incentives for public service delivery." *The Journal of Public Economics* 120:1.
- Ashraf, Nava, Oriana Bandiera, and Scott Lee. 2014. "Awards unbundled: evidence from a natural field experiment." *Journal of economic behavior and organization* 100:44–63. URL <http://search.proquest.com/docview/1523807233/>.
- Bandura, Albert. 1986. *Social foundations of thought and action : a social cognitive theory*. Prentice-Hall series in social learning theory. Englewood Cliffs, N.J.: Prentice-Hall.
- Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukherji, Marc Shotland, and Michael Walton. 2016. "Mainstreaming an Effective Intervention: Evidence from Randomized Evaluations of "Teaching at the Right Level" in India." *NBER Working Paper Series* :22746URL <http://search.proquest.com/docview/1831255938/>.

- Barrera-Osorio, Felipe and Dhushyanth Raju. 2017. "Teacher performance pay: Experimental evidence from Pakistan." *Journal of Public Economics* 148 (C):75–91.
- Bau, Natalie and Jishnu Das. 2020. "Teacher Value Added in a Low-Income Country †." *American Economic Journal: Economic Policy* 12 (1):62–96.
- Baumeister, Roy F. 1999. *The self in social psychology*. Key readings in social psychology. Philadelphia, PA ; Hove: Psychology Press.
- Beaman, Lori, Esther Duflo, Rohini Pande, and Petia Topalova. 2012. "Female leadership raises aspirations and educational attainment for girls: a policy experiment in India." *Science (New York, N.Y.)* 335 (6068):582.
- Belloni, Alexandre et al. 2014. "Pivotal Estimation Via Square-Root LASSO in Non-Parametric Regression." *The Annals of Statistics* 42 (2):757–788.
- Besley, Timothy and Maitreesh Ghatak. 2005. "Competition and Incentives with Motivated Agents." *American Economic Review* 95 (3):616–636.
- . 2008. "Status Incentives." *American Economic Review* 98 (2):206–211.
- Bold, Tessa, Deon Filmer, Gayle Martin, Ezequiel Molina, Christophe Rockmore, Brian Stacy, Jakob Svensson, and Waly Wane. 2017. "What Do Teachers Know and Do? Does It Matter? : Evidence from Primary Schools in Africa."
- Bénabou, Roland and Jean Tirole. 2002. "Self-Confidence and Personal Motivation." *The Quarterly Journal of Economics* 117 (3):871–915.
- . 2003. "Intrinsic and Extrinsic Motivation." *Review of Economic Studies* 70 (3):489–520.
- Callen, Michael et al. 2016. "The Political Economy of Public Sector Absence: Experimental Evidence from Pakistan." *NBER Working Paper Series* :22340URL <http://search.proquest.com/docview/1795921414/>.
- Chetty, Nadarajan et al. 2014. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." 104 (9).
- Das, Jishnu et al. 2012. "Learning Levels and Gaps in Pakistan: A Comparison with Uttar Pradesh and Madhya Pradesh." *Economic and Political Weekly* 47 (26/27):228–240.
- de Ree, Joppe, Karthik Muralidharan, Menno Pradhan, and Halsey Rogers. 2018. "Double for Nothing? Experimental Evidence on an Unconditional Teacher Salary Increase in Indonesia\*." *The Quarterly Journal of Economics* 133 (2):993–1039.

- Deci, Edward et al. 1999. "A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation." *Psychological Bulletin* 125 (6):627–668. URL <http://search.proquest.com/docview/203472909/>.
- Dewatripont, Mathias et al. 1999. "The economics of career concerns, Part I: comparing information structures." *Review of economic studies* 66(1) (226):183–198. URL <http://search.proquest.com/docview/38702024/>.
- Duflo, Esther et al. 2012. "Incentives Work: Getting Teachers to Come to School." *American Economic Review* 102 (4):1241–1278.
- Eden, Dov and Arie Aviram. 1993. "Self-Efficacy Training to Speed Reemployment: Helping People to Help Themselves." *Journal of Applied Psychology* 78 (3):352. URL <http://search.proquest.com/docview/1290321664/>.
- Fackler, Sina and Lars-Erik Malmberg. 2016. "Teachers' self-efficacy in 14 OECD countries: Teacher, student group, school and leadership effects." *Teaching and Teacher Education* 56.
- Frank, Robert H. 1985. *Choosing the right pond : human behavior and the quest for status*. New York ; Oxford: Oxford University Press.
- Frey, Bruno S. 1997. *Not just for the money : an economic theory of personal motivation*. Cheltenham: Edward Elgar.
- Frey, BS and S Neckermann. 2008. "Awards A View from Psychological Economics." *Zeitschrift Fur Psychologie-Journal Of Psychology* 216 (4):198–208.
- Gauri, Varun, Julian Jamison, Nina Mazar, Owen Ozier, Shomikho Raha, and Karima Saleh. 2018. "Motivating Bureaucrats Through Social Recognition: Evidence from Simultaneous Field Experiments." URL <http://search.proquest.com/docview/2063163854/>.
- Glewwe, P. and K. Muralidharan. 2016. "Improving Education Outcomes in Developing Countries: Evidence, Knowledge Gaps, and Policy Implications." In *Handbook of the Economics of Education*, vol. 5. Elsevier, 653–743.
- Glewwe, Paul et al. 2010. "Teacher Incentives." *American Economic Journal: Applied Economics* 2 (3):205–227.
- Haushofer, J, A John, and K Orkin. 2019. "Can Simple Psychological Interventions Increase Preventive Health Investment?" .

- Holmstrom, Bengt and Paul Milgrom. 1991. "Multitask Principal - Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design." *Journal of Law, Economics, and Organization* 7:24. URL <http://search.proquest.com/docview/1300226943/>.
- I-SAP. 2017. "Public Financing of Education in Pakistan: 2010-11 to 2016-17." URL [http://i-saps.org/upload/report\\_publications/docs/1496496299.pdf](http://i-saps.org/upload/report_publications/docs/1496496299.pdf).
- John, Oliver et al. 2008. *Paradigm shift to the integrative big five trait taxonomy: History, measurement, and conceptual issues*, vol. Vol. 3. 114–158.
- Kosfeld, Michael and Susanne Neckermann. 2011. "Getting More Work for Nothing? Symbolic Awards and Worker Performance." *American Economic Journal: Microeconomics* 3 (3):86–99.
- Kosfeld, Michael, Susanne Neckermann, and Xiaolan Yang. 2014. "Knowing that You Matter, Matters! The Interplay of Meaning, Monetary Incentives, and Worker Recognition." *SSRN Electronic Journal* .
- Lazear, Edward and Sherwin Rosen. 1981. "Rank-Order Tournaments as Optimum Labor Contracts." *The Journal of Political Economy* 89 (5):841. URL <http://search.proquest.com/docview/1290576073/>.
- Lee, Scott. 2018. "Intrinsic Incentives: A Field Experiment on Leveraging Intrinsic Motivation in Public Service Delivery." *SSRN Electronic Journal* .
- Luthans, F. and A.D. Stajkovic. 1999. "Reinforce for performance: The need to go beyond pay and even rewards." *Academy of Management Executive* 13 (2):49–57.
- Mansoor, Zahra. 2019. "The Double-Edged Sword of Non-Financial Incentives (Pilot- Part I): Evidence from the Education Sector in Pakistan." :22746 URL <http://search.proquest.com/docview/1831255938/>.
- Markham, Steven E. et al. 2002. "Recognizing Good Attendance: A Longitudinal, Quasi-Experimental Field Study." *Personnel Psychology* 55 (3):639–660.
- Mullainathan, Sendhil and Eldar Shafir. 2013. *Scarcity : why having too little means so much*. London.
- Muralidharan, Karthik and Venkatesh Sundararaman. 2011. "Teacher Performance Pay: Experimental Evidence from India." *Journal of Political Economy* 119 (1):39–77.
- Perry, James. 1996. "Measuring public service motivation: an assessment of construct reliability and validity." *Journal of public administration research and theory* 6 (1):5–24. URL <http://search.proquest.com/docview/39010044/>.

- Popova, Anna et al. 2016. "Training Teachers on the Job : What Works and How to Measure It."
- Rammstedt, Beatrice and Oliver P John. 2007. "Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German." *Journal of Research in Personality* 41 (1):203–212.
- Rivkin, Steven G. et al. 2005. "Teachers, Schools, and Academic Achievement." *Econometrica* 73 (2):417–458.
- Rockoff, Jonah E. 2004. "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data." *American Economic Review* 94 (2):247–252.
- Soto, Christopher J and Oliver P John. 2017. "Short and extra-short forms of the Big Five Inventory–2: The BFI-2-S and BFI-2-XS." *Journal of Research in Personality* 68 (C):69–81.
- Swann, William. 1996. *Self-Traps: The Elusive Quest for Higher Self-Esteem*. W H Freeman Co; First Printing edition (April 1, 1996).
- Tanguy, B, Stefan Dercon, K Orkin, and AS Taffesse. 2014. "The future in mind: Aspirations and forward-looking behaviour in rural Ethiopia." .
- Tanwir, Maryam. 2014. "Gender neutrality and the Pakistani bureaucracy." *Journal of International Women's Studies* 15 (2):143.
- Zee, Marjolein and Helma M. Y Koomen. 2016. "Teacher Self-Efficacy and Its Effects on Classroom Processes, Student Academic Adjustment, and Teacher Well-Being: A Synthesis of 40 Years of Research." *Review of Educational Research* 86 (4):981–1015.

# Appendix

## Appendix A: Tables

## A. 1: Descriptive Statistics

	(1)	(2)	(3)	(4)	(5)
	Mean	Sd	p0.25	p0.50	p0.75
<b>Basic teacher characteristics</b>					
Age	45.54	10.31	37	49	54
Gender (=1 if male)	0.43				
Salary	77604.47	31779.54	51000	71000	97328
Years of experience	19.99	10.94	10	22	30
Years of education	15.72	0.83	16	16	16
Married (=1 if married)	0.90				
Total teachers in a session	27.38	6.48	23	26	31
<b>Basic job characteristics</b>					
Job Grade	15.53	2.58	15	16	17
Time till next promotion (in yrs)	6.06	4.83	2	5	10
HT's school's enrollment capacity	467.05	480.86	189	317	555
School Location of HT (=1 if urban)	0.23				
<b>Baseline Performance</b>					
<i>normalised</i>					
Pre Test Scores (Incentivized)	-0.15	1.01	-0.78	-0.23	0.32
Pre Test Scores (Overall)	-0.02	1.03	-0.57	-0.07	0.44
<i>Percent</i>					
Pre Test Scores (Overall)	33.92	13.56	26.67	33.33	40.00
<b>Non-cognitive traits</b>					
<i>Personality traits &amp; Self-efficacy</i>					
BFI Index	0.01	0.55	-0.32	0.02	0.35
Openness	0.01	1.00	-0.63	0.02	0.68
Extraversion	0.01	1.00	-0.71	-0.13	1.04
Conscientiousness	0.01	1.10	-0.07	-0.07	0.27
Agreeableness	-0.00	1.00	-0.77	0.00	0.76
Neuroticism	-0.00	0.99	-0.93	0.31	0.93
Self-Efficacy	-0.01	0.99	-0.68	-0.12	0.92
<i>Motivational Orientation</i>					
Extrinsic Motivation	0.25				
Intrinsic Motivation	0.41				
Pro-social Motivation	0.31				
<i>Other intrinsic measures</i>					
PSM Index	0.00	0.38	-0.25	-0.01	0.25
Donation in hypothetical game (total PKR 10,000)	4052	2876	2000	4000	5000
Observations	3394				

Notes: Pretest scores, overall personality index, each individual personality trait, and self-efficacy are normalized against the control group.

## A. 2: Randomization Balance - All Treatments

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
	Control	Peer	Career	Peer + Career	Career + C-Peer	C-Career	C-Peer + C-Career	Peer-Career	Peer-Peer + Peer-Career	Peer-Career	Peer-Peer + Peer-Career	Peer-Career	Peer-Peer + Peer-Career
Age	44.79 (0.90)	46.70 (0.88)	46.55 (0.83)	46.16 (0.97)	45.39 (0.97)	0.03**	0.03**	0.15	0.53	0.96	0.58	0.19	0.18
Gender (=1 if male)	0.42 (0.05)	0.57 (0.05)	0.50 (0.05)	0.56 (0.06)	0.45 (0.06)	0.03**	0.12	0.03**	0.67	0.26	0.94	0.05**	0.39
Salary	69009 (3578)	74779 (3408)	77471 (3670)	74336 (3403)	70669 (3403)	0.07	0.02**	0.11	0.61	0.45	0.89	0.19	0.07
Years of Education	15.66 (0.06)	15.72 (0.06)	15.73 (0.06)	15.72 (0.06)	15.73 (0.06)	0.30	0.17	0.29	0.19	0.74	0.95	0.74	0.99
Married (=1 if married)	0.91 (0.02)	0.95 (0.02)	0.93 (0.02)	0.93 (0.02)	0.94 (0.02)	0.09	0.47	0.57	0.14	0.17	0.16	0.43	0.52
<b>Basic job characteristics</b>													
Time till next promotion (in yrs)	6.05 (0.44)	6.05 (0.48)	5.93 (0.48)	6.36 (0.43)	5.97 (0.43)	1.00	0.75	0.44	0.81	0.75	0.44	0.81	0.91
HT's school's enrollment capacity	237 (26.65)	267 (26.55)	330 (35.78)	256 (31.67)	246 (31.67)	0.32	0.04**	0.64	0.79	0.15	0.77	0.51	0.06
School Location of HT (=1 if urban)	0.11 (0.03)	0.15 (0.04)	0.17 (0.04)	0.07 (0.03)	0.08 (0.03)	0.35	0.19	0.30	0.39	0.68	0.08	0.11	0.07
<b>Baseline Performance</b>													
Pre Test (Incentivized)	-0.11 (0.15)	-0.13 (0.17)	-0.18 (0.15)	-0.28 (0.14)	-0.33 (0.14)	0.87	0.51	0.14	0.08	0.72	0.29	0.15	0.18
Pre Test (Overall)	-0.04 (0.13)	0.02 (0.15)	-0.05 (0.13)	-0.09 (0.12)	-0.14 (0.14)	0.70	0.84	0.64	0.33	0.58	0.45	0.24	0.40
<b>Non-Cognitive Traits</b>													
Overall BFI Index	0.09 (0.04)	0.05 (0.03)	0.06 (0.04)	0.08 (0.04)	0.03 (0.04)	0.21	0.34	0.69	0.07	0.78	0.46	0.57	0.39
Self-efficacy Index	-0.04 (0.07)	-0.03 (0.07)	-0.03 (0.07)	-0.04 (0.07)	-0.02 (0.07)	0.87	0.95	0.99	0.80	0.91	0.86	0.93	0.82
Intrinsic Motivation	0.51 (0.04)	0.48 (0.04)	0.52 (0.04)	0.50 (0.04)	0.51 (0.04)	0.35	0.73	0.90	0.97	0.22	0.53	0.32	0.75
Extrinsic Motivation	0.20 (0.03)	0.21 (0.03)	0.17 (0.03)	0.20 (0.03)	0.21 (0.03)	0.61	0.26	0.80	0.51	0.06	0.80	0.88	0.06
Pro-social Motivation	0.27 (0.03)	0.29 (0.03)	0.29 (0.03)	0.28 (0.03)	0.26 (0.03)	0.54	0.52	0.90	0.56	0.97	0.67	0.27	0.23
Joint F-Test						0.27	0.54	0.70	0.30	0.94	0.75	0.53	0.11
Observations	716	649	687	635	707								

Notes: The first five columns report the mean and standard errors of the four recognition treatments and the control group. The last eight columns show equality of means between the control group and the treatment group, and between each treatment, for each variable of interest. Estimates are significant at the \*\*5%, and \*\*\*1% level

### A. 3: Attrited Sample and Treatments

	(1)
	Attrited(=1 if sample attrited)
Peer	-0.007 (0.01)
Career	-0.011 (0.01)
Peer PLUS	-0.010 (0.01)
Career PLUS	-0.013 (0.01)
Observations	3493
R-squared	0.223
PDS LASSO Controls	No
District FE	Yes

Notes: Errors clustered at the training session level which is the unit of randomization. Estimates are significant at the \*10%, \*\*5%, and \*\*\*1% level.

#### A. 4: Balance across Attrited and Main Sample

	(1)	(2)	(3)
	Attrited Sample	Main Sample	P-value difference
<b>Basic teacher characteristics</b>			
Age	47.04 (1.51)	45.85 (0.69)	0.40
Gender (=1 if male)	0.43 (0.07)	0.50 (0.04)	0.23
Salary	77039 (5158)	73151 (2883)	0.39
Years of Experience	21.08 (1.69)	20.10 (0.90)	0.52
Years of Education	15.75 (0.11)	15.72 (0.05)	0.75
Married (=1 if married)	0.92 (0.03)	0.93 (0.04)	0.66
<b>Basic job characteristics</b>			
Time till next promotion (in yrs)	4.64 (0.82)	6.10 (0.38)	0.04**
HT's school's enrollment capacity	237 (26.65)	267 (26.55)	0.32
School Location of HT (=1 if urban)	0.11 (0.03)	0.15 (0.04)	0.35
<b>Baseline Performance</b>			
Pre Test Scores (normalised)	-0.09 (0.18)	-0.20 (0.14)	0.39
<b>Non-Cognitive Traits</b>			
Overall BFI Index	0.01 (0.07)	0.06 (0.03)	0.50
Self-Efficacy Index	0.13 (0.12)	-0.03 (0.06)	0.15
Intrinsic Motivation	0.42 (0.06)	0.50 (0.03)	0.11
Extrinsic Motivation	0.23 (0.06)	0.19 (0.02)	0.55
Pro-social Motivation	0.34 (0.06)	0.28 (0.02)	0.28
Joint F			0.81
Observations	100	3394	

Notes: Errors are clustered at the training session level which is the unit of randomization. The first two columns present the means for the attrited and the main sample, whereas the third column presents the p-value difference for each variable of interest. Estimates are significant at the \*10%, \*\*5%, and \*\*\*1% level.

A. 5: Non-Consenting Sample and Treatments

	(1)
	Non-Consent(=1 if did not consent)
Peer	0.004 (0.05)
Career	-0.039 (0.05)
Peer PLUS	0.008 (0.05)
Career PLUS	0.030 (0.05)
Observations	3394
R-squared	0.029
PDS LASSO Controls	No
District FE	Yes

Notes: Errors clustered at the training session level which is the unit of randomization. Estimates are significant at the \*10%, \*\*5%, and \*\*\*1% level.

## A. 6: Balance across Non-Consenting and Consenting Trainees

	(1)	(2)	(3)
	Non-consenting Sample	Consenting Sample	P-value difference
<b>Basic teacher characteristics</b>			
Age	48.54 (1.15)	45.89 (0.69)	0.01***
Gender (=1 if male)	0.53 (0.08)	0.50 (0.04)	0.66
Salary	80358 (5034)	73162 (2881)	0.11
Years of Experience	23.16 (1.25)	20.12 (0.89)	0.00***
Years of Education	15.68 (0.08)	15.71 (0.05)	0.62
Married (=1 if married)	0.93 (0.02)	0.93 (0.01)	0.94
<b>Basic job characteristics</b>			
Time till next promotion (in yrs)	6.45 (0.48)	6.08 (0.38)	0.34
HT's school's enrollment capacity	237 (26.65)	267 (26.55)	0.32
School Location of HT (=1 if urban)	0.11 (0.03)	0.15 (0.04)	0.35
<b>Baseline Performance</b>			
Pre Test Scores (normalised)	-0.36 (0.16)	-0.21 (0.14)	0.05**
<b>Non-Cognitive Traits</b>			
Overall BFI Index	0.04 (0.05)	0.06 (0.03)	0.72
Self-Efficacy Index	-0.11 (0.09)	-0.02 (0.06)	0.26
Intrinsic Motivation	0.44 (0.06)	0.49 (0.03)	0.23
Extrinsic Motivation	0.20 (0.04)	0.21 (0.03)	0.72
Pro-social Motivation	0.33 (0.04)	0.27 (0.02)	0.08
Observations	207	3187	

Notes: Errors are clustered at the training session level which is the unit of randomization. The first two columns present the means for the non-consenting and consenting sample, whereas the third column presents the p-value difference for each variable of interest. Estimates are significant at the \*10%, \*\*5%, and \*\*\*1% level.

### A. 7: Treatment Effects on Other Secondary Outcomes

	Intrinsic Motivation		External Locus		Department Credibility	
	(1)	(2)	(3)	(4)	(5)	(6)
Peer	0.040 (0.045)	0.069 (0.042)	-0.031 (0.062)	-0.075 (0.066)	0.055 (0.065)	0.042 (0.061)
Career	-0.011 (0.048)	0.002 (0.042)	0.009 (0.061)	-0.022 (0.064)	0.060 (0.069)	0.068 (0.062)
Peer PLUS : Net Impact of Frame	-0.036 (0.046)	-0.023 (0.049)	-0.078 (0.066)	-0.054 (0.071)	0.051 (0.061)	0.081 (0.061)
Career PLUS : Net Impact of Frame	0.098* (0.054)	0.092* (0.049)	-0.114 (0.073)	-0.131* (0.071)	0.121 (0.078)	0.085 (0.067)
Observations	3316	3316	3286	3286	3303	3303
Adjusted R <sup>2</sup>	0.179	0.207	0.168	0.187	0.004	0.122
PDS LASSO controls	No	Yes	No	Yes	No	Yes
District Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Errors clustered at the training session level which is the unit of randomization. All regressions are an ANCOVA estimation with baseline values of the dependent variable (except department credibility for which we did not have a baseline) and district FE. Controls include trainee-level teacher controls, master trainer controls, and enumerator controls. All dependent variables are normalized by the mean and standard deviation of the control group. Estimates are significant at the \*10%, \*\*5%, and \*\*\*1% level.

A. 8: Correlations between Actual and Predicted Overconfidence

	(1)	(2)
	Predicted Overconfidence	Actual Overconfidence
Actual overconfidence	0.31	1.00
Post Test Scores	-0.37	-0.73

Notes: Predicted overconfidence is estimated by predicting actual overconfidence through LASSO. Actual overconfidence is constructed as a continuous variable that is the difference between teacher beliefs of how well they scored on the test and actual post-test score at endline.

A. 9: Treatment Effects on Predicted Overconfidence

	(1)	(2)	(3)	(4)
Peer	0.794 (0.694)	-0.449 (0.387)		
Career	0.477 (0.666)	-0.213 (0.386)		
Peer PLUS : Net Impact of Frame	2.065*** (0.660)	1.503*** (0.368)		
Career PLUS : Net Impact of Frame	1.665** (0.669)	0.920** (0.386)		
Peer and Career			0.642 (0.588)	-0.090 (0.486)
PLUS: Net Impact of Frame			1.844*** (0.492)	1.869*** (0.395)
Observations	2963	2953	2963	2963
Adjusted R <sup>2</sup>	0.297	0.662	0.296	0.600
PDS LASSO controls	No	Yes	No	Yes
District Fixed Effects	Yes	Yes	Yes	Yes

Notes: Errors clustered at the training session level which is the unit of randomization. The dependent variable is predicted overconfidence. All regressions include district FE. Controls include trainee-level teacher controls, master trainer controls, and enumerator controls. Estimates are significant at the \*10%, \*\*5%, and \*\*\*1% level.

A. 10: Mediation Analysis

	(1)	(2)
	Post-Test Score	De-mediated Post-Test Score
Peer and Career	0.183 (0.137)	0.133 (0.131)
PLUS: Net Impact of Frame	-0.242** (0.106)	-0.108 (0.108)
Observations	3392	2938
Adjusted R <sup>2</sup>	0.141	0.083
PDS LASSO controls	Yes	Yes
District Fixed Effects	Yes	Yes

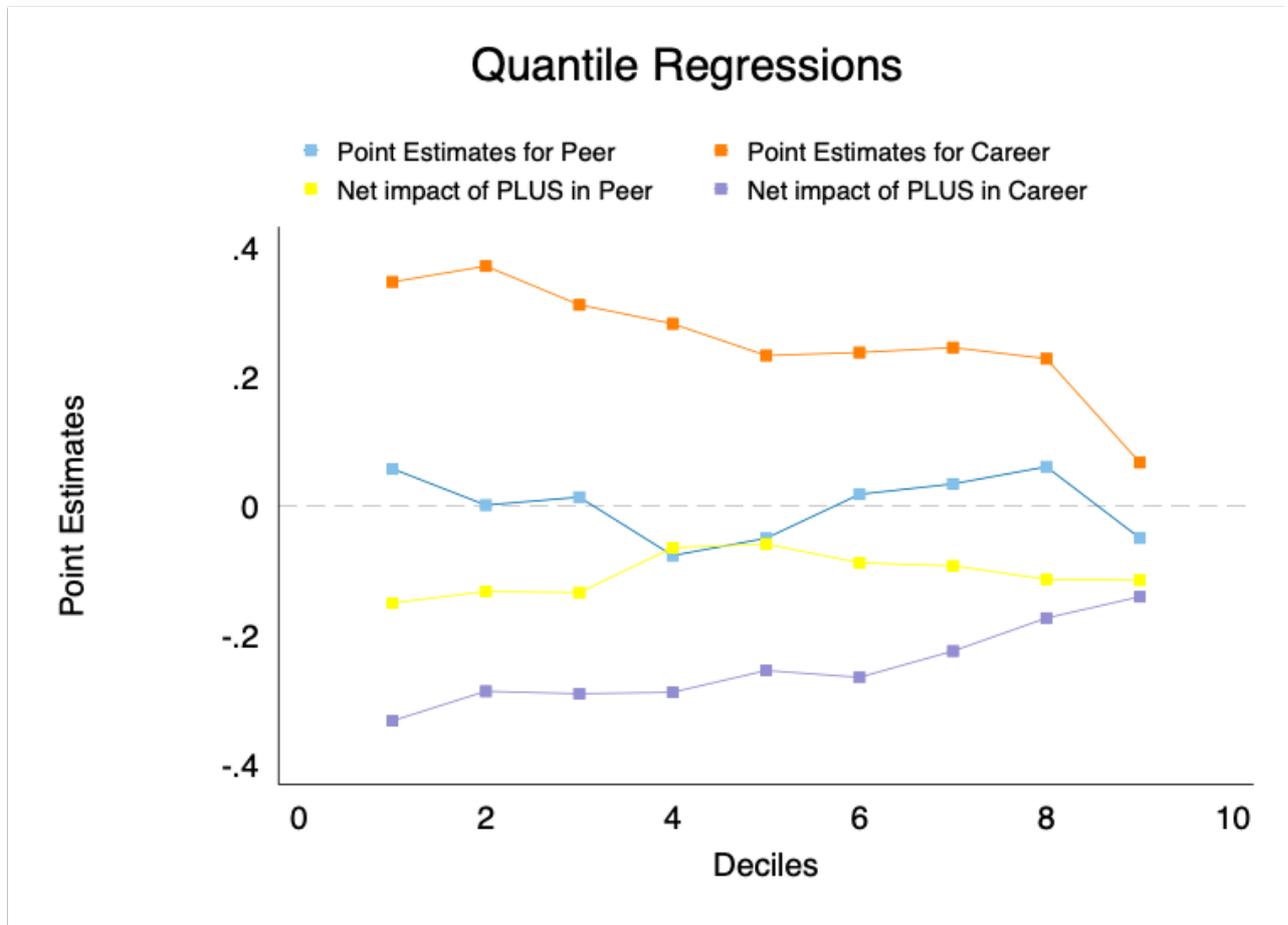
Notes: Errors are clustered at the training session level which is the unit of randomization. Both regressions are an ANCOVA estimation with the baseline value of the dependent variable and district FE. In column 1, the dependent variable is the post-test score. In column 2, the dependent variable is the de-mediated post-test score which is calculated by: 1) regressing the main outcome on treatment, pre-treatment controls, the mediator, interaction between the mediator and treatment, and interaction between the mediator and all other pre-treatment variables; 2) calculating the de-mediated post-test scores which is the predicted outcome excluding all coefficients that include the mediator fixed at a specific value. Controls include trainee-level teacher controls, master trainer controls, and enumerator controls. Estimates are significant at the \*10%, \*\*5%, and \*\*\*1% level.

### A. 11: Heterogeneous Treatment Effects - Post Test Scores and Overconfidence

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Post Test	Overconf.	Post Test	Overconf.	Post Test	Overconf.	Post Test	Overconf.
Male x Peer + Career	0.056 (0.154)	2.115 (3.179)						
Female x Peer + Career	0.213 (0.140)	-3.447 (2.827)						
Male x Net PLUS	-0.302** (0.139)	5.774** (2.631)						
Female x Net PLUS	-0.141 (0.099)	4.855** (2.011)						
Below Med. Ability x Peer + Career			0.215 (0.173)	-3.214 (3.064)				
Above Med. Ability x Peer + Career			0.013 (0.115)	1.775 (2.325)				
Below Med. Ability x PLUS			-0.226* (0.116)	5.983*** (2.068)				
Above Med. Ability x PLUS			-0.175 (0.110)	4.059* (2.072)				
Below Med. Experience x Peer + Career					0.070 (0.137)	0.225 (2.633)		
Above Med. Experience x Peer + Career					0.210 (0.145)	-2.626 (2.780)		
Below Med. Experience x PLUS					-0.164 (0.111)	3.735* (1.972)		
Above Med. Experience x PLUS					-0.243** (0.112)	6.494*** (2.124)		
Below Med. Personality x Peer + Career							0.099 (0.140)	0.426 (2.872)
Above Med. Personality x Peer + Career							0.170 (0.137)	-2.278 (2.365)
Below Med. Personality x PLUS							-0.216* (0.116)	5.603*** (2.089)
Above Med. Personality x PLUS							-0.198** (0.100)	4.473** (1.882)
Observations	3394	3072	3394	3072	3394	3072	3382	3062
Adjusted R <sup>2</sup>	0.138	0.067	0.095	0.053	0.108	0.055	0.099	0.075
PDS LASSO controls	No	No	No	No	No	No	No	No
District Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

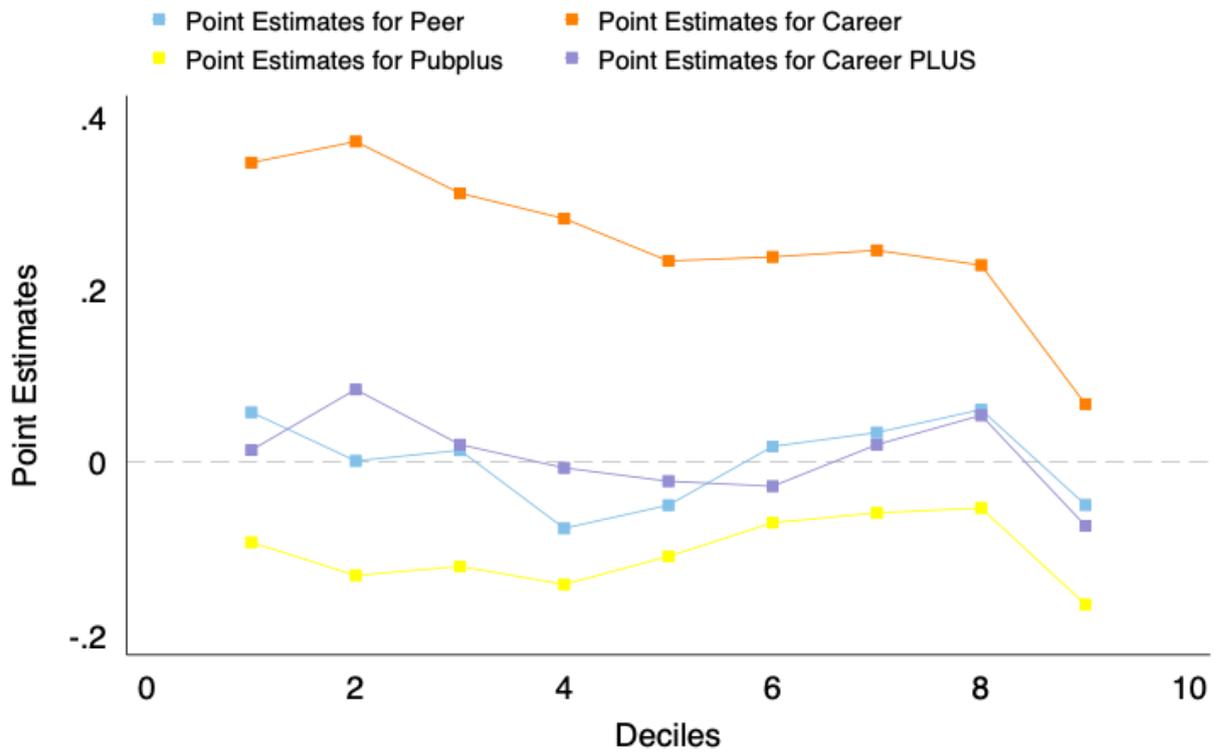
Notes: Errors clustered at the training session level which is the unit of randomization. All regressions include district FE. Controls include trainee-level teacher controls, master trainer controls, and enumerator controls. Estimates are significant at the \*10%, \*\*5%, and \*\*\*1% level.

## Appendix B: Figures

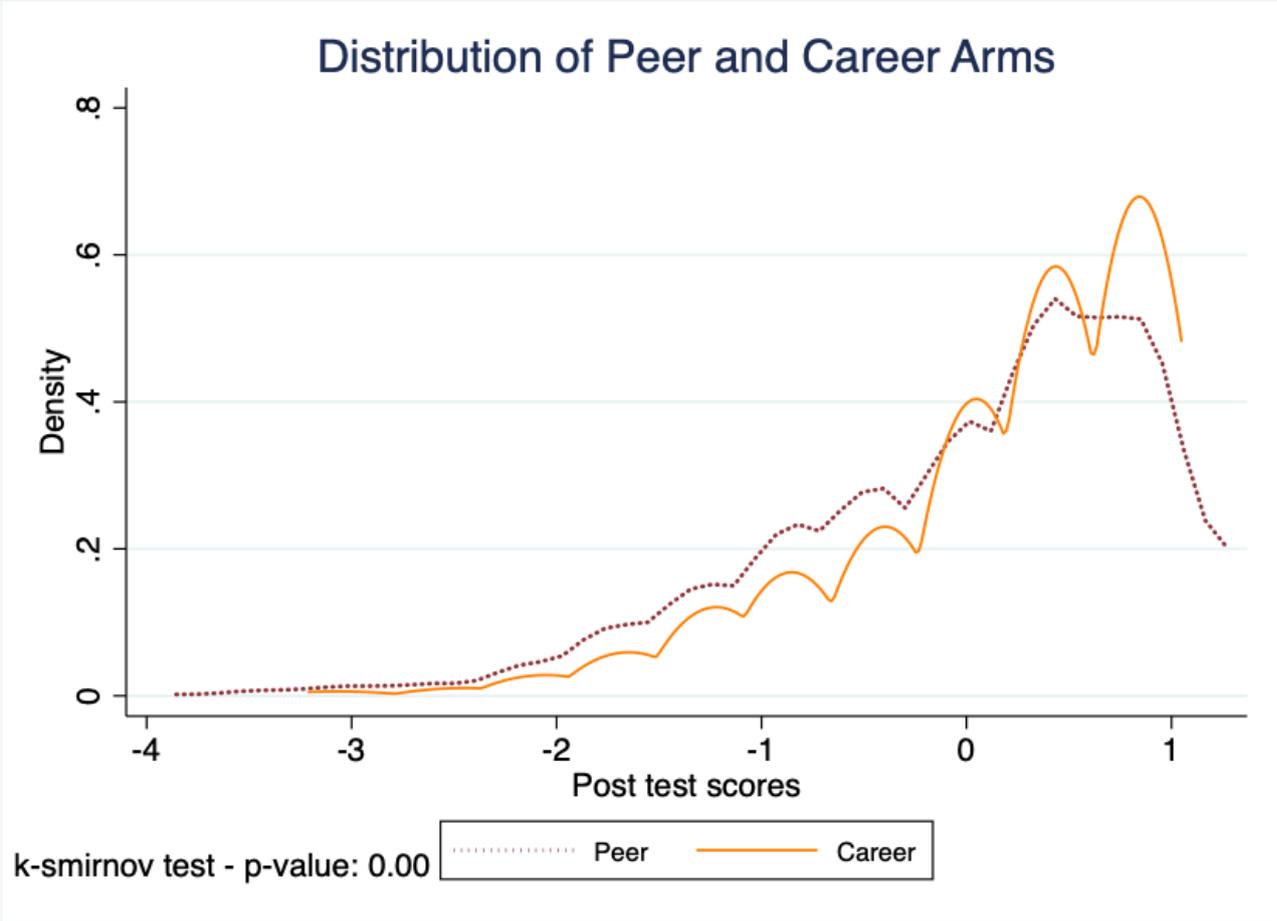


B. 1: Quantile Regression Estimates

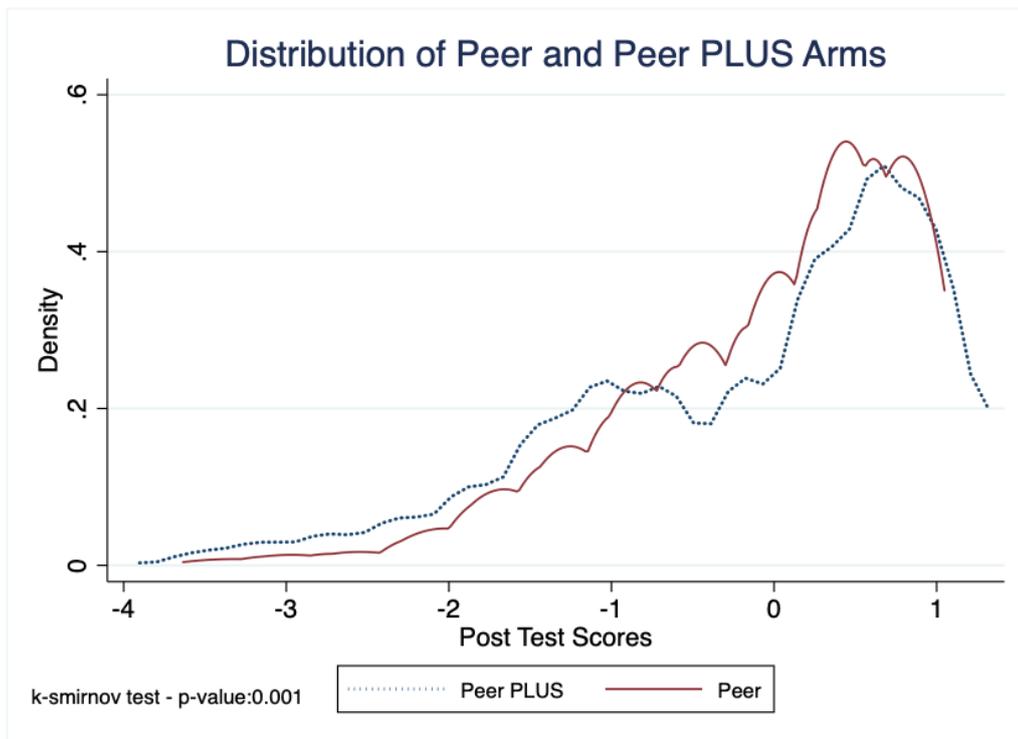
## Quantile Regressions



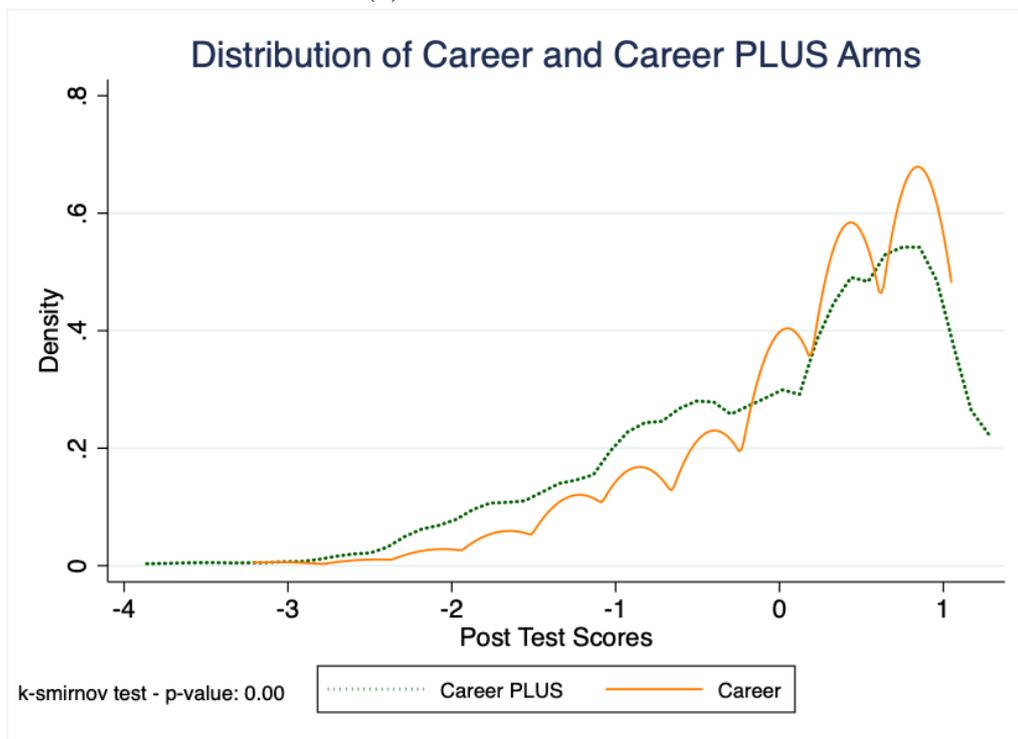
B. 2: Quantile Regression Estimates



B. 3: K smirnov-test: Peer and Career Distribution

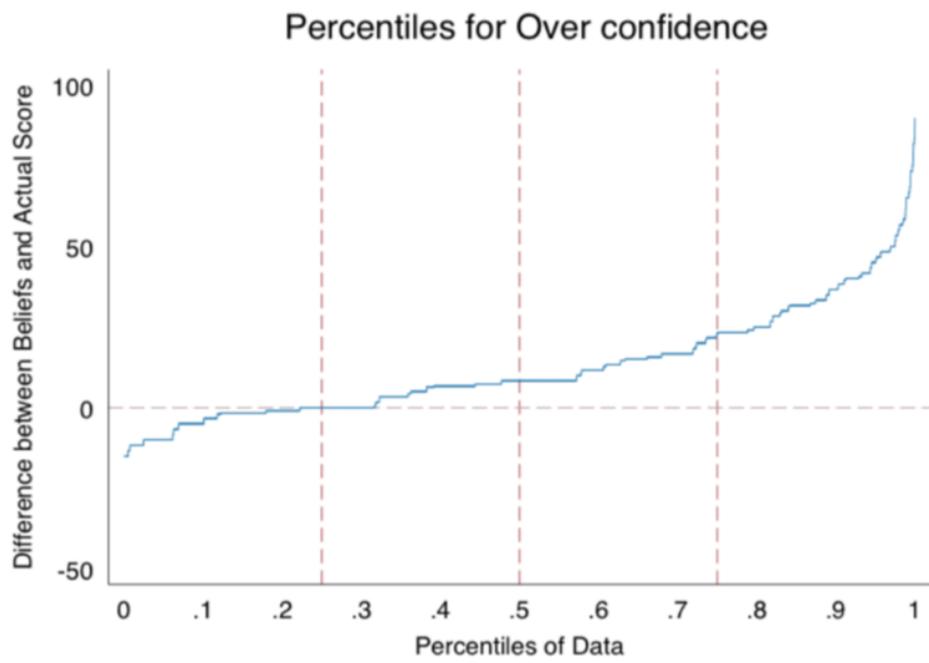


(a) Peer and Peer PLUS



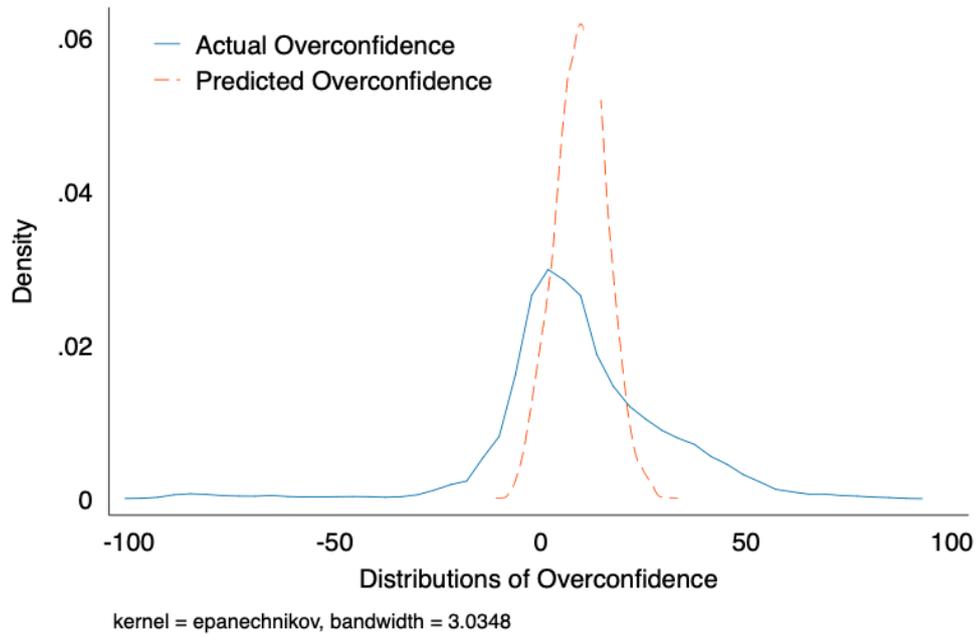
(b) Career and Career PLUS

B. 4: K smirnov-test: Peer/Career and PLUS Counterparts



B. 5: Qplot for Overconfidence

### Distribution of Actual and Predicted Overconfidence



B. 6: Distribution of Actual and Predicted Overconfidence

## Appendix C: Training Details

Training Topics	
<b>sr.no</b>	<b>Topic</b>
1	Power of Coaching
2	Student Leadership
3	Pupil Voice
4	Protecting Children
5	Staff and Distributed Leadership
6	Co-curricular Activities
7	Staff Leave Rules

## Appendix D: Training Test Sample

Date: <input type="text"/> <input type="text"/> / <input type="text"/> <input type="text"/> / <input type="text"/> <input type="text"/>	School emiscode: <input type="text"/> <input type="text"/>	Teacher Name: _____ _____ Teacher CNIC: _____ _____
--	--	--

**Marks: 15**

**Time: 15 minutes**

### Section A:

***Encircle the right option for the given statements / questions.***

Q1.	In SHEEP Model "S" stands for ----- a) Size    b) Score    c) Stay Safe    d) Scale
Q2.	The most appropriate way of pupil voice in School is: a) Student Council    b) Sports Team    c) Monitors    d) Focus Group
Q3.	When a civil servant completes his continuous service of more than 10 years he may be granted extraordinary leave at a time for the maximum period of: a) Two years    b) Three years    c) Four years    d) Five years
Q4.	In Pupil Voice SLT stand for: a) Super Leading Team    b) Student Leading Team c) Senior Leadership Team    d) Student Learning Team
Q5.	Delayed development of the child is: a) Physical abuse    b) Emotional abuse    c) Sexual abuse    d) Neglect
Q6.	The 4MAT system or (4 Mode Application Techniques ) was developed by Mc Carthy in 1996 for a) Teaching    b) Learning    c) Teaching-Learning    d) Mentoring
Q7.	Disability Leave may be granted, outside the leave account up to a maximum ----- days. a) 720 days    b) 120 days    c) 180 days    d) 365 days
Q8.	The most commonly used Coaching Model is: a) SWOT Model    b) GROW Model    c) SMARTER Model    d) 4MAT Model

Q9.	The Term ECM stands for: a) Every Child Movement c) Early Childhood Motivation	b) Early Childhood Management d) Every Child Matters
Q10	Professional development consists of reflective activity designed to improve an individual's..... a) Attributes      b) Knowledge      c) Understanding and skills      d) All above mentioned	
Q11.	----- provides the means to develop school capacity and reduce the workload of head teacher freeing him/her to do those key things that only heads can do. a) Democratic Leadership   b) Transformational Leadership   c) Distributed leadership   d) Team work	
Q12.	Where did the idea of Coaching come from.....? a) Sports Psychology      b) Learning Psychology      c) Health Psychology      d) All of them	

**Section B:**

***Encircle the right option for the given statements / questions.***

Q1.	Which acronym of SHEEP we are considering in subsequent statement “Children and young people live in decent homes and sustainable communities”... A) Be safe      b) Be healthy      C Achieve Economic well-being      d) Make a positive contribution		
Q2.	Which is not included in three “Big Basic Skills of Coaching” ..... a) Listening      b ) Leading      c) Reviewing d) Questioning		
Q3.	Those learners who learn by observing, analysing, classifying and theorising are called..... a) “WHY” learners      b) “WHAT” learners      c) “HOW” learners      d) “What If” learners		

## Appendix E: Experimental Design details

Treatment	
Control	716
Peer Recognition	649
Career-based Recognition	687
Peer Plus	635
Career Plus	707
Total	3394

E. 1: Summary of Randomization



# عظیم اساتذہ عظیم کہانیاں

## مبشر خلیق

بہاولپور سے مبشر خلیق کہتے ہیں کہ  
”پچھلے برس سے لے کر اب تک  
مدرسہ ہجرت اور نظر انداز



کہے ہوئے بچے (neglected students)  
جیسے پٹانہ پڑا کا مقابلہ کرنا میری عظیم کامیابیوں میں سے  
ایک ہے۔ سکول میں کام کے بوجھ کی ذمہ داری کے باوجود  
مبشر نے لنگھ میں ہجرتی کے سبب سے طریقے دریافت  
کرنے کی جدوجہد جاری رکھی۔ ان کا کہنا ہے کہ کئی میڈیا  
کا استعمال پینٹر اساتذہ کے لیے ایک ذرا دلچسپ اور اپنے خواب  
سے کم نہیں۔ تاہم ان کے مٹی میڈیا طریقے اور اپنے سکول  
میں ایک کیمپ کیسٹ کا نام کی۔ ان کے اپنے ساتھی اساتذہ  
کی حوصلہ افزائی کی کہ وہ بھی (مٹی میڈیا کا استعمال)  
تیکھیں تاکہ وہ اپنے پڑھانے کے انداز میں ہجرتی اور  
تیکھیں۔ مبشر کا یقین ہے کہ پبلک ایجوکیشن سسٹم  
میں تبدیلی صرف کمنٹ (committed) اساتذہ  
کے ذریعے ہی لائی جاسکتی ہے جو اس طرح کی چھوٹی  
چھوٹی کامیابیوں سے سائنسی سٹیج ایجوکیشن میں اہم دست  
لاکھیں گے۔ وہ کہتے ہیں کہ ”پچھلے ایک ایسا عمل ہے جو  
مسئلہ حوصلہ افزائی کا باعث بن جاتا ہے۔ جب میں  
طالب علم تھا تو مجھے اساتذہ میرے آئیڈیل بن گئے، اور  
اب میں ان کا ہڈ پانے کے لئے کوشش کرتا ہوں۔“

## شہشاد اور پانہ



گورنمنٹ گرلز ہائی سکول کی شہشاد  
رہا نہ کہتی ہیں کہ ”اگر ایک پچھرا چھرا  
پڑھا رہی ہو تو اس بات سے کوئی فرق  
نہیں پڑتا کہ کلاس کمرے میں ہے یا کھلے میدان میں۔“  
جب سے ان کے سکول کی ادارت کو غیر ملکی ادارہ قرار دیا گیا  
ہے، شہشاد نے نشان لیا کہ وہ اسے بچوں کی تعلیم پر اثر  
انداز نہیں ہونے دین گی۔ آج 4 سال بعد وہ اپنے سکول  
کو 600 بچوں کے ساتھ نہایت کامیابی کے ساتھ کھلے  
میدان میں چلا رہی ہیں۔ حال ہی میں انہیں ان کے  
ایک طالب علم نے فون کر کے بتایا کہ اس نے بورڈ میں  
دوسری پوزیشن حاصل کی ہے۔ اسی طرح ایک دوسرے  
طالب علم نے انہیں بتایا کہ اس نے ایم ایس سی (MSc)  
کی ڈگری مکمل کر لی ہے۔ شہشاد کہتی ہیں کہ ”مدرسہ ہجرت  
کے ساتھ سکول چلانا بہت مشکل ہے مگر میں نے اپنے  
طلباء اور اساتذہ کی حوصلہ افزائی کے لیے سخت جدوجہد  
کی۔“ سب اساتذہ میں کچھ نہ کچھ مٹائی کر دکھانے کی  
استعداد اور نظر موجود ہوتا ہے۔

## نانہ بشارت



نانہ بشارت کہتے ہیں کہ جب ایک مرحہ  
وہ اپنے سکول میں لنگھ میں  
ہجرتی لانے کا مزم کر لیں، تو کوئی  
چیز ان کے راستے کی رکاوٹ نہیں بن سکتی تھی کہ یہ  
والدین کی (اپنے بچوں کی تعلیم میں) عدم دلچسپی ہی کیوں  
نہ ہو۔ جب نانہ نے کئی مرحہ پڑھانا شروع کیا تھا تو  
اس وقت ان کے سکول میں صرف 20 بچے تھے، سکول  
میں سہولیات کی کمی تھی اور بچوں کی تعلیمی کارکردگی انتہائی  
بری تھی۔ نانہ نے طلباء اور ان کے والدین کو بہتر طریقے  
سے گلے کیلئے سخت محنت کی، جی کہ کسی مرحہ وہ اتوار کے  
روز بھی سکول آتی رہیں۔ پانہ آفران کو سکول میں بچوں اور  
ان کے والدین کے ساتھ ایک مشہور تعلق قائم کرنے کی  
صورت میں واضح فرق نظر آئے شروع ہو گیا۔ آج ان کے  
پاس 200 بچے پڑھ رہے ہیں اور ان کی تعلیمی کارکردگی  
کے بارے میں کئی بات شائد ہوتی ہیں۔

E. 2: Self-Efficacy Frame Hand-out

Appendix F: Snapshots of Training





# 3

## Carrots or Sticks: The Impact of Incentives and Monitoring on the Performance of Public Extension Staff

# Carrots or Sticks: The Impact of Incentives and Monitoring on the Performance of Public Extension Staff

Garance Genicot\*, Zahra Mansoor<sup>§</sup> and Ghazala Mansuri <sup>†</sup>.

October, 2020

## ABSTRACT

Like many other frontline public service departments, agriculture extension service delivery is also subject to the classic agency problem where staff is hard to monitor, and due to the nature of farmer extension tasks where some dimensions are easily measurable and some are hard to measure, designing staff incentives is not straightforward. Partnering with the Agriculture Extension Department in Punjab Pakistan, we evaluate the impact of three alternative pay-for-performance schemes on extension outreach and quality. The performance schemes leverage the departmental digital monitoring system called AgriSmart and link incentives to objective metrics available on AgriSmart (Objective arm), supervisors' own subjective evaluation of staff (Subjective arm) to capture additional harder to measure dimensions, or supervisors' own subjective evaluation with an additional element of top-down monitoring to align supervisors' incentives with the objectives of the principal (Subjective Plus arm). Early results show that while all treatments improve performance on AgriSmart metrics, Subjective Plus also shows positive and significant effects on farmer experience of extension outreach and quality and farmer agricultural practices. In addition, we also find positive and significant effects on supervisor (AD) performance in the Subjective Plus arm, particularly on tasks that form an important input into how frontline staff conduct extension outreach. Additional data collection will explore whether the results in Subjective Plus are driven by the effective use of discretion by supervisors in their evaluations, an improvement in supervisor's own performance and practices, or both, before we can arrive at the policy and cost implications of the programme.

**Acknowledgements:** Ayesha Shahid provided invaluable research and implementation support to the project. We are also thankful to Saad Imtiaz, Usman Zahid, and Maria Qazi for research assistance. This research was made possible through ongoing support and insights from various counterparts at the Extension Department, the Agriculture Delivery Unit, and the Agriculture Ministry during the last four years. We gratefully acknowledge funding from the World Bank i2i, World Bank Strategic Research Programme, and the International Growth Center.

\* Georgetown University; § Dphil Candidate, Blavatnik School of Government; † The World Bank.

# 1 Introduction

Agriculture productivity is recognized as critical for growth and development, but it remains a challenge for many developing countries. Amongst many constraints to improving productivity, farmer access to timely and actionable information is clearly an important aspect where the information barriers are often the largest for smaller, poorer, and remote farmers (Conley and Udry, 2010; Duflo et al., 2011; Suri, 2011). This makes public extension critical given it remains the primary source of productivity information for farmers who need it the most.

However public extension, like many critical public services which rely on last mile service delivery, often tends to be plagued by weak incentive and monitoring systems. While monitoring public sector employees is a challenge everywhere, the problem is exacerbated in extension services due to the spatial spread of farmers. In addition, extension tasks have different dimensions where some are easily measurable (such as the number of farmer visits conducted or farmer trainings provided) and some are harder to measure (such as the quality of advice provided). This makes performance of public extension a classic agency problem where staff is hard to monitor and, due to the nature of extension tasks, the known advantages and disadvantages of objective versus subjective incentive systems are unclear. Incentives based on objective metrics, though effective, can often be distortionary by incentivizing more on dimensions that are easier to measure (Holmstrom and Milgrom, 1991; Gine et al., 2017). On the other hand, subjective measures allow supervisors to capture harder to measure dimensions of performance (Parent and Rebitzer, 1999), but can often result in inaccurate rankings due to collusion or various types of supervisor biases (Gibbs et al., 2004; Prendergast, 2007; Neal, 2011). At the same time, subjective evaluations can also result in the effective use of local information for evaluations instead of patronage if the supervisor incentives are aligned with the objectives of the principal (Prendergast and Topel, 1996). Against these considerations, there is limited evidence in extension services on the effectiveness of objective versus subjective incentive systems, and more specifically, on how subjective evaluations could possibly be strengthened by aligning supervisor incentives with the

objectives of the principal.<sup>1</sup>

Partnering with the Agriculture Extension Department in Punjab Pakistan, this study evaluates the impact of a province-wide pay-for-performance programme in improving extension outreach and quality, and ultimately farmer agricultural practices. In particular, the study aims to answer whether performance incentives can improve extension outreach and quality, and if so, whether the impact of incentives that rely solely on objective metrics is larger or smaller than incentives that allow for the use of supervisor discretion to assign incentives. In addition, it also aims to explore whether the use of discretion is more effective if supervisor incentives are aligned with the objectives of the principal.

The questions of the study are directly policy relevant. Public extension is the primary source of information for smallholder farmers in Punjab which comprise more than 60% of the total farm households in the province.<sup>2</sup> Agricultural growth in Punjab has been on a downward trend. The observed gaps in yield between progressive and average farmers highlight the significant potential gains in overall productivity by improving public extension services for smaller and more remote farmers (Punjab Agriculture Policy, 2018). However, when it comes to improving extension services, the Extension Department has historically lacked access to any objective data on staff outreach and service quality, which limits its ability effectively monitor staff through the center. At the same time, direct line managers of extension agents - called Assistant Directors (ADs) - have private information on the quality of staff performance through their day-to-day engagement with their staff, spot checks, and farmer feedback calls. In this context, the questions around the merits and demerits of objective versus subjective incentive systems are directly relevant. Prior to the roll-out of the incentive programme, the Extension Department developed a comprehensive digital performance management system called AgriSmart anchored in smart-phone/tablet-based

---

<sup>1</sup>Within agriculture extension, limited rigorous evidence on the impact of publicly provided extension services on farm productivity or on adoption rates does not help. Most of the existing studies have problems of causal attribution and lack the data to cleanly measure outcomes of interest. Anderson and Feder (2007) provide an excellent review of this early literature and argue for the need for well-designed evaluations. Recent papers have begun to fill this gap and are producing more encouraging evidence of the potential for agricultural extension (see Fernando, 2016; Beaman et al., 2018).

<sup>2</sup>[www.agripunjab.gov.pk](http://www.agripunjab.gov.pk).

daily activity reporting by all field staff as a way to improve staff monitoring.

Leveraging the AgriSmart system, the department designed an incentive programme and randomized three different pay for performance schemes (and a control group) across 126 tehsils in Punjab. Tehsils are an administrative unit below districts in Punjab and are managed by Assistant Directors (ADs) in the Extension Department. In the first set of tehsils (Objective Arm), all field staff are assigned a bonus using only the metrics available from the AgriSmart portal. In the second set of tehsils (Subjective Arm), the supervisor (AD) can assign a bonus to all staff under his/her tehsil and use the data from the portal as per his/her discretion. The third arm (Subjective plus) mirrors the second arm but adds an element of top-down monitoring to align the AD incentives with the objectives of the principal. In order to do this, a report on the performance of the AD's tehsil is shared with the top leadership of the department (Secretary, Director General, and Divisional Directors) every month.

Early results from the evaluation of the incentive programme yields three main lines of results. First, we observe positive and significant treatment effects on AgriSmart outreach measures (such as days present, hours worked, villages visited, and farmers visited) across all arms. An overall performance index normalizes each AgriSmart indicator by the mean and standard deviation of the control group and then computes an average of all normalized measures. We find a positive and significant treatment effect of  $0.42\sigma$  on the AgriSmart overall performance index for the Objective arm,  $0.19\sigma$  for the Subjective arm, and  $0.29\sigma$  for the Subjective Plus arm. The Objective arm performs significantly better than both the Subjective and the Subjective Plus arm on the overall AgriSmart performance index. This is intuitive given the Objective arm incentive is explicitly linked to these indicators and also consistent with existing evidence which shows that incentives work when they are linked to outcome measures that can be measured objectively and clearly (Dufflo et al., 2012; Khan et al., 2016).

Second, we find positive and significant effects on measures of extension outreach and quality (as captured through farmer surveys that are not measurable through AgriSmart) in the Subjective Plus arm. A farmer outreach index normalizes each outreach indicator (such as whether the farmer knows the frontline agent, whether the farmer attends extension activity, or whether

public extension is the main source of information for the farmer) by the mean and standard deviation of the control group and then computes an average of all normalized measures. We find positive and significant treatment effects of  $0.10\sigma$  on the overall farmer outreach index in the Subjective Plus arm. We also find that out of those who attend extension meetings, farmers in the Subjective Plus arm are 14% points more likely to report the meetings to be useful as compared to the control group. However, we do not find such effects in the Objective or the Subjective arm. We observe the same trend of treatment effects on our secondary outcomes such as farmer practices and yield. In the Subjective Plus arm, we find positive and significant effects on farmers changing their farming practices (such as using different inputs and plantation techniques) and some indicative effects of farmers reporting higher yield as compared to the control group. Similar to the farmer reported indicators of extension outreach and quality, we do not find such effects in the Objective or the Subjective arm.

Third, given Assistant Directors (ADs) who are the direct line supervisors of frontline extension agents conduct several managerial tasks that form an important input into how staff carry out their extension activities, we observe AD performance on these metrics. In particular, we observe AD performance on scheduling and assigning Farmer Training Programmes (FTPs), which is the central extension outreach task conducted by frontline agents. Our results show that ADs in the Subjective Plus arm schedule and assign more FTPs to frontline agents and these effects are significantly different from the Objective and the Subjective arms.

Supplementary evidence explains why we observe treatment effects on measures of extension outreach and quality in the Subjective Plus arm across different measures (i.e. AgriSmart measures of outreach and farmer reported measures of outreach and quality); and why the treatment effects on AgriSmart outreach measures in the Objective and Subjective arms do not translate to farmer reported measures of extension outreach and quality. Given the only difference in the Subjective Plus arm is the AD-layer of monitoring, we hypothesize two potential channels of AD behaviour change that could have resulted in improved extension outreach and quality in the Subjective Plus arm. First, AD's own performance and practices could have changed which may have effected extension agents' performance. Second, AD's bonus assignment behaviour

may have become more efficient which may have effected agents' performance. We demonstrate some evidence for the first channel through our treatment effects on FTP scheduling. Where AD bonus assignment behaviour is concerned, we find significantly different behaviours in bonus assignments between the Subjective and Subjective Plus arm. In particular, we find that ADs in the Subjective arm are more likely to make repeat assignments of bonuses to the same individuals overtime as compared to the Subjective Plus arm. While we are unable to confirm whether this indicates bias or repeat assignments to high ability staff, it indicates differences in bonus assignment behaviour that requires further inquiry.

We also explore mechanisms underlying how the Objective and Subjective incentive schemes may have worked to understand our results. Preliminary evidence indicates that the Subjective Plus arm shows a higher coefficient on indicators of outreach effort on the extensive margin (i.e. expanding outreach to more farmers), although the coefficients are not significantly different from other treatments. On the other hand, the Objective arm shows evidence of higher outreach effort on the intensive margin (i.e. repeating the same set of farmers more times), with the effects being significantly higher compared to the Subjective and Subjective Plus arm. This indicates that the designs of Subjective Plus and Objective arms may have encouraged different types of outreach effort. Second, we find that staff is significantly less satisfied in the Objective arm as compared to the other arms, which could in part explain why we do not observe treatment effects on farmer reported measures of quality in the Objective arm.

Overall, our results show that while all treatments improve performance on AgriSmart outreach measures, Subjective Plus also improves performance on farmer reported experience of extension outreach and quality, and farmer-level outcomes. In our context, given ADs are the direct supervisors of extension staff and provide important input to frontline staff on how service is to be delivered, an AD level monitoring layer in Subjective Plus appears to work by aligning AD incentives with the objectives of the principal. The evidence, however, is not entirely conclusive as to whether the positive treatment effects observed in Subjective Plus are driven by checking bias of supervisors in their bonus assignments or via improving supervisors' own performance and management practices. For example, if the impact observed in Subjective Plus is primarily

driven through the AD performance and practices channel instead of incentives for frontline workers, this could have significant cost implications given the Subjective Plus (and Subjective) arm cost the government twice as much as the Objective arm.

Further data collection will explore these and additional mechanisms - such as AD preferences for assigning bonuses, additional measures of AD performance, and whether the schemes induced different types of trade-offs (such as effort on the extensive versus intensive margins or few longer versus many shorter extension trainings) and their related impacts on farmer practices and outcomes. The choice of the incentive scheme will ultimately depend on the government's objectives for extension service delivery and the costs and benefits associated with such trade-offs.

Two recent studies have similarities to our study and context. Callen et al. (2016) analyse the impact of a randomly assigned smartphone monitoring system on public health worker absenteeism in Punjab. They find positive treatment effects of the technology, but only where political influence and bureaucratic interference is limited. Khan et al. (2016) use an experimental approach to evaluate the impact of incentives in the property tax department in Punjab and show that rewarding tax collectors on revenue increased total revenue, but also resulted in more collusion via increasing tax collectors' bargaining power relative to taxpayers. Our study complements these studies but also adds a unique angle by evaluating the impact of monitoring and performance-based pay schemes in another public service where the agent task has both observable and non-observable components, performance is hard to measure, and supervisors hold important private information about agent performance.

Our research contributes to several themes of literature. It contributes to the overall literature on performance pay in the public sector, where tasks have observable and non-observable dimensions, are contextual, and hard to measure (see Burgess and Ratto, 2003; Manning, Hasnain, and Pierskalla, 2012 for surveys). Our results also contribute to the literature on rules-based bureaucracy versus discretion. While the challenges of discretion in subjective evaluation systems are well documented (Baker et al., 1994; Prendergast, 1999; Gibbs et al., 2004), we show circumstances under which they may or may not work well. Finally, our results also contribute to the literature on how practices and performance of middle-management may relate to frontline

service delivery (Rasul and Rogger, 2018).

This paper is structured as follows. Section 2 describes the relevant literature while Section 3 presents the theoretical framework backing our design. Section 4 outlines the contextual setting, Section 5 presents the experimental design, randomization, and implementation, and Section 6 presents an outline of the main data sources and outcome measures. Section 7 presents early results and Section 8 concludes.

## 2 Literature

In this section, we outline the key pieces of literature that support our conceptual framework presented in Section 3.

Incentives that tie performance to objective and measurable metrics of performance can align agent effort with the objectives of the principal on those dimensions and improve overall performance (see Bandiera et al., 2007; Khan et al., 2016). However, settings where agents carry out tasks that have both easy to measure and hard to measure dimensions, objective measures can also distort effort with agents exerting more effort on the dimension of task that is easily measurable (Holmstrom and Milgrom, 1991; Andreoni et al., 2016; Gine et al., 2017). In other cases, challenges of measurement can dampen the effect of incentives leading to gaming, fraud, and collusion (Banerjee et al., 2008). In addition, objective performance measures can be rigid and impose undue risk on the employee by not taking external shocks into account. Under these conditions, an evaluation based on an objective metric alone could be noisy, and a risk-averse agent might not respond to incentive pay (Rajan and Reichelstein, 2009).

On the other hand, subjective evaluations are able to capture more dimensions of performance but are subject to supervisor biases. Gibbs et al. (2004) argue that one of the advantages of subjectivity is that it allows evaluators to exploit additional information about conditions that arise after the formal reward plan is set. However, they find that subjective evaluations are

subject to the evaluators' personal biases, and the feedback from them is often vague and uninformative. They identify two common biases in subjective evaluations: centrality and leniency. The 'centrality bias' refers to the practice of offering all employees ratings that differ little from the norm and the 'leniency bias' refers to offering employees ratings that are skewed towards the top of the scale (Gibbs et al., 2004).

Where supervisor discretion is concerned, a key question is whether the incentives of the supervisor are aligned with the objectives of the organization (Holmstrom, 1980). Supervisors can have their own incentives to evaluate workers in ways that may not be desired by the organization (Prendergast, 1999). If supervisors are not explicitly rewarded for accurate performance ratings, they are unlikely to be motivated to invest time in gathering information since they receive little benefit from conducting more accurate evaluations (Baker et al., 1994). Despite these challenges, Prendergast and Topel (1996) argue that while discretion leaves room for inaccuracy and bias, it can also result in the use of local information rather than patronage if the supervisor incentives are aligned with the objectives of the principal.

### **3 Conceptual Framework**

In this section, we present a theoretical framework to conceptualize our experimental design. Key hypotheses that we aim to test follow directly from the framework.

Agriculture service delivery involves tasks which have some easily measurable and some hard to measure dimensions. Two key tasks include individual farmer visits for one on one advice and group-level farmer training programmes (FTPs) that are catered to different crops and seasons in the agricultural cycle. These outreach tasks have some dimensions that are easily measurable and observable through AgriSmart (such as number of farmers contacted for individual advice, or number of FTPs conducted). However, some dimensions of effort are harder to measure because they have a quality dimension (for instance repeating visits when needed while also reaching out to new farmers, providing quality advice to farmers, and paying attention to their

concerns). Supervisors of frontline agents often have more information regarding their staff performance along these harder to measure dimensions, through their monitoring of staff, contact with community members, or direct feedback provided by staff.

For simplicity, assume that there are two dimensions of effort: one dimension of outreach effort that is observable and easy to measure that we label  $e$ , and another dimension that is harder to observe and measure that we call *quality* and denote by  $q$ .

A level of outreach  $e$  and quality  $q$  by a field worker generates a social benefit of  $f(e, q)$  (i.e. farmers' welfare or productivity) that is not directly observable. The marginal product of each type of effort is strictly positive and decreasing. An incentive scheme determines the probability  $p_i(e, q)$  that a worker  $i$  with effort levels  $(e, q)$  gets a bonus  $B$ .

Assume that a worker puts a weight  $\alpha_i$  over his financial gain (his *extrinsic motivation*), and a weight  $1 - \alpha_i$  over the social benefits generated by his activities (his *intrinsic motivation*). In addition, there is a cost to effort given by  $c(e, q)$ , which is strictly increasing in effort.

The principal cares about maximizing the social benefit  $f(e, q)$  taking into account the effort costs  $c(e, q)$ . To do so, he/she chooses a bonus scheme  $\{p(e, q)\}_{\forall e_i, q_i}$  (at an overall cost of  $C(\sum_{i'} p_{i'}(e_{i'}, q_{i'}) B_{i'})$ ) and to a feasibility constraint (he/she can condition the scheme only on what is observable).

After the principal has set the bonus scheme, the worker chooses his/her effort levels  $(e, q)$  to maximize utility:

$$\alpha p_i(e, q) B + (1 - \alpha) f(e, q) - c(e, q). \quad (1)$$

Under the regular assumptions (differentiability and concavity), the first order conditions would characterize the worker's choice:

$$\begin{cases} \alpha \frac{\partial p_i(e, q)}{\partial e} B + (1 - \alpha) \frac{\partial f(e, q)}{\partial e} - \frac{\partial c(e, q)}{\partial e} = 0 \\ \alpha \frac{\partial p_i(e, q)}{\partial q} B + (1 - \alpha) \frac{\partial f(e, q)}{\partial q} - \frac{\partial c(e, q)}{\partial q} = 0. \end{cases} \quad (2)$$

Clearly if  $\alpha = 0$ , the worker would choose the optimal effort and a zero bonus would be optimal.

Under a purely objective scheme, bonuses depend solely on outreach effort  $e$ . It is easy to see that if  $\alpha > 0$ , an increase in  $B$  increases  $i$ 's outreach effort  $e$ . Whether it encourages or discourages quality depends on complementarity vs substitutability between outreach  $e$  and quality  $q$  in the production and cost functions. For instance, if the social benefits are independent but  $c_{12} > 0$  then an increase in  $B$  decreases  $q$ .

Under a subjective scheme, supervisors choose the allocation of bonuses to the field workers. Supervisors have the same information on  $e$  as supervisors in an objective scheme. Assume for now that supervisors observe  $q$  (alternatively we can assume that they have only partial information about  $q$  and potentially at some cost). Subject to a feasibility constraint (i.e. conditioning on what is observable), the supervisor selects an allocation of the bonus  $B$  for each  $i$ . The resulting mapping from effort to bonus implied by the supervisor bonus allocation behaviour determines  $p_i(e, q)$  for individual  $i$ . A given supervisor bonus allocation results in workers effort choices  $(\hat{e}, \hat{q})$  that maximizes their utility.

Supervisor preferences are defined over a private component and over the net social benefit generated by the workers. In particular, let  $\beta_j \in [0, 1]$  be the weight that supervisor  $j$  puts on his/her private preferences over the allocation of bonuses and  $1 - \beta_j$  be the weight that he/she puts on the net social benefit generated. The private component of  $j$ 's preference is captured by  $V_j(\mathbf{p}) = \sum_i \omega_{ji} u(p_i(\hat{e}_i, \hat{q}_i)B)$  where  $\omega_{ji}$  is the weight that supervisor  $j$  puts on worker  $i$  and where  $u(0) = 0$ ,  $u' > 0$  and  $u'' < 0$ . This formulation allows us to capture a wide range of supervisors' biases such as favoritism ( $\omega_{ji} > 0$  for only some select individuals), or centrality bias (where  $\omega_{ji} = \omega_j > 0$  for all  $i$ ) (as in Gibbs et al., 2004).

All together, these preferences mean that supervisor  $j$  chooses an allocation of bonuses that aims at maximizing:

$$\beta_j V_j(\mathbf{p}) + (1 - \beta_j) \left( \sum_i (f(e_i, q_i) - c(e_i, q_i)) - C \left( \sum_{i'} p_{i'}(e_{i'}, q_{i'}) B_i \right) \right) \quad (3)$$

If  $\beta_j = 0$ , supervisor  $j$ 's objective is to maximize the social surplus. In contrast,  $\beta_j = 1$  represents a supervisor whose preference over the allocation of bonuses is unrelated to the social benefit potentially generated by the workers. In addition since this supervisor does not internalize the cost of the bonuses, he/she has an incentive to be overly generous in the distribution of the bonuses (the leniency bias). Note that if the social benefits of  $q$  and  $e$  are independent but  $c_{12} > 0$ , then supervisors would select to incentivize both dimensions of effort.

Note that when subjective incentive schemes align the incentives of the supervisors with the objective of the principal (extension head quarters), it decreases the weights of private benefit from bonus allocation  $\beta$ .

The framework presented above outlines the key differences between incentive schemes that are purely objective, subjective, or subjective where supervisor incentives are aligned with the objectives of the principal. This framework can be extended to include the effort of supervisors themselves in two possible ways. First, supervisors provide inputs into extension service delivery. For example, supervisors (ADs) schedule and assign farmer training programmes (FTPs) to extension workers, thereby increasing their outreach. Second, supervisors may engage in (costly) monitoring of their workers in order to observe  $q$ .

Following our framework, we generate the following hypotheses:

Hypothesis I: All incentive schemes (objective or subjective) should increase easy to measure outreach indicators ( $e$ ).

Hypothesis II: Easy to measure outreach indicators ( $e$ ) will improve more in objective incentive schemes that can be directly linked to such measures versus no incentive schemes or subjective schemes.

Hypothesis III: Harder to measure effort ( $q$ ) will increase in subjective schemes that allow for more dimensions of quality to be captured relative to no incentive schemes or purely objective schemes.

Hypothesis IV: Both dimensions of effort ( $e$  and  $q$ ) will increase more in subjective schemes that can align supervisor incentives with the objectives of the principal versus purely subjective schemes.

Hypothesis V: Subjective schemes will distribute more bonuses compared to objective schemes.

In addition to these hypotheses, we will also try to answer a range of additional questions. First, the extent to which objective schemes will improve extension quality will depend on the strategic complementarity versus substitutability of  $e$  and  $q$ . By observing impacts on both outreach and quality, we will be able to assess the extent to which the two are complements or substitutes. Second, incentive schemes where supervisors have access to both objective and subjective metrics, we will explore whether the two sources of information operate as complements or substitutes (Gibbs et al., 2004). Third, we will explore whether different objective and subjective incentive schemes vary in their outreach effort on the intensive (i.e. making more visits to the same set of farmers) and extensive (i.e. expanding outreach to a greater set of farmers) margins.

## 4 Context

Out of the total cultivated land in Pakistan, 57% lies in Punjab. Hence improving agricultural productivity is at the forefront of any policy reform in the province.<sup>3</sup> Agricultural productivity in Punjab has slumped over the last couple of years, with little or no growth in yield (Punjab Agriculture Department, 2018; Pakistan Economic Survey, 2019). This has primarily been attributed to poor farming practices, low technology adoption, and low level of crop diversification (Punjab Agriculture Department, 2018). Smallholder farmers, that constitute more than 60% of total farm households, particularly register low yields (Punjab Agriculture Department, 2018). Access to relevant and timely information for these farmers could potentially have huge gains for agricultural productivity and reduction in poverty. But information barriers for these smaller and more remote farmers are large, and public extension remains their primary source

---

<sup>3</sup><http://www.cabi.org/projects/project/10880>.

of information.

## **4.1 Agriculture Extension Department in Punjab**

In Punjab, the Provincial Agriculture Extension Directorate is the main entity responsible for getting timely and productive information to farmers. The department has a frontline workforce of up to 2500 workers that have a presence down to the village level - starting from field assistants (FAs) at the union council level who cover 7-8 villages on average, agricultural officers (AOs) at the markaz level who cover 7-8 union councils, assistant directors at the tehsil level (ADs), deputy directors (DDs) at the district level, and directors at the divisional level (See Figure 1).

The AOs and FAs are the key frontline staff. Their daily tasks include a variety of activities - the Farmer Training Program (FTP), a monthly tour plan according to which AOs/FAs are required to conduct group training programs in villages under their jurisdiction, is one of the primary extension tasks. In addition, they also conduct individual farmer visits for advice on use of agricultural inputs, setting up demonstration plots, or departmental activities such as reporting and office desk work. At times of crises, they are often assigned non-departmental tasks such as school or dengue monitoring. While both AOs and FAs work separately to provide advice to farmers at the markaz and/or village level, they work together to provide FTPs (with the AO being the technical lead and the FA playing a support role). As Figure 1 shows, ADs are the direct supervisors of the frontline workers in each tehsil. ADs carry a range of managerial tasks that form an important input into how extension is carried out. The most important of these tasks is the scheduling and assigning of FTPs to each staff member based on which these activities are conducted. In addition, the ADs are also mandated to conduct surprise visits and staff meetings at the tehsil head quarters to carry out procedural quality assurance.

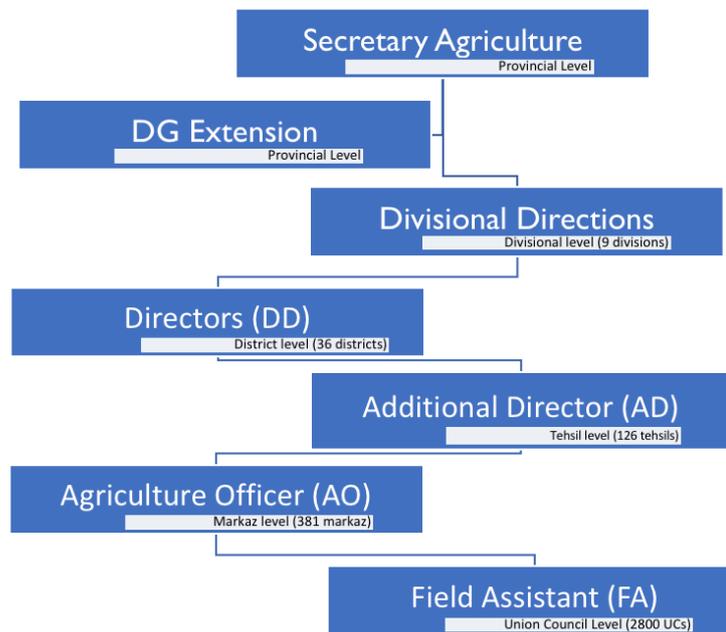


Figure 1: HIERARCHICAL STRUCTURE

Though the department has extensive institutional structures in place, historically it has lacked management tools that can provide objective and verifiable data on staff outreach and quality of service. The only tool for monitoring is surprise field visits by supervisors but these have also been ad-hoc in nature. Due to the lack of objective and verifiable data on staff time-use and service quality, it has been hard to design any performance-linked incentives, and extension workers have few, if any, incentives to reach farmers off the beaten track.

In addition, there is a perception that public extension services have little to offer and have been captured for the most part by larger landowners under the guise of support to “lead farmers” and that on balance this suits field staff who have little incentive, in any case, to reach out to small farmers. At the same time, extension is a large and costly department for the government and its effectiveness has become critical both for efficient utilization of government budgets and more importantly for improving agricultural productivity.

## 4.2 Travelling and Daily Allowance

One type of incentive that does exist in the system is that frontline staff are eligible to claim travelling and daily allowance (TADA) through the non-salary TADA budget head that each tehsil receives with every budget release. The *de jure* requires that field staff only claim TADA for distances that are greater than a 16 km radius from any staff's work jurisdiction. However *de facto*, supervisors approve TADA claims for regular work activities within the 16 km radius with the justification that workers need some TADA to do their jobs.

Our analysis of the historical TADA data shows that this budget has been disbursed unequally both across and within districts. Figure 2 depicts average TADA claims and payments by districts for FY 15-16. This shows significant variation across districts in the average per staff TADA payments. The magnitude of this variation can be seen in the following comparison – the average payment for Lodhran (PKR 61,904) is almost ten times as much as that in Narowal (PKR 6,271). This is despite the former having a slightly smaller number of frontline staff as compared to the latter. There are also significant differences in the percentage of unmet claims across designations, with the lowest tiers in the departmental hierarchy (i.e. frontline field force) having the highest proportion of unmet claims.<sup>4</sup>

This ad hoc assignment of TADA disbursement acts as a deterrent to the effectiveness of field staff, particularly when expenses incurred as part of outreach effort are not honored. The department is well aware that the current system has created inequities, across districts, across staff within districts, and across staff at different levels in the hierarchy, and hence recognizes the need to make this budget more transparent, accountable, and incentive responsive.

---

<sup>4</sup>Based on authors' calculation of the historical TADA data.

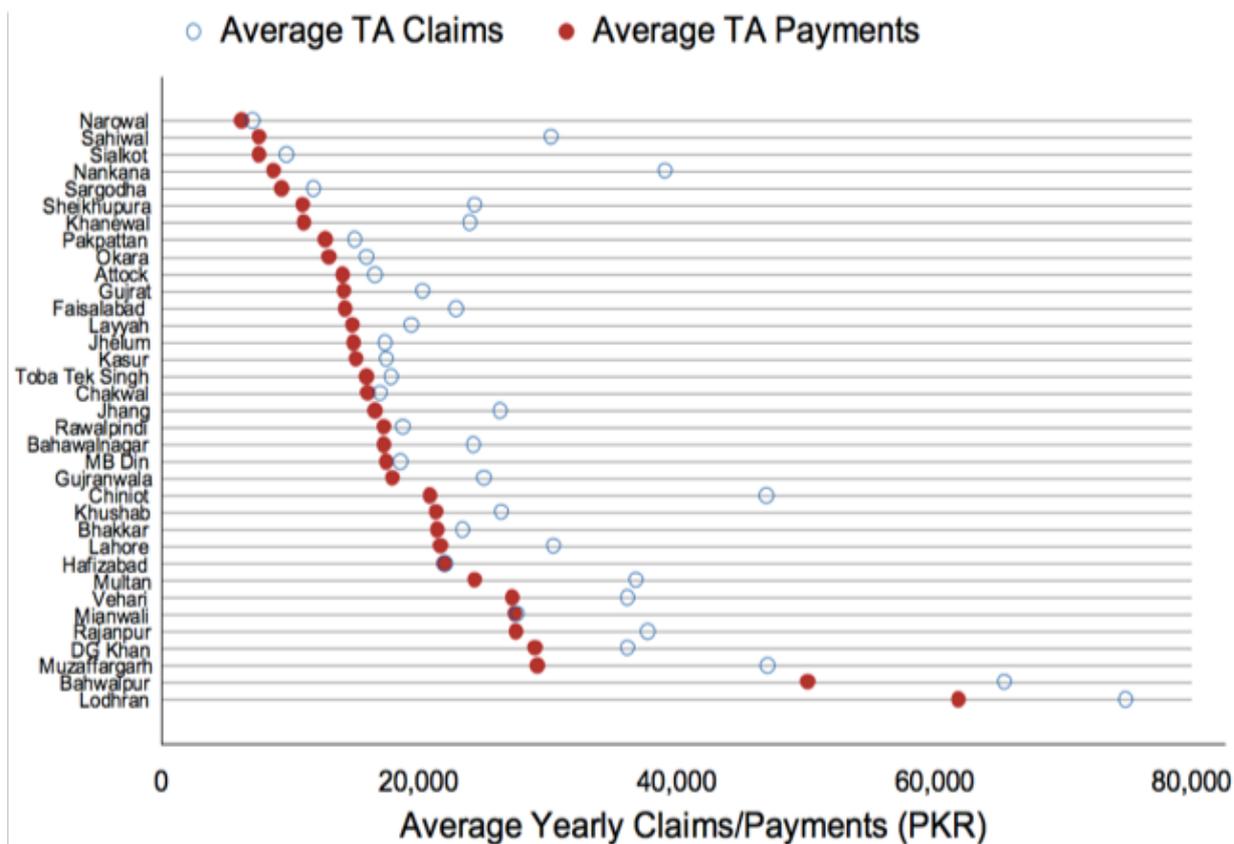


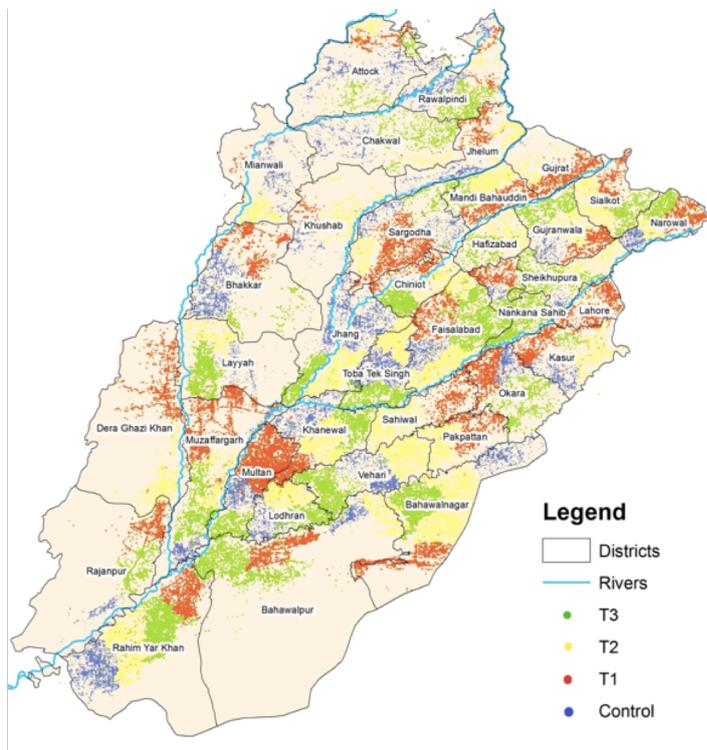
Figure 2: DISTRIBUTION OF TA DA PAYMENTS AND CLAIMS IN PUNJAB 2015-2016

### 4.3 The AgriSmart Programme

In 2014, the Ministry of Agriculture began a programme to improve public extension service delivery through ICT based monitoring of agricultural extension workers under a World Bank Programme called the Punjab Public Sector Management Reform Program (PPMRP). The programme was implemented by the Punjab Institute of Technology Board (PITB), which was the technology counterpart for the PPMRP. The AgriSmart Programme which was developed under the PPMRP is a comprehensive performance e-management system which entails a smartphone-based application where field staff are required to log their daily time-use and a management dashboard for supervisors that caters to various managerial tasks such as assigning farmer train-

ing programmes (FTPs), approving leaves, and making HR updates.

The AgriSmart programme (including the AgriSmart portal and the accompanying smartphone application) is now active all over Punjab. All field staff send in daily activities that are geo-tagged, photo and time-stamped, which allows for automated calculation of time-spent on activities and daily attendance. Although farmer visits and village visits are self-reported, geo-tagging of activities and audits of reported data through farmer call backs discourage misreporting. Figure 3a shows a visualization of extension activity data points for a 3-month period in 2019. The accompanying AgriSmart dashboard enables two-way digital communication between ADs and field staff in the form of online leave requests (both submission and approval) and online scheduling and assignment of farmer training programmes. Figure 3b below shows an image of the homepage of the monitoring application which the field staff uses to enter activity data or file leave requests.



(a) AGRISMART DATA POINTS



(b) HOMEPAGE

Figure 3: AGRISMART SYSTEM

## 5 Experimental Design and Randomization

The Extension Department rolled out three different incentive schemes. All incentive schemes include a performance-based bonus. The bonus includes four different categories: base payment, low bonus, high bonus, and no bonus. The bonus amounts under each of the categories were pre-determined by the department. The base payment was set at roughly 7% of the monthly salary and was meant to be paid to field staff for basic attendance to cover their basic traveling expenditure. This amount was bench-marked against historical average TA claims made by field

staff.<sup>5</sup> The low bonus and high bonus amounts were set at an additional 7% and 15% of salary respectively and were to be paid out for good performance (see Table C.1 in Appendix C for incentive amounts for each bonus category).

A unique feature of the programme is that it is designed to be budget neutral. All bonus payments are made through the pre-existing TADA budget in an attempt to make the existing budget more accountable and incentive responsive.

## 5.1 Treatment Arms

*Objective Arm:* In tehsils assigned to the Objective arm, extension staff bonuses are determined solely by measures of their performance as logged on AgriSmart. To qualify for base payment, the worker must meet his/her basic compliance thresholds for days and hours worked in a month. For low bonus or high bonus, the worker must meet additional thresholds in terms of villages and farmer visits, distance travelled, percentage of time spent on extension, and proportion of completed farmer training programmes. Thresholds for meeting base, low, and high bonus for AOs and FAs were determined by the departmental headquarter and are detailed out in Table C.2 and C.3 in Appendix C. The indicators of performance for both AOs and FAs are similar but differ quantitatively based on their geographical scope.

A system of audit with farmer call backs is operated by a third party to verify the quality of the AgriSmart data to discourage misreporting. ADs are able to view the performance metrics and bonus category of each worker in their tehsil but have no discretion over the assigned bonus for any worker. Ultimately, comparing the Objective arm with the Control arm helps us assess whether linking bonuses to objective measures of performance increases extension workers' outreach and quality of services.

*Subjective Arm:* In tehsils assigned to the Subjective arm, the AD has discretion over each worker's bonus category. To make this decision, the AD may use his/her subjective evaluation (as

---

<sup>5</sup>This was initially bench marked against FY15-16 but in FY18-19 the amounts were revised upwards.

gathered through surprise visits, farmer feedback, or interaction with staff), the staff performance summary on AgriSmart, or some combination of the two, but has complete discretion over how he/she assigns payments. The AD makes bonus assignments to staff on the dashboard on a specific *staff payment assignment page* that is available through his/her dashboard log in (see Figure C.4 in Appendix C). Against each staff member, the performance metrics of AgriSmart are available to the AD (without specific thresholds), but viewing them is optional. The AD is expected to assign bonuses to staff on a monthly basis.

The potential advantage of the Subjective arm comes from the AD's ability to monitor dimensions of performance that are not easily quantifiable by the AgriSmart system. On the other hand, the subjectivity could reduce transparency and leave room for bias. Comparing the Subjective arm with the Control arm helps us assess whether linking bonuses to subjective performance evaluations increases extension worker outreach and quality of services over the Control as well as in comparison to the Objective arm.

*Subjective Plus Arm:* In tehsils assigned to the Subjective Plus arm, the AD bonus assignment process mirrors the Subjective arm. The main difference with the Subjective arm is that a report on the tehsil's performance is provided to the ADs' supervisors including Director General Extension, Secretary Agriculture at the provincial headquarters, and Divisional Directors on a monthly basis (See Figure C.5 in Appendix C for a snapshot of the report). This report includes AgriSmart summary statistics of worker performance, AD compliance on a set of AgriSmart Standard Operating Procedures (SOPs) (such as whether he/she scheduled FTPs and approved staff leaves), and indicators of data quality from the AgriSmart audit.

The motivation behind this arm is to see whether the career concerns of supervisors can be useful for aligning their incentives with the objectives of the principal. Comparing the Subjective Plus arm to the Subjective Arm allows us to assess if aligning supervisors' incentives with the objectives of the principal helps mitigate supervisor biases in purely subjective evaluations and/or improves supervisor performance and practices.

*Control Arm:* In tehsils assigned to the Control arm, no bonus payments are made. Staff continue

to obtain travel and daily allowances as per the existing claim system, with no check on accuracy. ADs in this arm have access to the dashboard for uploading farmer training programmes, approving leaves, updating HR details, but do not have access to staff performance summaries as in the treatment groups.

Figure 4 below presents a simplified version of the programme Theory of Change to highlight differences between our incentive schemes. The figure shows that to conduct extension service delivery, extension frontline staff and supervisors both make inputs. While all incentive schemes offer incentives to frontline workers, it is only the Subjective Plus arm that incentivizes ADs as well through an additional layer of monitoring.

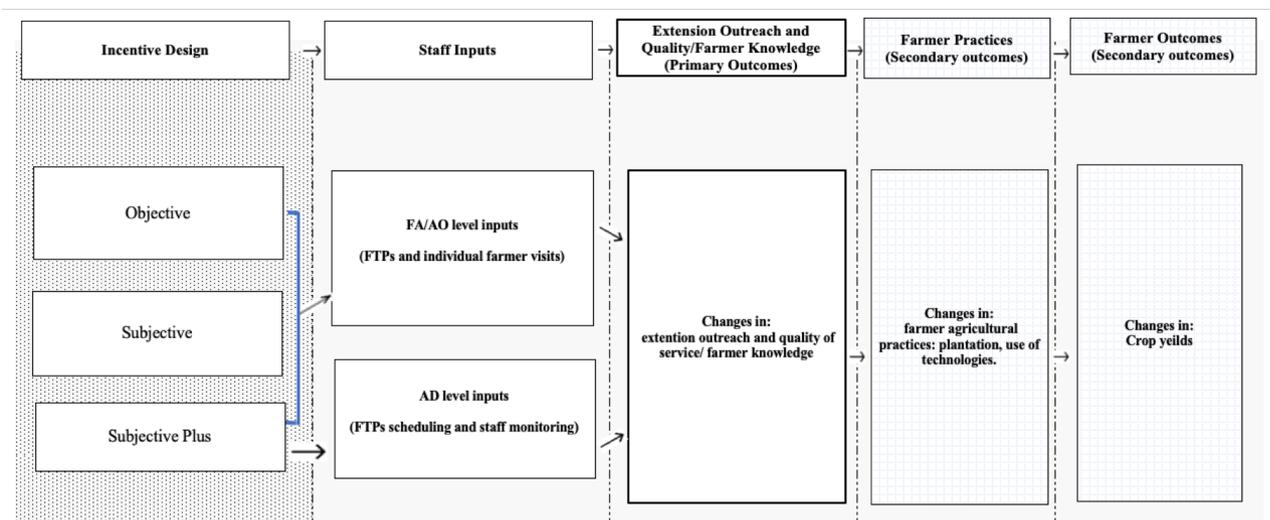


Figure 4: SIMPLIFIED THEORY OF CHANGE

## 5.2 Randomization and Balance Checks

The experiment is implemented across the entire province of Punjab. Stratifying by district, we randomize the 126 tehsils in the province into one of the three treatment arms, and a control group. We check for balance across field staff characteristics (such as age, salary, and years of

experience), baseline performance metrics of field staff that were gathered through the AgriSmart system (such as days worked, hours worked, village visits, and farmer visits), and tehsil level variables (distribution of TADA payments, proportion of villages in tehsils with access to metal road). Table A.2 in Appendix A presents the results for equality of means between different experimental arms for each of these variables as at March 2018 prior to the roll-out in April 2018. We find that treatments are balanced across a range of field staff and tehsil level variables. Only three test fails the equality of means test, however that is a failure rate of less than 5% which is well within the acceptable range.<sup>6</sup>

During the course of the programme, attrition took place if workers transferred out of extension or retired. Table A.3 shows that the treatments are also balanced on these variables for the ITT employees and Table A.4 and Table A.5 show that attrition is not related to any of our treatment groups. We are not concerned with substantial spillovers because randomization is at the tehsil level (creating substantial geographic separation across treatment arms) and the department rarely redistributes resources across tehsils. However, staff can be transferred across tehsils which we are able to track. In case of transfers, we rely on ITT estimates.

### 5.3 Implementation and Policy Challenges

The experiment was rolled out in April 2018, starting with department wide trainings to 2,596 FAs and AOs, and the 126 ADs at the tehsil level. The trainings took place treatment-wise and sequentially across all districts and included: 1) an overview of the Agrismart application along with a technical training on how to use the application for daily log of activities; 2) a review of the performance-based incentive programme. The treatment-wise training for ADs focused on the use of the dashboard to carry out specific managerial functions such as approval of leaves, assigning FTPs, and assigning payments to field staff in the Subjective and Subjective Plus arms.

Figure 5 below presents the project timeline which highlights key aspects during the last two

---

<sup>6</sup>The joint F tests are also all greater than 0.19.



payments being approved and filed by ADs given the loss of discretion over what was seen earlier as a fairly flexible budget, the dip in the budget led to lower payments of bonuses than expected. However, the payments against Agrismart were consistently logged on the dashboard and were seen as an accumulating liability to the extension agents.

In 2019-2020, budget allocation improved with a 75% jump (see Figure 6) as did the payments against bonuses (see Figure 7). At the same time, the department began working towards a sustainable policy change that could streamline the distribution of bonuses from the TADA budget after the evaluation was completed. This included the approval of a separate bonus head through which part of the TADA budget could be utilized for bonus payments (see Appendix E for details of the approved policy change).

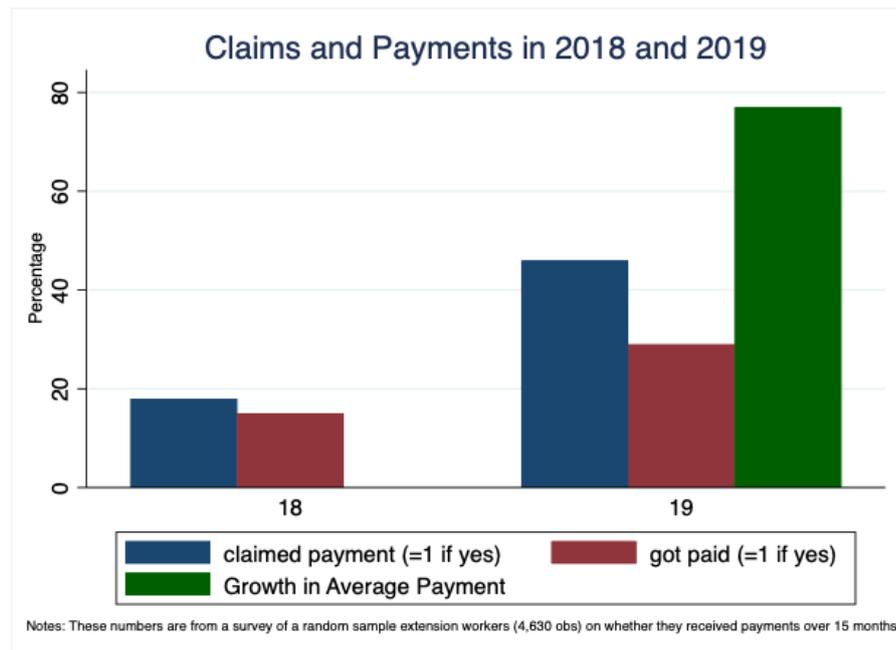


Figure 7: CLAIMS AND PAYMENTS

## 6 Performance Measurement and Data Sources

### 6.1 Primary and Secondary Outcomes

A well-designed incentive scheme should improve the outreach and quality of agricultural extension services delivered by extension agents. Ultimately, improved delivery of agricultural extension services would also help raise yields and improve farmer practices (if quality public extension is indeed causally related to improved farmer practices).

Our primary outcomes of interest are access and quality of public extension to farmers. This includes measures of extension outreach - such as days and hours worked, villages visited, farmers visited, farmer training programmes (FTPs) conducted - as captured by AgriSmart, and wider measures of access (that may not be captured on AgriSmart) such as whether farmers know the AO/FA and use public extension. For quality of extension, we consider measures such as farmer feedback on the usefulness of extension services, quality of information provided, and farmer knowledge. Since we are also interested in capturing the long-term effects of quality public extension on farmer practices, we also aim to collect measures of yield for all major and minor crops and farmer practices such as use of inputs, pesticides, seeds, and fertilizers, but see these as our secondary outcomes.

To understand causal mechanisms for the impact of our treatments, we collect a range of measures of behaviors and attitudes at the AD (tehsil) and FA/AO level. At the AD level, we capture AD management practices, job satisfaction and motivation, perceptions of the performance management programme, sources of information for AD bonus assignment behaviour (in Subjective arms), and basic compliance of ADs with their Agrismart SOPs such as scheduling and assigning FTPs and approving leaves. At the FA/AO level, we capture job satisfaction and motivation, their attitudes and behaviours on the job, and perceptions regarding the performance management system.

Finally, for heterogeneous treatment effects we capture a range of baseline measures at the AD,

FA/AO, village, and farmer level. At the AD level, these include measures of basic AD profile (such as age, gender, years of experience), AD management practices, career ambition, personality types, and experimental survey games that measure pro-sociality and tendency to cheat. At the FA/AO level, we collect measures of their profile, job satisfaction, pro-sociality, personality types, self-efficacy, and prior levels of TADA claims and payments.<sup>7</sup> At the farmer/village level, we collect measures of land size, education level, and accessibility to metal roads.

## 6.2 Data Sources

We use a combination of administrative and survey data to get at our measures of interest.

*AgriSmart.* Our main source of extension outreach information is the administrative smartphone data collected through the AgriSmart application that field staff use on a daily basis. The AgriSmart dashboard also calculates measures at the AD level such as whether they assigned FTPs, approved leaves, and routinely assigned payments to staff (in Subjective and Subjective Plus) which we refer to as the AD SOPs. All extension outreach data (and AD SOPs) is collated at the employee (or AD) month level which gives us a monthly employee panel over the period under analysis.

Given the incentive programme relied on the uptake of technology across extension staff (especially for the Objective arm), descriptive statistics in Table A.1 show that the uptake of the AgriSmart system among extension staff was high, with greater than 90% of staff regularly reporting their time use on the system. Non-reporting staff was mostly the result of connectivity issues, tablet maintenance, or loss of tablets.

*Audit.* To corroborate the veracity of the reported data in AgriSmart and to disincentivize misreporting, there is also a monthly farmer phone survey to farmers whose numbers are reported in the Agrisart data by extension staff – we call this the AgriSmart audit. The survey is

---

<sup>7</sup>We capture personality traits using the Big Five Inventory John et al. (2008); pro-social motivation is captured using Perry (1996).

conducted by randomly selecting one farmer reported by each extension worker every month. The audit questionnaire is a short survey that includes questions like whether the farmer knows the AO/FA and whether the activity reported in the smartphone data was indeed conducted with the reported farmer.

*Farmer Surveys.* To capture our treatment effects on access and quality of extension, and more broadly on farmer practices, we use a government dataset of 1.8 million farmers that was collated across Punjab as part of a province-wide farmer registration exercise in 2019. Using this database of “external” farmers as our sampling frame, we conduct monthly call back surveys to farmers 18 months into the programme, starting January 2020. The survey questions ask about the quality of extension in the farmer’s village, access to public extension, and farming practices. This survey was paused in March 2020 due to the COVID-19 outbreak but will be resumed when the current situation allows. This survey will continue up until December 2020 around which the evaluation endline is planned.<sup>8</sup>

*Staff surveys.* We also conduct an additional monthly survey of three randomly selected extension staff from each tehsil. Covering around 15% of the total staff strength, this monthly survey includes questions on satisfaction with the AgriSmart system and whether payments against bonuses were made. The latter question is important to track since the bonuses are paid from a prior TADA budget head and despite notifications from the central government, lapses in payments were expected. We are currently in the process of collecting administrative data on payments to all staff through the TADA budget head in the financial year 2019-2020 which will give us a broader coverage.

*Baseline and endline surveys.* For heterogenous treatment effects, we conducted a baseline of FAs/AOs and ADs prior to the roll-out of the incentive programme. To capture causal mechanisms, we planned to roll-out a shorter mid-line to FAs/AOs and ADs in March 2020 followed by an endline in December 2020. Our plans for a midline were paused due to COVID-

---

<sup>8</sup>We conduct a range of checks on this farmer database including whether the database has balance across treatments and whether it is representative across tehsils by comparing it with the Punjab Agriculture Census (See Appendix D for details).

19. We plan to launch our endline survey with FA/AOs and ADs in November 2020.

Appendix D includes a detailed note on the sampling strategy for the various monthly surveys.

## 7 Results

In this section, we discuss preliminary results on our primary and secondary outcomes based on the outcome data collected since the implementation of the programme in April 2018.

### 7.1 Treatment Effects on Extension Outreach - AgriSmart Data

To estimate the causal impact of our treatments on AO/FA extension outreach, we use outreach indicators as captured in the AgriSmart data, which is a monthly staff-level panel from April 2018 to March 2020. To estimate treatment effects, we do an ANCOVA estimation as shown below:

$$y_{itdm} = \alpha + \gamma \underline{y}_{itd} + \sum_j \beta_j T_{itd}^j + \zeta X_{itd} + \nu_d + \mu_m + \epsilon_{itdm} \quad (4)$$

Where  $y_{itdm}$  is our AgriSmart outreach indicator (number of days and hours worked, village visits, farmer visits, FTP conducted, or an overall AgriSmart performance index) for extension worker  $i$  in tehsil  $t$ , district  $d$ , and month  $m$ ;  $\underline{y}_{itd}$  is the baseline indicator of  $y_{itdm}$  for the ANCOVA estimation;  $T_{itd}^j$  is the treatment indicator that takes the value 1 if the tehsil  $t$  of extension worker  $i$  is assigned to treatment  $j$  and 0 otherwise;  $X_{itd}$  is a vectors of AO/FA and tehsil controls at baseline; and  $\nu_d$  and  $\mu_m$  are the district fixed effects and month dummies. Standard errors are clustered at the tehsil level, the level of randomization of our treatments (Abadie et al., 2017).  $\beta_j$  is the coefficient of interest for each of our treatments.

Table 1 shows positive and significant treatment effects on number of days worked, hours worked, village visits, individual farmer visits, and the overall AgriSmart performance index in all three treatment arms. For compliance measures such as days and hours worked (columns 1 and 2), we find that in comparison to the control group, AOs/FAs in the Objective arm work almost a day more and an hour longer daily each month on average, AOs/FAs in the Subjective arm work half a day more and 30 minutes longer daily on average, and AOs/FAs in the Subjective Plus arm work half a day more and an hour longer daily on average. For outreach indicators such as village visits and individual farmer visits (columns 3 and 4), the Objective arm does significantly better than both the Subjective and Subjective Plus arms. In comparison to the control group, AOs/FAs in the Objective arm make 9 more village and 18 more farmer visits each month whereas AOs/FAs in the Subjective arm make 3 more village and 6 more farmer visits, and AOs/FAs in the Subjective Plus arm make 4 more village and 9 more farmer visits. Completion of Farmer Training Programmes (FTPs), which is a key task for the department, is only significant for the Subjective Plus arm (column 5) - AOs/FAs in the Subjective Plus arm schedule and complete at least one additional FTP as compared to the control group. To address challenges around multiple hypothesis testing, an overall AgriSmart performance index normalizes each AgriSmart indicator by the mean and standard deviation of the control group and then computes an average of all normalized measures (as proposed by Anderson, 2008). Treatments effects on the performance index show that in comparison to the control group, the Objective arm improves AgriSmart outreach by  $0.42\sigma$ , the Subjective arm by  $0.19\sigma$ , and the Subjective Plus arm by  $0.29\sigma$  (column 6). The higher treatments effects on the Objective arm are significantly different from the Subjective and Subjective Plus arm. These treatments effects persist after q-value adjustments which are also presented in Table 1.

The higher treatment effects in the Objective arm are consistent with our second hypothesis which predicts such effects given the incentives under this arm are explicitly linked to AgriSmart measures as opposed to the Subjective and Subjective Plus arm. At the same time, the positive effects in the Subjective and Subjective Plus arm suggest that even though these incentives were not directly linked to the Agrismart measures, the fact that ADs had access to these measures for making bonus assignments might have motivated staff to improve their performance on these

measures. Finally, notice that FTP scheduling is an AD-level task. The fact that the design of the Subjective Plus arm induces a layer of top-down monitoring on the ADs can explain why the significant treatment effects on FTP completion are only observed in this arm.

Although we observe significant treatment effects on AgriSmart outreach indicators, we also test whether these effects stay consistent or diminish overtime. Quarter by quarter treatment effects in Table A.7 show that these effects stay consistent over time (See Figure 8 below). Figures B.1 and B.2 in the appendix show time trends of key outreach indicators.

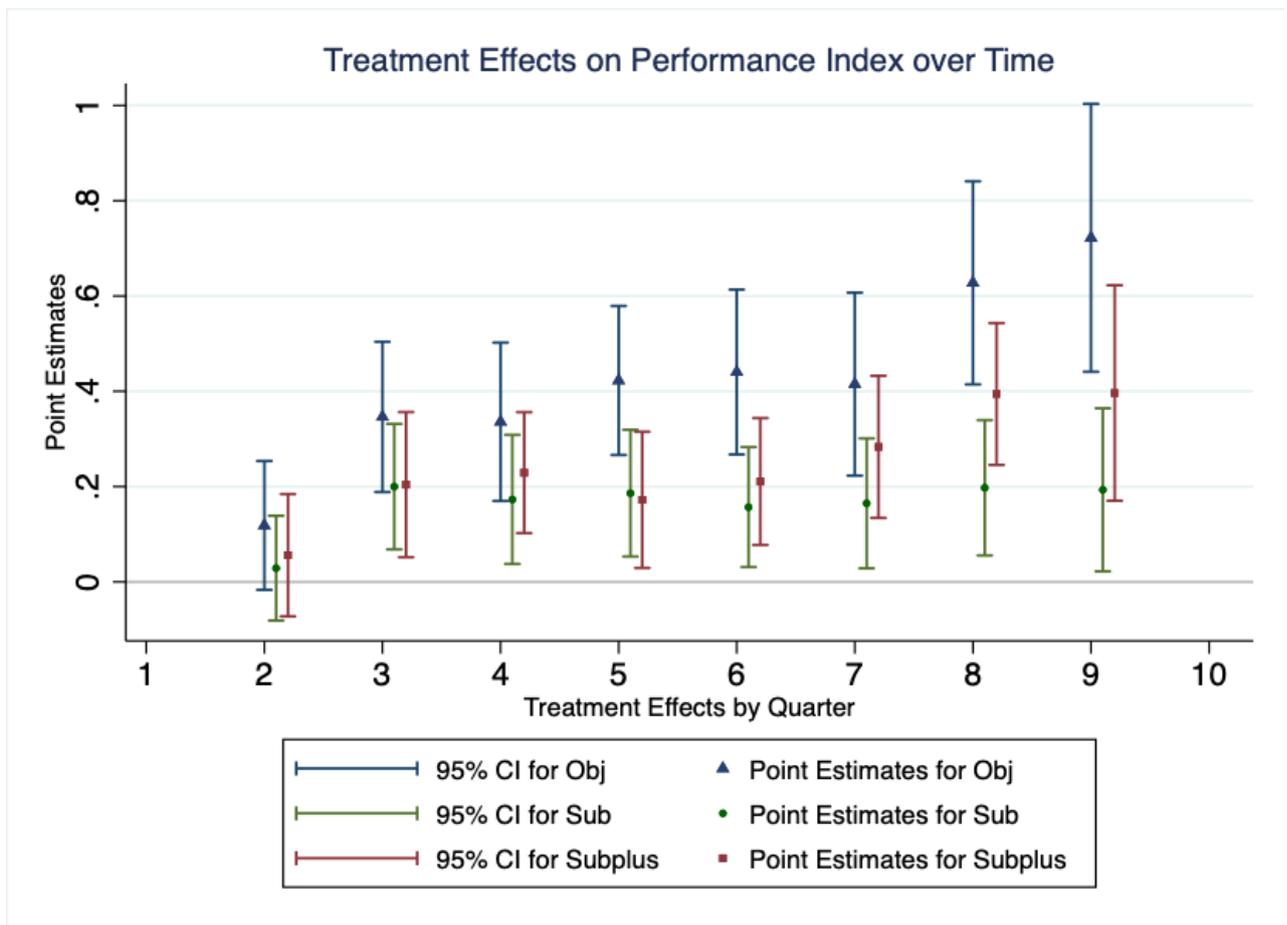


Figure 8: TREATMENT EFFECTS ON PERFORMANCE INDEX

While extension activities are self-entered into the smartphone application by extension agents, there are several built-in checks such as time-stamped activities, geo-tagged activities and photos, automated distance calculation, and monthly audits of the AgriSmart data which disincentivize misreporting. Despite these checks to discourage misreporting, the following questions remain: a) staff across all treatments may have misreported data by adding erroneous village visits or farmer phone numbers; b) staff in the Objective arm may have tried to game the system by putting in just enough effort to meet thresholds for bonuses; and c) staff across all treatment may have simply adopted “better” reporting behaviour on the AgriSmart system instead of actually improving outreach. In Appendix F, we leverage the large volume of reported farmer phone numbers and staff activity database in the Agrismart data along with farmer responses in the Agrismart audit surveys to rule out a) and b). Further information on AO/FA AgriSmart reporting behaviour through the endline survey will aim to assess whether c) could be driving our results.

## **7.2 Treatment Effects on Extension Outreach and Quality - Farmer Surveys**

To estimate the causal impact of our treatments on farmer experience of extension access and quality, we use monthly data from our (external) farmer surveys that were administered after 18 months into the programme. These measures are distinct from the AgriSmart outreach measures in two ways. First, they capture *quality* of extension which is not captured under AgriSmart. Second, the outreach measures under the farmer surveys (such as whether the farmer knows the AO/FA or attended an extension meeting) essentially measure extension access and spread given these surveys are based on random selection of farmers from the AO/FA village who may or not be have been part of the AO/FA regular visitation lists. To estimate treatment effects, we use the following estimation:

$$y_{itdm}^{farmer} = \alpha + \sum_j \beta_j T_{itd}^j + \zeta X_{itd} + \nu_d + \mu_m + \epsilon_{itdm} \quad (5)$$

where  $y_{itd}^{farmer}$  is an access or quality indicator based on farmer experience for farmer  $i$  in tehsil  $t$ , district  $d$ , and month  $m$ ;  $T_{itd}^j$  is the treatment indicator that takes the value 1 if the tehsil for farmer  $i$  is assigned to treatment  $j$  and 0 otherwise;  $X_{itd}$  is a vector of AO/FA, farmer, and village controls in tehsil  $t$  and district  $d$ ; and  $\epsilon_{itdm}$  is the error term. As before,  $\nu_d$  and  $\mu_m$  are the district fixed effects and month dummies, and we cluster errors at the tehsil level. While we do not have baseline data for these outcomes, we check for balance on a range of fixed farmer characteristics and find that we have balance across key farmer variables (see Table A.8).

Table 2 shows significant treatment effects on farmer outreach and quality in the Subjective Plus arm. To capture farmer reported outreach, we capture data across a range of indicators such as whether the farmer knows the AO/FA, whether the farmer attended an extension meeting, or whether public extension is the primary source of information for the farmer. We compute a farmer outreach index that normalizes each of these outreach indicators by the mean and the standard deviation of the control group and then compute an average of all normalized measures (as in Anderson, 2008). We find positive and significant treatment effects of  $0.10\sigma$  on the overall farmer outreach index in the Subjective Plus arm, with these effects being significantly different from the Objective and Subjective arm (column 1). We also find that out of those who attend extension meetings, farmers in the Subjective Plus arm are 14% points more likely to report the meetings to be useful as compared to the control group. These treatment effects persist after adjusting q-values for multiple hypothesis testing across treatments (see notes in Table 2).

We do not find any significant effects in the Objective or the Subjective arm. The null effects in the Objective arm imply that AgriSmart measures of extension are neither complements nor substitutes of outreach and quality measures that are based on farmer experience. They also indicate that staff in the Objective arm may have simply improved outreach on the intensive margin (i.e. making more visits to the same set of farmers) instead of the extensive margin (i.e. expanding outreach to a greater set of farmers). Given the only difference between the Subjective

and Subjective Plus arm is the AD-level monitoring layer in the latter, the null effects on the Subjective arm indicate differences in AD behaviour between the Subjective and the Subjective Plus arms - both in terms of AD performance on key extension tasks such as scheduling of FTPs and AD bonus assignment behaviour. In addition, these differences in treatment effects also imply possible differences in how staff motivation and job satisfaction may have been effected across different treatment arms. We address these question in more detail in section 7.5 where we discuss the mechanisms underlying our results.

Overall, these results highlight that across measures of extension outreach and quality, the Subjective Plus arm works better than the Objective and Subjective arms. However, it is not clear whether these effects in the Subjective Plus arm are driven by better AD performance on tasks that form an essential input into extension outreach and quality or by better bonus assignment behaviour (or both). We explore this further in Sections 7.4 and 7.5

### 7.3 Treatment Effects on Farmer Practices and Yield

In this section, we report the treatment effects of our incentive schemes on our *secondary outcomes* - farmer practices (such as cropping techniques and use of technology) and yield of major crops.<sup>9</sup> We use the monthly data from our (external) farmer surveys for these outcomes and use the same estimation as (5).

Table 3 reports an encouraging set of results for the Subjective Plus arm. Column 2 shows that in comparison to the control group, farmers in the Subjective Plus arm are 4% points more likely to change their prior agricultural practices such as use of inputs/seeds and plantation practices. In addition, we also find a marginally significant impact on yield which is relatively harder to move. We find no significant treatment effects on the Objective or the Subjective arm. These results are consistent with the effects we find on farmer experience of extension outreach and quality in the section above.

---

<sup>9</sup>Based on the Pakistan Agriculture Census (2010), more than 90% of farmers grow wheat which is why we look at yield of wheat.

## 7.4 Treatment effects on AD Performance

In this section, we investigate whether the positive treatment effects in Subjective Plus are driven by better AD performance on tasks that form an essential input into extension outreach.

The AgriSmart programme captures two key performance metrics of ADs. First, whether ADs schedule and assign FTPs to extension staff. This directly impacts the number of Farmer Training Programmes (FTPs) that are implemented by extension workers. Second, whether ADs approve leaves submitted by staff. Given approved leave days are counted as present days for staff in the system, this directly effects whether staff is able to meet their monthly compliance measures for the Objective arm to qualify for base pay or any bonus. We use monthly panel data from the AgriSmart programme to compute these measures of AD performance. To estimate the impact of our treatments, we estimate the following equation:

$$y_{tdm}^{Ad} = \alpha + \sum_j \beta_j T_{tdm}^j + \zeta X_{td} + \nu_d + \mu_m + \epsilon_{tdm} \quad (6)$$

Where  $y_{tdm}^{Ad}$  is the AD performance measure for AD in tehsil  $t$ , district  $d$ , and month  $m$ .  $T_{tdm}^j$  is the treatment indicator that takes the value 1 if tehsil  $t$  is assigned to treatment  $j$  and 0 otherwise;  $X_{td}$  is a vector of AD and tehsil controls in tehsil  $t$  and district  $d$ ; and  $\epsilon_{tdm}$  is the error term. As before, we include district fixed effects, month dummies, and cluster errors at the tehsil level. Given the tehsil reports in Subjective Plus were introduced 9 months into the programme, we also estimate a variant of (6) with pre report and post report Subjective Plus treatment effects.

Table 4 shows that the treatment effects on FTP scheduling are positive and significant for the Subjective Plus arm. In comparison to the control group, ADs in Subjective Plus assign almost 15% points more FTPs to FAs and 12% points more FTPs to AOs (columns 4 and 5). On the other hand, treatment effects for ADs in the Objective arm are only marginally significant with around 6% points more FAs and AOs being assigned FTPs while the effects are insignificant in

the Subjective arm (columns 4 and 5). Where leave approval is concerned, in comparison to the control group ADs in the Objective arm approve 18% points more leaves while ADs in the Subjective Plus arm approve around 11% points more leaves (column 2). These effects persist after q-value adjustments for multiple tests across treatments (see notes in Table 4). Once again, we find no treatment effects for ADs in the Subjective arm.

The significant effects on leave approval (and the marginally significant effects on FTP scheduling) in the Objective arm are consistent with the explanation that staff may have created pressure on the ADs to perform given the incentives in this arm are directly linked to these measures. The strong treatment effects on FTP scheduling observed in Subjective Plus are indicative of how the additional layer of top-down monitoring report worked in aligning AD performance with the AgriSmart programme and the broader goals of extension. This is also evident in the p-value differences in the treatment effects of the pre-report and post-report periods of the Subjective Plus arm that are shown in Table 4.

## 7.5 Mechanisms

In this section, we begin to explore why Subjective Plus has consistent treatment effects across different measures of extension outreach and quality (both as measured through AgriSmart and farmer experience). Given the key difference in the Subjective Plus arm is the AD-level layer of monitoring, we hypothesize two potential channels of AD behaviour to understand our results. First, AD's own management practices and performance may have improved, and these may have had effects on the performance of extension agents. Second, ADs in the Subjective Plus arm may have made more efficient bonus allocations resulting in improved performance of extension agents. We provide some evidence for the first channel in section 7.4 above. In this section, we provide some preliminary evidence on the second channel.

We also try to understand underlying mechanisms for how the Objective versus Subjective schemes work. In particular, we explore two potential mechanisms. First, we hypothesize

whether the Objective arm, that explicitly incentivizes *number* of farmer visits and/or trainings, encourages more effort on the intensive margin (i.e. making more visits to the same set of farmers) compared to the Subjective arms which may encourage effort on both the intensive and the extensive margin (i.e. expanding extension outreach to additional farmers). Second, we explore if staff in the Objective arm is more dissatisfied due to the rigidity or inflexibility of the scheme's formula (as implied by Rajan and Reichelstein, 2009) versus the Subjective arms.

### 7.5.1 AD Bonus Assignment Behaviour

Our conceptual framework in Section 3 hypothesizes that an AD level layer of monitoring in the Subjective Plus arm should reduce the weight of personal preferences of ADs over bonus allocations as compared to the Subjective arm. In this section, we explore mechanisms to understand if AD bonus allocation behaviour differs between these two arms.

*Distribution of Bonuses across Arms.* We begin by assessing the average amount of bonus that is assigned across treatment arms. Table 5 shows that while the Objective arm allocates 63% of staff to base, 3% to low and 4% to high bonus, and 31% to no payments, the proportion of low and high bonus is much higher in the Subjective and Subjective Plus arms. The Subjective arm assigns 37% to base, 34% to low bonus, 20% to high bonus, and 8 % to no payments; while the Subjective Plus arm assigns 34% to base, 39% to low bonus, 21 % to high bonus, and 9% to no payment. Column 1 in Table 6 shows that the average bonus amount assigned in the Subjective and Subjective Plus arms is almost double of what is assigned in the Objective arm. While this is consistent with the literature on subjective evaluations which argues that supervisors can often be subject to a leniency bias (Gibbs et al., 2004), in our context it is unclear if the high amounts of bonus allocation represent leniency or supervisors assigning bonuses more efficiently (especially in the Subjective Plus arm).

*AD Bonus Assignment Behaviour in Subjective and Subjective Plus.* While ADs in both the Subjective and Subjective Plus arms assign similar amounts of bonus on average, in this section we try to to understand whether ADs differ in how they make bonus assignments in these two

arms. We use the monthly panel data (over a two-year period) on bonus allocations of staff that is routinely recorded in the AgriSmart system to construct two types of indicators. First, we construct an employee-level *bonus variation* indicator which calculates the variance in each individual's bonus over the two-year period. Second, we construct an AD-level *tendency to repeat* indicator which first estimates an initial set of bonus recipients in the first three months of the programme and then calculates the proportion of individuals that are repeated every month from this initial set of bonus recipients in the following months. While a low variance in individual bonus or a high proportion of repeat individuals could indicate bias, it could also be indicative of repeat assignments to high ability individuals. Irrespective, we use these indicators to explore whether there is a difference in the bonus assignment behaviour of ADs in the Subjective and Subjective Plus arm.

Column 2 in Table 6 shows that the coefficient on the variance of individual bonuses is higher in the Objective and Subjective Plus arms as compared to the Subjective arm, with the coefficients of Subjective and Subjective Plus arm being significantly different at the 10% level. The variance in the bonus assignments of staff in the Subjective Plus arm is in fact closer to the Objective arm as opposed to the Subjective arm. This indicates that supervisors' in the Subjective Plus arm may have exercised their discretion differently as compared to the Subjective arm. Column 3 in Table 6 shows that while ADs in the Subjective Plus arm repeat 65% of their high bonus assignments to bonus recipients from the first three months, ADs in the Subjective arm repeat 90% of their high bonus assignments. These proportions are also significantly different from each other at the 1% level (see Figure 9 below). While we cannot confirm whether the difference is due to bias or rewarding high ability/performance, the results in column 2 and 3 indicate differences in the bonus allocation behaviour between the Subjective and Subjective Plus arm ADs which require further inquiry.

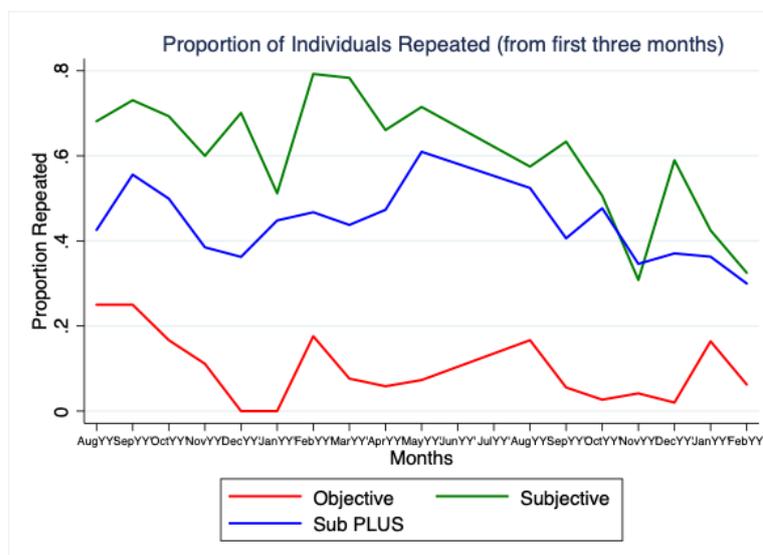


Figure 9: PROPORTION OF INDIVIDUALS REPEATED FOR BONUS

Endline data will aim to collect detailed information on AD preferences for making bonus allocations. This will include AD and staff data on caste and ethnic background, AD self-reported data on metrics and information against which bonuses were rewarded, and additional information on quality of service delivery. Future analysis will explore whether the differences in bonus assignment behaviour presented above are due to AD preferences for bias versus rewarding on quality, and if the latter is true, AD valuation of objective versus subjective sources of information for making assignments.<sup>10</sup>

### 7.5.2 Objective versus Subjective Schemes

*Outreach Effort on the Intensive and Extensive Margins.* While the Objective arm is explicitly linked to number of farmer visits and/or trainings provided, the Subjective arms provide super-

<sup>10</sup>Preliminary analysis in Table A.13 tests whether Objective information in AgriSmart predicts bonus assignments by ADs in the Subjective and Subjective plus arms. Currently, we find that Objective information predicts bonus assignments in both arms.

visors the flexibility to capture (additional) harder to measure dimensions of performance. We explore whether these treatment design differences encourage more effort on the intensive margin for the Objective arm and more effort on the extensive margin for the Subjective arms.

To explore this hypothesis, we leverage the large AgriSmart database of over 4 million farmer phone numbers that are entered by staff while conducting individual farmer visits and group-based FTPs. We use this database to construct four types of indicators - total farmer numbers reported by staff, total unique numbers reported, total unique numbers reported in the post treatment period on top of the pre-treatment numbers, and average number of times a number is repeated. While the first three indicators provide indication of effort on the extensive margin, the last indicator provides evidence of effort on the intensive margin.

Columns 1-3 in Table 7 show evidence of greater staff effort on the extensive margin across all treatments, but with two interesting takeaways. First, we find that the Subjective Plus arm has the highest positive and significant coefficient on total unique farmers reached and unique farmers reached on top of the pre-treatment farmer set (columns 2 and 3). This is followed by the Objective arm which is also positive and significant, and then by the Subjective arm which is only marginally significant. While AgriSmart measures of individual farmer visits and village visits in Table 1 showed higher treatment effects in the Objective arm as compared to the other two arms, the results in this section indicate that with individual farmer visits and FTPs combined, staff in the Subjective Plus arm may have exerted higher effort to expand their farmer outreach. Second, we find that while the average number of times a farmer number is repeated is quite low across all treated groups (see column 4), the likelihood of the Objective arm to repeat a number is significantly higher than the Subjective and the Subjective Plus arm (see p-value differences in the Table 7). This provides indicative evidence for our hypothesis that the Objective arm may have encouraged higher effort by staff on the intensive margin in comparison to the other treatments. Future analysis will exploit the project GIS data to analyse the frequency of activities in specific geographical zones to explore this hypothesis further.

*Staff Satisfaction.* We are also interested in understanding how the Objective and Subjective incentive schemes may have effected staff motivation and satisfaction. Objective metrics tend to

be rigid and can impose undue risk on agents (Rajan and Reichelstein, 2009), whereas subjective metrics tend to be more flexible.

We use our monthly extension staff surveys that were initiated in January 2019 to understand staff satisfaction with the AgriSmart system. With regards to staff satisfaction, the survey includes two short questions - first, whether staff believe AgriSmart helps improve extension outreach; and second, the extent to which staff feel satisfied with the system. For the first question, we construct a dummy variable that equals 1 if staff is satisfied with the programme; for the second question, we normalize the categorical satisfaction variable by the mean and standard deviation of the control group and observe treatment effects in standard deviation units.

Table 8 shows that staff in the Objective arm show  $0.33\sigma$  lower satisfaction with the AgriSmart system as compared to the control group (column 2), with these effects being significantly different from the Subjective and Subjective plus arm. On the other hand, we find no significant treatment effects on the Subjective or the Subjective Plus arm. These results are consistent with the literature on how objective metrics are unable to cater to shocks and can impose undue risks on staff resulting in no response to incentives (Rajan and Reichelstein, 2009).

Our endline survey data for AOs and FAs will further explore perceptions around the advantages and/or disadvantages of each of the incentive schemes in more detail. We will also explore whether the rigidity imposed by the Objective scheme can, in part, explain why we do not observe treatment effects on outreach and quality captured through farmer-level surveys.

## 8 Conclusion

This paper evaluates the impact of a large scale randomized controlled trial in Punjab, Pakistan that aims to measure the impact of three different pay-for-performance schemes on extension outreach and quality. Leveraging the extension department's comprehensive digital performance management system called AgriSmart, the incentive schemes link incentives to objective metrics on AgriSmart (Objective arm), supervisors' own subjective evaluation (Subjective arm), and supervisors' own subjective evaluation with an element of top-down monitoring to align supervisors' incentives with the objectives of the principal (Subjective Plus).

Our results show that while all treatments improve performance on AgriSmart outreach measures, Subjective Plus also improves performance on farmer reported experience of extension outreach and quality, and farmer-level outcomes. However, these effects are not observed in the Objective or the Subjective arm. In addition, supervisors (ADs) in Subjective Plus also shows positive and significant effects on their managerial task of scheduling farmer training programmes (FTPs), which is an essential extension activity carried out by frontline staff. The evidence on Subjective Plus, however, is not entirely conclusive as to whether the positive treatment effects are driven by checking bias of supervisors in their bonus assignments or via improving supervisors' own performance and management practices, or both. For example, if the impact observed in Subjective Plus is primarily driven through the AD performance and practices channel instead of incentives for frontline workers, this could have significant cost implications for the government.

Further data collection will explore additional mechanisms underlying our results - such as AD preferences for assigning bonuses, additional measures of AD performance, measures of staff motivation and satisfaction across treatment arms, and whether the schemes induced different types of trade-offs (such as effort on the extensive versus intensive margins or few longer versus many shorter extension trainings) and their related impacts on farmer practices and outcomes. The choice of the incentive scheme will ultimately depend on the government's objectives for extension service delivery and the costs and benefits associated with such trade-offs.

Ultimately, our results will have direct implications for incentives in public services where staff is hard to monitor due to their spatial spread, where supervisors hold important contextual information, and where tasks have both easily measurable and hard to measure dimensions. Our results will also contribute to the literature on rules-based bureaucracy versus discretion. In particular, it will show circumstances under which subjective evaluations are effective or ineffective, how supervisors use their private sources of information versus objective sources of information, and the kind of trade-offs induced by purely objective metrics versus subjective metrics.

Table 1: Treatment Effects on Extension Outreach - AgriSmart Measures

	(1)	(2)	(3)	(4)	(5)	(6)
	Days Worked	Hours Worked	Village visits	Individual Farmer Visits	FTP Completed	Performance Index
Objective	0.952*** (0.266)	0.988*** (0.259)	9.239*** (1.719)	18.634*** (3.150)	0.765 (0.615)	0.423*** (0.065)
Subjective	0.721*** (0.262)	0.548** (0.214)	3.279*** (1.085)	6.591** (2.609)	0.573 (0.542)	0.186*** (0.056)
Subjective Plus	0.765*** (0.287)	1.018*** (0.238)	4.768*** (1.187)	9.473*** (2.424)	1.639*** (0.579)	0.290*** (0.052)
<i>Mean of dep. var</i>	21.159	9.530	19.782	22.058	4.871	-0.004
<i>P-Value Differences</i>						
Obj-sub	0.41	0.03**	0.00***	0.00***	0.76	0.00***
Obj- subplus	0.54	0.89	0.02**	0.01***	0.19	0.07*
sub - subplus	0.89	0.01***	0.23	0.35	0.05**	0.09*
Observations	49157	49157	49157	49157	19315	49157
Adjusted R <sup>2</sup>	0.405	0.233	0.219	0.133	0.026	0.176
Controls	No	No	No	No	No	No
District Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Standard errors clustered at the tehsil level (the unit of randomization) are within parenthesis. All regressions are an ANCOVA estimation with baseline values of the dependent variable, district FE, and year-month dummies. Regressions are based on ITT. The dependent variables in the first 5 columns are based on a monthly average. The performance index in column 6 converts all AgriSmart measures (except FTP completed) into z scores and computes their average. Regressions are robust to inclusion of controls (presented in Table A.5). We compute FDR adjusted q values for the multiple tests across treatment arms for the performance index and the FTP completed variable which show that the treatment effects remain consistent (adjusted q values for the performance index: treatments - 0.001, 0.002, 0.001; tests across treatments - 0.003 0.084, 0.09; adjusted q- values for FTP completed: treatments - 0.324, 0.352, 0.03; tests across treatments - 0.76, 0.32, 0.153). Estimates are significant at the \*10%, \*\*5%, and \*\*\*1% level

Table 2: Treatment Effects on Extension Access and Quality - Farmer Surveys

	(1)	(2)	(3)	(4)
	Farmer Outreach Index	Meeting Useful	Length Meeting	Knowledge Score
Objective	-0.025 (0.030)	-0.003 (0.038)	-4.800 (5.999)	0.140 (0.089)
Subjective	0.021 (0.028)	0.051 (0.037)	-4.610 (5.449)	0.011 (0.085)
Subjective Plus	0.096*** (0.033)	0.141*** (0.036)	-1.501 (5.359)	0.072 (0.088)
<i>Mean of dep. var</i>	-0.015	0.633	80.553	0.061
<i>P-Value Differences</i>				
Obj-sub	0.14	0.17	0.96	0.09*
Obj- subplus	0.00***	0.00***	0.50	0.39
sub - subplus	0.04**	0.01***	0.48	0.42
Observations	6835	1057	1057	3309
Adjusted R <sup>2</sup>	0.009	0.009	0.007	0.018
PDS LASSO controls	No	No	No	No
District Fixed Effects	Yes	Yes	Yes	Yes

Notes: Standard errors clustered at the tehsil level (the unit of randomization) are within parenthesis. All regressions include district FE and year-month dummies. Regressions are based on ITT. Regressions are robust to inclusion of controls (presented in Table A.8). We compute FDR adjusted q values for the multiple tests across treatment arms (adjusted q values for farmer outreach index: treatments - 0.448, 0.448, 0.012; tests across treatments - 0.21 0.001, 0.08; adjusted q- values for Usefulness: treatments - 0.94, 0.20, 0.001; tests across treatments - 0.20, 0.001, 0.03). Estimates are significant at the \*10%, \*\*5%, and \*\*\*1% level

Table 3: Treatment Effects on Farmer Practices and Outcomes

	(1)	(2)	(3)
	Planted New Crop	Changed Practice	Wheat Yield (in kg)
Objective	-0.003 (0.010)	0.020 (0.016)	5.320 (37.667)
Subjective	-0.014 (0.013)	-0.003 (0.017)	21.694 (31.254)
Subjective Plus	0.016 (0.011)	0.037** (0.016)	80.544* (42.890)
<i>Mean of dep. var</i>	0.117	0.254	1272.14
<i>P-Value Differences</i>			
Obj-sub	0.37	0.10*	0.66
Obj- subplus	0.12	0.32	0.11
sub - subplus	0.02**	0.03**	0.15
Observations	6395	6395	5832
Adjusted R <sup>2</sup>	0.001	0.006	0.005
Controls	No	No	No
District Fixed Effects	Yes	Yes	Yes

Notes: Standard errors clustered at the tehsil level (the unit of randomization) are within parenthesis. All regressions include district FE and year/month dummies. Regressions are based on ITT. Regressions are robust to inclusion of controls (presented in Table A.8). We compute FDR adjusted q values for the multiple tests across treatment arms (adjusted q values for changed technology practice: treatments - 0.327, 0.855, 0.09; tests across treatments - 0.2, 0.384, 0.09). Estimates are significant at the \*10%, \*\*5%, and \*\*\*1% level

Table 4: Treatment Effects on AD Performance

	% Leaves Approved		FAs FTP Assigned		AOs FTP Assigned		Total Scheduled FTPs	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Objective	0.182*** (0.042)	0.182*** (0.041)	0.068* (0.038)	0.068* (0.038)	0.066* (0.035)	0.066* (0.034)	-0.005 (1.052)	0.005 (1.053)
Subjective	0.060 (0.044)	0.059 (0.043)	0.057 (0.039)	0.057 (0.039)	0.022 (0.039)	0.020 (0.038)	-0.136 (0.782)	-0.129 (0.782)
Subjective Plus	0.040 (0.043)		0.108*** (0.039)		0.061* (0.031)		1.415* (0.750)	
Before Tehsil Report x Sub Plus		-0.063 (0.055)		0.051 (0.044)		-0.015 (0.041)		0.939 (0.868)
After Tehsil Report x Sub Plus		0.105** (0.045)		0.145*** (0.043)		0.118*** (0.037)		1.704** (0.762)
<i>Mean of dep. var</i>	0.732	0.732	0.443	0.443	0.697	0.697	8.43	8.43
<i>P-Value Differences</i>								
Obj-Sub	0.00***	0.00***	0.79	0.79	0.20	0.18	0.88	0.88
Obj- Subplus	0.00**		0.33		0.87		0.12	
Sub - Subplus	0.60		0.17		0.27		0.02**	
Obj- report x Subplus		0.05**		0.08*		0.18		0.06*
Sub- report x Subplus		0.26		0.03**		0.03**		0.01***
Subplus - report x Subplus		0.00***		0.01***		0.01***	0.20	
Observations	2529	2529	2939	2939	1934	1934	2208	2208
Adjusted R <sup>2</sup>	0.103	0.113	0.135	0.139	0.164	0.168	0.057	0.057
PDS LASSO controls	No	No	No	No	No	No	No	No
District Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Standard errors clustered at the tehsil level (which is the unit of randomization) and presented in parenthesis. All regressions include district FE and year/month dummies. The observations are based on a AD-level monthly panel. Regressions are based on ITT. Treatment effects are robust to the inclusion of controls (Table A.8). We compute FDR adjusted q values for the multiple tests across treatment arms (adjusted q values for columns 2, 4, 6 and 8: Leaves approved - treatments - 0.001, 0.31, 0.05 0.08; tests across treatments - 0.004, 0.046, 0.455, 0.004; adjusted q- values for FA FTP: treatments - 0.862, 0.296, 0.881, 0.051; tests across treatments - 0.296, 0.051, 0.296, 0.051; adjusted q- values for AO FTP: treatments - 0.235, 0.699, 0.125, 0.016; tests across treatments - 0.159, 0.125, 0.019, 0.016; adjusted q- values for FTP scheduled: treatments - 0.672, 0.683, 0.148, 0.036; tests across treatments - 0.498, 0.112, 0.001, 0.627). Estimates are significant at the \*10%, \*\*5%, and \*\*\*1% level

Table 5: Distribution of Bonus Across Treatments

	(1)	(2)	(3)	(4)	(5)
	Base Bonus	Low Bonus	High Bonus	No Payment	Not Assigned
Objective	0.625	0.026	0.041	0.309	
Subjective	0.372	0.341	0.204	0.083	0.20
Subjective Plus	0.339	0.368	0.208	0.085	0.19
Observations	34624	34624	34624	34624	26145

Notes: The distribution of bonuses is calculated over all months starting from May 2018 to March 2020. The observations are fewer than Table 1 because: a) we do not have payment assignment data for the first month of roll-out; and b) these observations do not include controls.

Table 6: Treatment Effects on Bonus Assignment Behaviour

	(1)	(2)	(3)
	Average Bonus Amount	Variation in Bonus Amount	Prop. Staff Repeated for Bonus
Objective	2191.508*** (565.916)	1615.479*** (77.709)	0.301*** (0.101)
Subjective	4247.535*** (581.095)	1513.741*** (148.930)	0.901*** (0.101)
Subjective Plus	3997.316*** (578.916)	1749.751*** (101.129)	0.654*** (0.116)
<i>P-value differences</i>			
Obj - Sub	0.00***	0.45	0.00***
Obj-Subplus	0.00***	0.15	0.00***
Sub - Subplus	0.47	0.06*	0.01***
Observations	1929	1901	596
Adjusted R <sup>2</sup>	0.705	0.712	0.702
PDS LASSO controls	No	No	No
District Fixed Effects	Yes	Yes	No
Division Fixed Effects	No	No	Yes
Month dummies	No	No	Yes

Notes: Our estimation excludes the control group given staff in those tehsils was not assigned monthly payments in the AgriSmart system. The constant is not included in the estimation and hence the coefficients represent averages within treatment groups. In columns 1 and 2, we include district dummies to control for district effects and cluster errors at the tehsil level. The first column reports the average bonus amount assigned to staff on a monthly basis and the second column reports the variation in the bonus assigned to staff overtime. The data set for the first two estimations is collapsed at the employee level. The third column reports the proportion of staff repeated each month by ADs for a high bonus from the initial set of high bonus recipients in the first three months. The data set for this estimation is collapsed at the month-tehsil level. Regressions are based on ITT. Estimates are significant at the \*10%, \*\*5%, and \*\*\*1% level

Table 7: Treatment Effects on Intensive and Extensive Margins of Outreach Effort

	(1)	(2)	(3)	(4)
	Total reported numbers	Total unique numbers	Total unique from pre-treat	Average times number repeated
Objective	130.717*** (45.455)	78.481** (32.582)	115.326*** (37.928)	0.345*** (0.081)
Subjective	103.431** (46.314)	61.495* (32.081)	74.596* (38.016)	0.218*** (0.064)
Subjective Plus	171.172*** (40.960)	105.715*** (28.353)	152.163*** (33.609)	0.187*** (0.071)
<i>Mean of dep. var</i>	531.340	370.776	393.972	2.177
<i>P-value differences</i>				
Obj - Sub	0.63	0.67	0.39	0.07*
Obj-Subplus	0.44	0.46	0.41	0.05**
Sub - Subplus	0.17	0.18	0.06*	0.88
Observations	2179	2179	2169	2179
Adjusted R <sup>2</sup>	0.015	0.011	0.016	0.012
PDS LASSO controls	No	No	No	No
District Fixed Effects	Yes	Yes	Yes	Yes

Notes: Standard errors clustered at the tehsil level (the unit of randomization) are within parenthesis. The observations are collapsed at the employee level. Regressions are based on ITT. Estimates are significant at the \*10%, \*\*5%, and \*\*\*1% level

Table 8: Treatment Effects on Staff Satisfaction

	(1)	(2)
	Has Agrismart improved Extension (=1 if yes)	Satisfied with AgriSmart (normalised)
Objective	0.049* (0.027)	-0.323*** (0.075)
Subjective	-0.017 (0.029)	-0.087 (0.074)
Subjective Plus	0.000 (0.026)	-0.105 (0.067)
<i>P-value differences</i>		
Obj - Sub	0.04**	0.00***
Obj-Subplus	0.10*	0.00***
Sub - Subplus	0.57	0.81
Observations	4317	4317
Adjusted R <sup>2</sup>	0.013	0.016
PDS LASSO controls	No	No
District Fixed Effects	Yes	Yes

Notes: Standard errors clustered at the tehsil level (the unit of randomization) are within parenthesis. The data set is a monthly employee level panel comprised of a randomly selected subset of employees every month. Regressions are based on ITT. Estimates are significant at the \*10%, \*\*5%, and \*\*\*1% level

## References

- Abadie, Alberto, Susan Athey, Guido W Imbens, and Jeffrey Wooldridge. 2017. *When Should You Adjust Standard Errors for Clustering?* [electronic resource]. Working paper series (National Bureau of Economic Research : Online) ; working paper no.24003. Cambridge, Mass.: National Bureau of Economic Research.
- Anderson, Jock and Gershon Feder. 2007. "Chapter 44 Agricultural Extension." *Handbook of Agricultural Economics* 3:2343–2378.
- Anderson, Michael L. 2008. "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects." *Journal of the American Statistical Association* 103 (484):1481–1495. URL <http://www.tandfonline.com/doi/abs/10.1198/016214508000000841>.
- Andreoni, James, Michael Callen, Yasir Khan, and Charles Sprenger. 2016. "Using Preference Estimates to Customize Incentives: An Application to Polio Vaccination Drives in Pakistan." *NBER Working Paper Series* :22019URL <http://search.proquest.com/docview/1768594362/>.
- Baker, George et al. 1994. "Subjective performance measures in optimal incentive contrac." *The Quarterly Journal of Economics* 109 (4):1125. URL <http://search.proquest.com/docview/210978266/>.
- Bandiera, Oriana et al. 2007. "Incentives for Managers and Inequality Among Workers: Evidence From a Firm-Level Experiment\*." *The Quarterly Journal of Economics* 122 (2):729. URL <http://search.proquest.com/docview/210985392/>.
- Banerjee, Abhijit et al. 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 <http://search.proquest.com/docview/37031156/>.
- Beaman, Lori, Ariel Benyishay, Jeremy Magruder, and Ahmed Mobarak. 2018. "Can Network Theory-based Targeting Increase Technology Adoption?" *NBER Working Paper Series* :24912URL <http://search.proquest.com/docview/2089872947/>.
- Burgess, Simon and Marisa Ratto. 2003. "The Role of Incentives in the Public Sector Issues and Evidence." *Oxford Review of Economic Policy* 19 (2):285–300.
- Callen, Michael et al. 2016. "The Political Economy of Public Sector Absence: Experimental Evidence from Pakistan." *NBER Working Paper Series* :22340URL <http://search.proquest.com/docview/1795921414/>.

- Conley, Timothy G and Christopher R Udry. 2010. "Learning about a New Technology: Pineapple in Ghana." *American Economic Review* 100 (1):35–69.
- Department, Agriculture. 2018. "Punjab Agriculture Policy." URL <http://www.agripunjab.gov.pk/system/files/Punjab%20Agriculture%20Policy.pdf>.
- Duflo, Esther et al. 2011. "Nudging Farmers to Use Fertilizer: Theory and Experimental Evidence from Kenya." *American Economic Review* 101 (6):2350–2390.
- . 2012. "Incentives Work: Getting Teachers to Come to School." *American Economic Review* 102 (4):1241–1278.
- Fernando, A. Shawn, A. Nilesh; Cole. 2016. "Mobile'izing Agricultural Advice: Technology Adoption, Diffusion and Sustainability."
- Gibbs, Michael et al. 2004. "Performance Measure Properties and Incentives." *IDEAS Working Paper Series from RePEc* URL <http://search.proquest.com/docview/1698641182/>.
- Gine, Xavier et al. 2017. "Mission and the Bottom Line : Performance Incentives in a Multi-Goal Organization."
- Holmstrom, Bengt. 1980. "On The Theory of Delegation." Discussion Papers 438, Northwestern University, Center for Mathematical Studies in Economics and Management Science. URL <https://EconPapers.repec.org/RePEc:nwu:cmsems:438>.
- Holmstrom, Bengt and Paul Milgrom. 1991. "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design." *Journal of Law, Economics, Organization* 7:24–52.
- John, Oliver et al. 2008. *Paradigm shift to the integrative big five trait taxonomy: History, measurement, and conceptual issues*, vol. Vol. 3. 114–158.
- Khan, Adnan Q. et al. 2016. "Tax farming redux: experimental evidence on performance pay for tax collectors.(Report)." 131 (1):219.
- Manning, Nick, Zahid Hasnain, and Jan Henryk Pierskalla. 2012. *Public Sector Human Resource Practices to Drive Performance*. GET Note. World Bank, Washington, DC.
- Neal, Derek. 2011. "The Design of Performance Pay in Education." *NBER Working Paper Series* :16710URL <http://search.proquest.com/docview/1687829446/>.

- Parent, Daniel and James Rebitzer. 1999. "Job characteristics, wages, and the employment contract / Commentary." *Review - Federal Reserve Bank of St. Louis* 81 (3):13–34. URL <http://search.proquest.com/docview/227775076/>.
- Perry, James. 1996. "Measuring public service motivation: an assessment of construct reliability and validity." *Journal of public administration research and theory* 6 (1):5–24. URL <http://search.proquest.com/docview/39010044/>.
- Prendergast, Canice. 1999. "The Provision of Incentives in Firms." *Journal of Economic Literature* 37 (1):7–63.
- . 2007. "The Motivation and Bias of Bureaucrats." *American Economic Review* 97 (1):180–196.
- Prendergast, Canice and Robert H. Topel. 1996. "Favoritism in Organizations." *Journal of Political Economy* 104 (5):958–978.
- Rajan, Madhav and Stefan Reichelstein. 2009. "Objective versus subjective indicators of managerial performance.(Report)." *Accounting Review* 84 (1):209.
- 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.
- Suri, Tavneet. 2011. "Selection and Comparative Advantage in Technology Adoption.(Report)." *Econometrica* 79 (1):159.
- Survey, Pakistan Economic. 2019. "Pakistan Economic Survey." URL [http://www.finance.gov.pk/survey/chapters\\_19/2-Agriculture.pdf](http://www.finance.gov.pk/survey/chapters_19/2-Agriculture.pdf).

# Appendix

## Appendix A: Tables

A. 1: Descriptive Statistics

	(1)	(2)	(3)	(4)	(5)
	Mean	Sd	p0.25	p0.50	p0.75
<b>Basic FA/AO characteristics</b>					
Age	44.36	11.76	34.33	47.75	54.21
Salary	39448	27700	27350	37200	43419
Years of experience	20.27	12.28	10.13	24.71	31.44
<b>FA/AO Performance</b>					
Days Worked/month	13.33	6.86	9	14	19
Hours Worked/month	7.55	4.64	4.14	7.40	11.00
Village Visits/month	13.32	11.89	4	11	19
Farmer Visited/month	17.71	23.16	3	11	24
Small Farmer Visited/month	9.40	12.70	1	5	13
Prop. Time on Extension/month	0.55	0.30	0.34	0.59	0.79
Distance Travelled (in Kms)/month	368.44	1174	108	226	412
<b>Other Variables</b>					
TA Pay Gap (yearly)	16341	24649	4211	10587	21383
Avg Claim (yearly)	26750	40656	0	10475	31353
Avg Payment (yearly)	18463	32172	0	6300	18000
Prop. of Villages with Metal Road	0.83	0.11	0.76	0.84	0.91
<b>Agrismart Uptake</b>					
Active Users	90.41	3.21	89.18	91.26	92.70
Observations	2595				

## A. 2: Randomization Balance - All Treatments

	Mean				P-value Differences					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Control	Objective	Subjective	Subjective +	C - obj	C-sub	C-sub +	obj-sub	obj-sub+	Sub- Sub+
<b>FA/AO characteristics</b>										
Age	43.48 (0.89)	44.94 (0.53)	45.06 (0.54)	43.59 (0.69)	0.166	0.132	0.92	0.88	0.13	0.10
Salary	38865 (1721)	40829 (1198)	38727 (902)	39869 (1475)	0.35	0.94	0.66	0.18	0.62	0.51
Years in Service	18.89 (0.83)	20.63 (0.56)	21.14 (0.55)	19.83 (0.68)	0.08	0.02**	0.39	0.51	0.38	0.14
Days Worked/month	14.17 (0.76)	12.81 (0.84)	12.88 (0.88)	13.52 (0.98)	0.23	0.27	0.60	0.96	0.58	0.63
Hours Worked/month	7.62 (0.76)	7.60 (0.84)	7.20 (0.88)	7.92 (0.98)	0.96	0.52	0.65	0.55	0.62	0.33
Village Visits/month	13.16 (1.14)	13.03 (1.07)	13.43 (1.38)	14.33 (1.59)	0.94	0.88	0.54	0.82	0.50	0.67
Farmer Visits/month	17.80 (1.86)	16.97 (1.64)	17.14 (2.01)	19.97 (2.80)	0.74	0.81	0.51	0.95	0.36	0.41
Small Farmer Visits/month	9.57 (1.13)	10.00 (1.02)	8.50 (1.10)	9.86 (1.29)	0.78	0.50	0.86	0.32	0.93	0.42
Prop. Time on Extension/month	0.57 (0.02)	0.57 (0.02)	0.53 (0.04)	0.54 (0.02)	0.85	0.48	0.37	0.37	0.20	0.92
Distance Travelled (in Kms)/month	380.10 (76.30)	345.10 (57.20)	351.50 (61.08)	411.08 (91.12)	0.72	0.77	0.79	0.94	0.54	0.59
Avg Claim (yearly)	60014 (7888)	59357 (5444)	69976 (12825)	60414 (4937)	0.95	0.50	0.97	0.44	0.88	0.49
Avg Payment (yearly)	42949 (5477)	45620 (4486)	52670 (11159)	43031 (3040)	0.71	0.43	0.99	0.55	0.63	0.41
Joint F-Test					0.27	0.09	0.15	0.12	0.54	0.67
Observations	610	682	675	628						
<b>Tehsil Variables</b>										
Prop. of Villages with Metal Road	0.79 (0.02)	0.84 (0.02)	0.84 (0.02)	0.81 (0.02)	0.16	0.20	0.61	0.85	0.25	0.33
Prop. of Villages with Brick Street	0.36 (0.05)	0.40 (0.06)	0.52 (0.05)	0.44 (0.05)	0.07*	0.03**	0.28	0.83	0.39	0.26
Prop. of Villages with Canal	0.61 (0.07)	0.64 (0.07)	0.67 (0.06)	0.69 (0.06)	0.75	0.53	0.38	0.78	0.61	0.79
Joint F-Test					0.50	0.24	0.51	0.99	0.19	0.20
Observations	32	32	34	33						

Notes: The first 4 columns represent the mean and standard deviation in parenthesis of the different variables across treatment groups. The next 6 columns report the results for equality of means between each of the treatment arms. We also add report the p-value difference of the joint F-test of individual characteristics between each of the groups. Estimates are significant at the \*\*5%, and \*\*\*1% level.

### A. 3: Balance for ITT Employees

	Mean				P-value Differences					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Control	Objective	Subjective	Subjective +	C - obj	C-sub	C-sub +	obj-sub	obj-sub+	Sub- Sub+
<b>FA/AO characteristics</b>										
Age	42.95 (1.45)	43.88 (1.06)	43.79 (0.99)	43.48 (1.25)	0.35	0.44	0.72	0.90	0.61	0.72
Salary	38118 (2062)	39913 (1496)	37835 (1637)	39752 (1461)	0.32	0.87	0.18	0.08*	0.91	0.18
Years in Service	18.30 (1.39)	19.61 (1.06)	20.02 (0.98)	19.41 (1.19)	0.19	0.12	0.49	0.63	0.80	0.49
Days Worked/month	15.50 (0.94)	15.28 (0.87)	14.91 (0.86)	15.36 (0.84)	0.82	0.59	0.67	0.71	0.92	0.67
Hours Worked/month	7.74 (0.51)	8.52 (0.47)	7.84 (0.48)	8.45 (0.47)	0.12	0.87	0.30	0.21	0.91	0.30
Village Visits/month	12.17 (1.21)	14.25 (1.12)	13.36 (1.36)	14.23 (1.15)	0.10*	0.44	0.59	0.53	0.99	0.59
Farmer Visits/month	16.82 (2.51)	18.16 (2.10)	16.52 (2.41)	19.16 (2.56)	0.51	0.91	0.42	0.49	0.71	0.42
Small Farmer Visits/month	7.09 (1.55)	9.23 (1.31)	6.86 (1.31)	8.13 (1.56)	0.09*	0.88	0.44	0.09*	0.44	0.44
Prop. Time on Extension/month	0.52 (0.04)	0.56 (0.04)	0.50 (0.04)	0.50 (0.04)	0.26	0.54	0.86	0.11	0.08*	0.86
Distance Travelled (in Kms)/month	515.59 (110)	526.28 (95)	505.54 (119)	569.69 (100)	0.91	0.92	0.57	0.78	0.67	0.57
Avg Claim (yearly)	67846 (8793)	76492 (9576)	67924 (14074)	66276 (9555)	0.04**	0.99	0.84	0.25	0.06*	0.84
Avg Payment (yearly)	53191 (7897)	58700 (8728)	52269 (13444)	49564 (8230)	0.09*	0.89	0.70	0.35	0.02**	0.70
Joint F-Test					0.27	0.12	0.43	0.18	0.35	0.80
Observations	12246	13852	13986	12678						

Notes: The first 4 columns represent the mean and standard deviation in parenthesis of the different variables across treatment groups. The next 6 columns report the results for equality of means between each of the treatment arms. We also add report the p-value difference of the joint F-test of individual characteristics between each of the groups. Estimates are significant at the \*\*5%, and \*\*\*1% level.

#### A. 4: Attrited Sample and Treatments

---

	(1)
	Attrited(=1 if sample attrited)
Objective	-0.013 (0.029)
Subjective	-0.033 (0.030)
Subjective Plus	0.009 (0.035)
Observations	59708
R-squared	0.027
Controls	No
District FE	Yes

---

Notes: Errors are clustered at the tehsil level which is the unit of randomization. Estimates are significant at the \*10%, \*\*5%, and \*\*\*1% level.

### A. 5: Attrition and its Correlates

	(1)	(2)	(3)	(4)	(5)	(6)
	Control+ Obj	Control + Sub	Control + Subplus	Obj + Sub	Obj + Subplus	Sub + Subplus
Objective	-0.010 (0.028)			0.004 (0.030)		
Subjective		-0.008 (0.031)				-0.021 (0.024)
Subjective Plus			0.010 (0.032)		0.021 (0.029)	
Age	0.000 (0.001)	-0.000 (0.001)	-0.000 (0.001)	0.002 (0.001)	-0.003* (0.002)	-0.001 (0.001)
Salary	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
Yrs of Exp	-0.003* (0.001)	-0.003* (0.001)	-0.003* (0.002)	-0.002* (0.001)	-0.001 (0.002)	-0.003** (0.001)
Agrismart Index	-0.003 (0.010)	0.008 (0.014)	-0.001 (0.016)	0.005 (0.013)	-0.030** (0.013)	-0.018 (0.014)
Yearly TADA Claimed	-0.000 (0.000)	-0.000 (0.000)	-0.000 (0.000)	0.000 (0.000)	-0.000 (0.000)	-0.000 (0.000)
Yearly TADA Paid	-0.000* (0.000)	-0.000 (0.000)	-0.000 (0.000)	-0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
Age* Treat	-0.003* (0.002)	0.001 (0.002)	-0.001 (0.002)	-0.005** (0.002)	0.002 (0.002)	0.002 (0.002)
Salary * Treat	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	-0.000 (0.000)	-0.000 (0.000)
Yrs Exp * Treat	0.002 (0.002)	0.001 (0.002)	-0.000 (0.002)	0.002 (0.002)	-0.002 (0.002)	0.001 (0.002)
Agrismart Index * Treat	-0.016 (0.017)	-0.004 (0.020)	-0.020 (0.027)	-0.024 (0.019)	0.020 (0.019)	0.019 (0.021)
Yearly TADA Claimed * Treat	0.000 (0.000)	-0.000 (0.000)	-0.000 (0.000)	0.000 (0.000)	-0.000 (0.000)	0.000 (0.000)
Yearly TADA Paid * Treat	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
Joint Test	0.339	0.782	0.805	0.399	0.239	0.977
Observations	12987	12495	12665	13246	13416	12924
Adjusted R <sup>2</sup>	0.039	0.021	0.038	0.031	0.052	0.029
Controls	No	No	No	No	No	No
District Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Standard errors in parentheses. The joint test includes a test including the treatment dummy and all other treatment interaction variables. Estimates are significant at the \*\*5%, and \*\*\*1% level.

A. 6: Treatment Effects on Extension Outreach - Agrismart Measures (with controls)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Days Worked	Hours Worked	Village visits	Individual Farmer Visits	FTP Scheduled	FTP Completed	Performance Index
Objective	0.947*** (0.298)	0.773*** (0.260)	8.499*** (1.698)	18.350*** (3.385)	-0.265 (0.654)	0.224 (0.523)	0.402*** (0.070)
Subjective	0.818*** (0.250)	0.483** (0.224)	3.316*** (1.112)	6.639** (2.940)	-0.603 (0.507)	0.518 (0.461)	0.191*** (0.061)
Subjective Plus	0.588** (0.274)	0.885*** (0.236)	4.142*** (1.224)	8.443*** (2.411)	1.396** (0.614)	1.350*** (0.443)	0.237*** (0.051)
Observations	46212	46212	46212	46212	18022	18022	46212
Adjusted R <sup>2</sup>	0.405	0.239	0.260	0.163	0.119	0.129	0.200
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
District Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Standard errors clustered at the tehsil level (the unit of randomization) are within parenthesis. All regressions are an ANCOVA estimation with district FE and year-month dummies. Controls at the staff level include: age, salary, pre-treatment TADA claims, and TADA payments. Controls at the tehsil level include: access of villages to a metal road, proportion of land cultivated, and proportion of villages with access to brick streets. Regressions are based on ITT. Estimates are significant at the \*10%, \*\*5%, and \*\*\*1% level

### A. 7: Quarter by Quarter Effects on Agrismart Outcomes

	(1)	(2)	(3)	(4)	(5)	(6)
	Days Worked	Hours Worked	Village visits	Individual Farmer Visits	FTP Completed	Performance Index
Q1 * Obj	0.081 (0.334)	1.056*** (0.391)	1.461 (1.715)	2.669 (3.721)	-1.192 (1.406)	0.118* (0.069)
Q1 * Sub	0.662** (0.286)	0.249 (0.348)	0.316 (1.204)	-0.832 (3.195)	-3.456*** (0.759)	0.029 (0.056)
Q1 * Subplus	0.185 (0.391)	0.628* (0.349)	0.750 (1.341)	2.270 (3.120)	-2.934*** (0.990)	0.056 (0.066)
Q2 * Obj	0.956** (0.369)	1.318*** (0.317)	6.025*** (1.876)	13.123*** (3.756)	-0.055 (1.030)	0.346*** (0.080)
Q2 * Sub	0.965*** (0.355)	0.716** (0.278)	3.628** (1.398)	7.558** (3.125)	-1.086 (1.133)	0.200*** (0.067)
Q2 * Subplus	0.557 (0.472)	0.981*** (0.296)	3.378** (1.552)	9.054** (3.519)	0.160 (1.125)	0.204*** (0.078)
Q3 * Obj	0.909** (0.353)	0.586** (0.288)	7.344*** (1.872)	14.932*** (4.074)	0.701 (0.754)	0.336*** (0.085)
Q3 * Sub	0.897** (0.379)	0.417* (0.252)	4.251*** (1.517)	6.003* (3.357)	0.776 (0.769)	0.173** (0.069)
Q3 * Subplus	1.126*** (0.391)	0.671** (0.299)	5.004*** (1.349)	8.506*** (3.005)	1.892** (0.833)	0.229*** (0.065)
Q4 * Obj	0.729** (0.343)	0.813*** (0.294)	9.076*** (1.942)	21.052*** (3.742)	1.207 (1.078)	0.423*** (0.080)
Q4 * Sub	0.607 (0.377)	0.620*** (0.228)	4.248*** (1.415)	7.587** (3.364)	0.616 (0.768)	0.186*** (0.068)
Q4 * Subplus	0.435 (0.470)	1.159*** (0.289)	2.953* (1.594)	6.137** (2.930)	1.409 (0.853)	0.172** (0.073)

Quarter by Quarter Effects: Continued

	(1)	(2)	(3)	(4)	(5)	(6)
	Days Worked	Hours Worked	Village visits	Individual Farmer Visits	FTP Completed	Performance Index
Q5 * Obj	1.039*** (0.317)	1.152*** (0.317)	9.186*** (2.141)	19.041*** (3.767)	0.619 (0.794)	0.440*** (0.088)
Q5 * Sub	0.627* (0.326)	0.634*** (0.238)	3.238** (1.275)	6.079* (3.275)	0.908 (0.682)	0.157** (0.064)
Q5 * Subplus	0.669** (0.280)	1.321*** (0.262)	3.612** (1.671)	6.885** (3.123)	1.617** (0.690)	0.211*** (0.068)
Q6 * Obj	0.988*** (0.323)	1.092*** (0.333)	8.708*** (2.259)	17.602*** (5.293)	0.719 (0.960)	0.415*** (0.098)
Q6 * Sub	0.592* (0.330)	0.564** (0.282)	2.748** (1.312)	8.449** (3.861)	0.577 (1.008)	0.165** (0.070)
Q6 * Subplus	0.763** (0.385)	1.189*** (0.320)	5.681*** (1.775)	11.251*** (4.006)	2.479** (1.147)	0.283*** (0.076)
Q7 * Obj	1.364*** (0.370)	0.851*** (0.303)	15.191*** (2.896)	29.032*** (5.128)	0.587 (0.765)	0.628*** (0.109)
Q7 * Sub	0.764** (0.380)	0.662** (0.273)	3.920** (1.552)	7.989** (3.535)	0.581 (0.831)	0.197*** (0.072)
Q7 * Subplus	1.254*** (0.342)	1.140*** (0.300)	8.772*** (1.987)	15.702*** (3.705)	2.080*** (0.790)	0.394*** (0.076)
Q8 * Obj	1.541*** (0.478)	1.101*** (0.338)	17.570*** (3.835)	32.521*** (6.628)	1.187 (0.985)	0.722*** (0.143)
Q8 * Sub	0.730 (0.520)	0.746** (0.315)	3.620** (1.775)	7.937* (4.131)	0.512 (0.861)	0.193** (0.087)
Q8 * Subplus	1.301*** (0.446)	1.345*** (0.327)	8.531*** (3.117)	14.981** (5.869)	0.815 (0.776)	0.396*** (0.115)
Observations	49157	49157	49157	49157	19315	49157
Adjusted R <sup>2</sup>	0.406	0.233	0.229	0.139	0.026	0.246
PDS LASSO controls	No	No	No	No	No	No
District Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Standard errors clustered at the tehsil level (the unit of randomization) are within parenthesis. All regressions are an ANCOVA estimation with district FE and year-month dummies. Regressions are based on ITT. All regressions include the quarter dummy variables. Estimates are significant at the \*10%, \*\*5%, and \*\*\*1% level

### A. 8: Balance on Characteristics of Farmers (in the farmer surveys)

	Mean				P-value Differences					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Control	Objective	Subjective	Subjective +	C - obj	C-sub	C-sub +	obj-sub	obj-sub+	Sub- Sub+
Number of Observations/Tehsil	84.11 (63.11)	84.92 (62.35)	73.99 (55.88)	81.14 (58.22)	0.96	0.57	0.86	0.53	0.81	0.64
Age	47.17 (0.94)	48.66 (0.96)	48.46 (1.00)	47.16 (0.98)	0.01***	0.01***	0.99	0.68	0.01***	0.02**
Gender (=1 if male)	1.000 (0.002)	1.000 (0.002)	0.999 (0.002)	0.996 (0.001)	0.26	0.84	0.18	0.12	0.08*	0.30
Less than Matric (=1 if Yes)	0.62 (0.05)	0.59 (0.05)	0.61 (0.05)	0.62 (0.05)	0.28	0.96	0.89	0.19	0.10	0.84
Land Owned (In Acres)	11.94 (2.01)	11.10 (2.00)	10.04 (2.08)	11.54 (1.94)	0.48	0.07*	0.76	0.27	0.72	0.12
Land Rented (In Acres)	2.66 (1.14)	2.52 (1.15)	2.55 (1.18)	2.43 (1.09)	0.71	0.74	0.54	0.94	0.78	0.69
Water Quality (=1 if sweet)	0.55 (0.04)	0.55 ((0.04)	0.58 (0.04)	0.60 (0.04)	0.91	0.34	0.09*	0.30	0.07*	0.33
Joint F-Test					0.48	0.10	0.88	0.58	0.33	0.30
Observations	1946	2253	2234	2031						

Notes: The first 4 columns represent the mean and standard deviation in parenthesis of the different variables across treatment groups. The next 6 columns report the results for equality of means between each of the treatment arms. We also add report the p-value difference of the joint F-test of individual characteristics between each of the groups. Estimates are significant at the \*\*5%, and \*\*\*1% level.

A. 9: Treatment Effects on Extension Access and Quality - Farmer Surveys (with controls)

	(1)	(2)	(3)	(4)
	Farmer Outreach Index	Meeting Useful	Length Meeting	Knowledge Score
Objective	-0.036 (0.030)	-0.002 (0.040)	-4.619 (5.945)	0.128 (0.087)
Subjective	0.018 (0.028)	0.046 (0.037)	-4.819 (5.616)	0.017 (0.085)
subjective Plus	0.094*** (0.032)	0.140*** (0.038)	-2.200 (5.444)	0.068 (0.086)
Observations	6552	1020	1020	3301
Adjusted R <sup>2</sup>	0.016	0.009	0.017	0.038
PDS LASSO controls	Yes	Yes	Yes	Yes
District Fixed Effects	Yes	Yes	Yes	Yes

Notes: Standard errors clustered at the tehsil level (the unit of randomization) are within parenthesis. All regressions include district FE and year-month dummies. Regressions are based on ITT. Controls include farmer age, land size, education, and gender. Estimates are significant at the \*10%, \*\*5%, and \*\*\*1% level

A. 10: Treatment Effects on Farmer Practices and Outcomes (with controls)

	(1)	(2)	(3)	(4)	(5)
	Planted New Crop	Increased area of crop	Decreased area of crop	Changed Practice	Wheat Yield (in kg)
Objective	-0.003 (0.011)	-0.035** (0.017)	-0.031* (0.017)	0.016 (0.016)	4.641 (36.986)
Subjective	-0.014 (0.013)	-0.037** (0.017)	-0.032* (0.018)	-0.003 (0.017)	21.289 (30.554)
Subjective Plus	0.015 (0.011)	-0.009 (0.017)	0.004 (0.018)	0.035** (0.016)	76.084* (42.432)
Observations	6367	6367	6367	6367	5807
Adjusted R <sup>2</sup>	0.003	0.010	0.010	0.015	0.006
Controls	Yes	Yes	Yes	Yes	Yes
District Fixed Effects	Yes	Yes	Yes	Yes	Yes

Notes: Standard errors clustered at the tehsil level (the unit of randomization) are within parenthesis. All regressions include district FE and year-month dummies. Regressions are based on ITT. Controls include farmer age, land size, education, and gender. Estimates are significant at the \*10%, \*\*5%, and \*\*\*1% level

A. 11: Treatment Effects on AD Performance (with controls)

	% Leaves Approved		FAs FTP Assigned		AOs FTP Assigned		Total Scheduled FTPs	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Objective	0.170*** (0.041)	0.169*** (0.041)	0.005 (0.030)	0.005 (0.030)	-0.002 (0.033)	0.001 (0.034)	0.404 (0.723)	0.405 (0.723)
Subjective	0.043 (0.037)	0.042 (0.036)	0.040 (0.029)	0.041 (0.029)	-0.009 (0.040)	-0.007 (0.040)	-0.199 (0.545)	-0.200 (0.545)
Subjective Plus	-0.032 (0.034)		0.029 (0.029)		0.011 (0.034)		1.481** (0.627)	
Before Tehsil Report x Sub Plus		-0.136*** (0.048)		-0.012 (0.032)		-0.090** (0.042)		1.253 (0.767)
After Tehsil Report x Sub Plus		0.046 (0.040)		0.059* (0.033)		0.097** (0.040)		1.632** (0.636)
Observations	12691	12691	36586	36586	3683	3683	18463	18463
Adjusted R <sup>2</sup>	0.078	0.088	0.079	0.079	0.171	0.179	0.130	0.130
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
District Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Standard errors clustered at the tehsil level (the unit of randomization) are within parenthesis. All regressions include district FE and year-month dummies. Regressions are based on ITT. Controls include AD and FA/AO level characteristics. Estimates are significant at the \*10%, \*\*5%, and \*\*\*1% level

A. 12: Treatment Effects on Audit Data

	(1)	(2)	(3)	(4)	(5)
	Survey Hit Rate	Know Field Staff	Attend Activity	Group Activity Useful	Ind. Activity Useful
Objective	-0.004 (0.011)	0.039* (0.021)	0.004 (0.022)	-0.013 (0.022)	-0.001 (0.009)
Subjective	0.002 (0.010)	0.022 (0.020)	-0.026 (0.020)	-0.025 (0.021)	-0.016* (0.008)
Subjective Plus	0.023** (0.011)	-0.012 (0.023)	-0.023 (0.023)	-0.035 (0.022)	-0.015 (0.010)
Observations	22042	14487	12961	2444	8177
Adjusted R <sup>2</sup>	0.120	0.002	0.007	0.015	0.019
PDS LASSO controls	No	No	No	No	No
District Fixed Effects	Yes	Yes	Yes	Yes	Yes

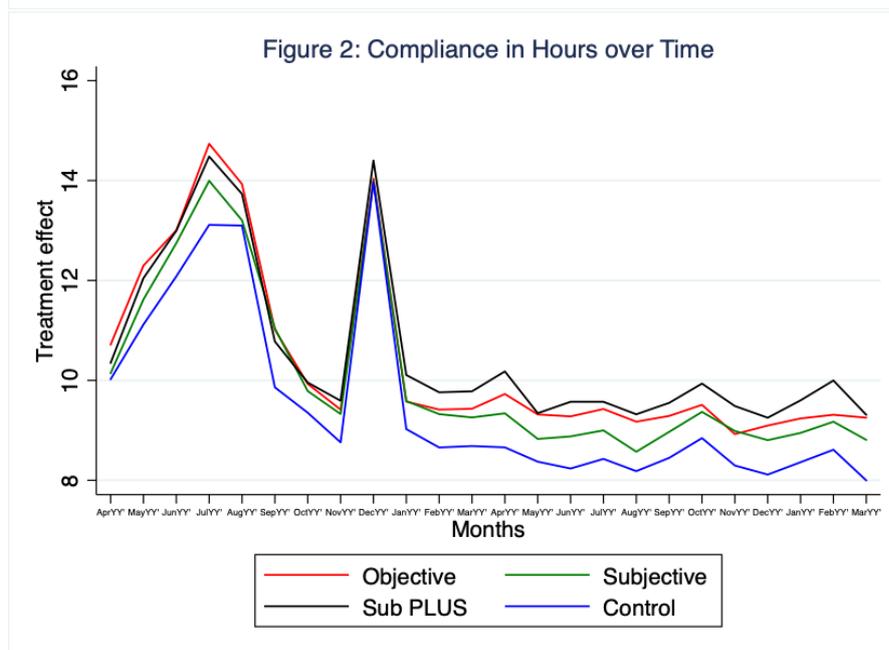
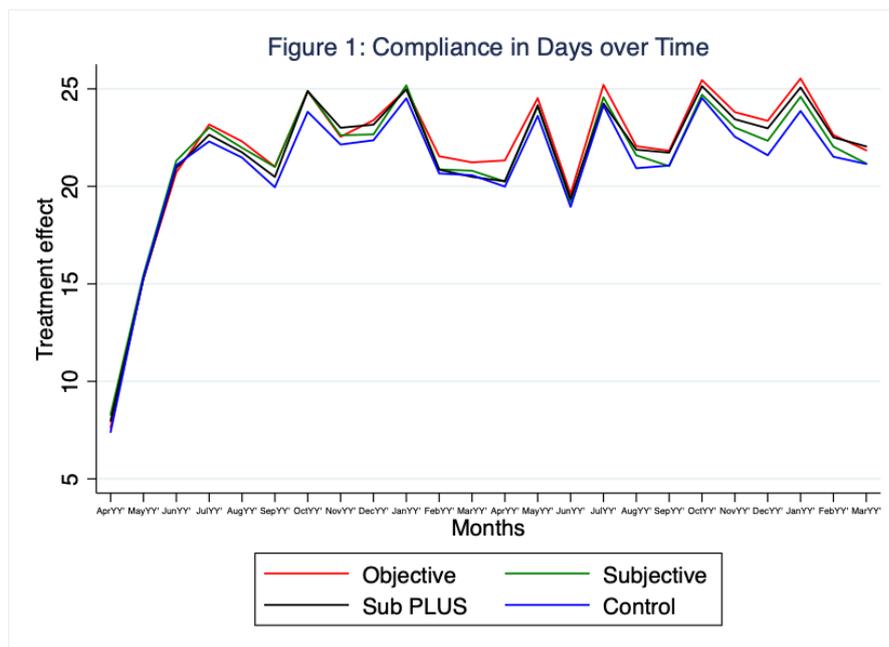
Notes: Standard errors clustered at the tehsil level (the unit of randomization) are within parenthesis. All regressions include district FE and year-month dummies. Regressions are based on ITT. Controls include AD and FA/AO level characteristics. Estimates are significant at the \*10%, \*\*5%, and \*\*\*1% level

### A. 13: Predictors of Bonus

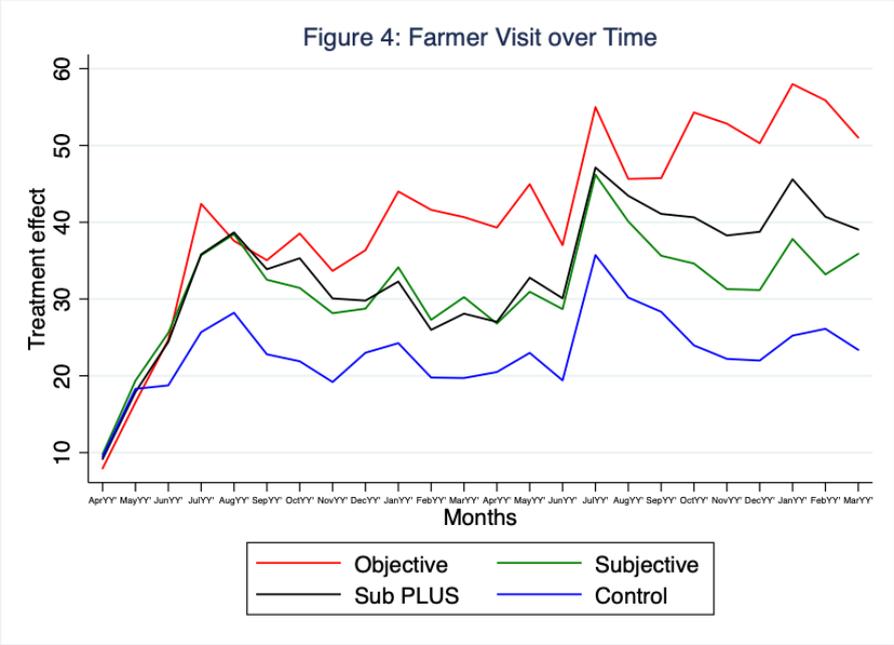
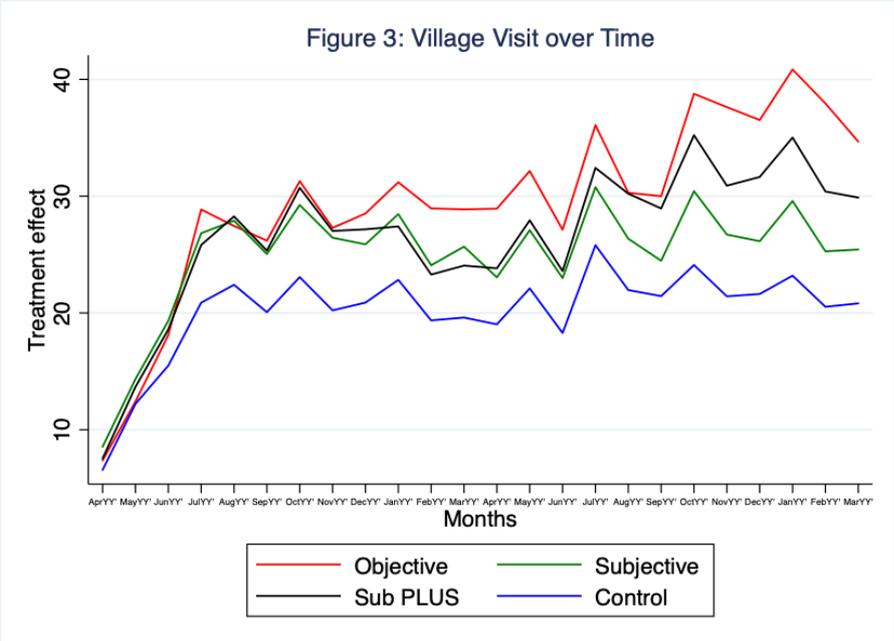
	(1)	(2)	(3)
	Objective	Subjective	Subjective Plus
Experience	-0.001 (0.001)	-0.001 (0.001)	-0.000 (0.001)
Salary	0.000 (0.000)	0.000 (0.000)	0.000** (0.000)
Percent Compliance (Days)	0.014 (0.009)	0.139*** (0.036)	0.119*** (0.041)
Percent Compliance (Hours)	0.013 (0.012)	-0.027 (0.032)	0.009 (0.024)
Above Med. Village Visits	0.063*** (0.017)	0.075*** (0.026)	0.104*** (0.023)
Above Med. Travel Spread	0.016* (0.008)	0.037*** (0.012)	0.049*** (0.012)
Above Med. Dept. Tasks	-0.051** (0.019)	-0.020 (0.021)	0.000 (0.020)
Above Med. Computer Saviness	0.010 (0.015)	0.014 (0.023)	0.011 (0.015)
Personal advice from AD	0.005 (0.009)	-0.025 (0.022)	0.004 (0.028)
Work advice from AD	-0.009 (0.026)	-0.034 (0.024)	0.007 (0.025)
Had meals with AD	0.024 (0.015)	0.041* (0.022)	0.051*** (0.015)
Knew AD before Job	0.073 (0.068)	-0.041 (0.061)	0.050** (0.023)
Observations	8818	7338	6826
Adjusted R <sup>2</sup>	0.044	0.056	0.066
PDS LASSO controls	No	No	No
District Fixed Effects	Yes	Yes	Yes

Notes: Standard errors clustered at the tehsil level are within parenthesis. The dependent variable is a dummy variable equal to 1 if staff was assigned a bonus. The first column reports key predictors of bonus in the objective arm, the second column reports them for the subjective arm, and the third column for the subjective plus arm. The data set is structured at the employee month level. Regressions are based on ITT. Estimates are significant at the \*10%, \*\*5%, and \*\*\*1% level

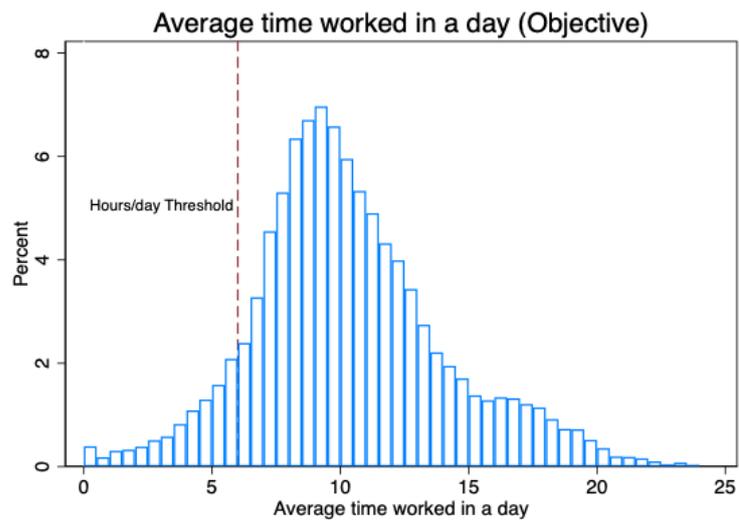
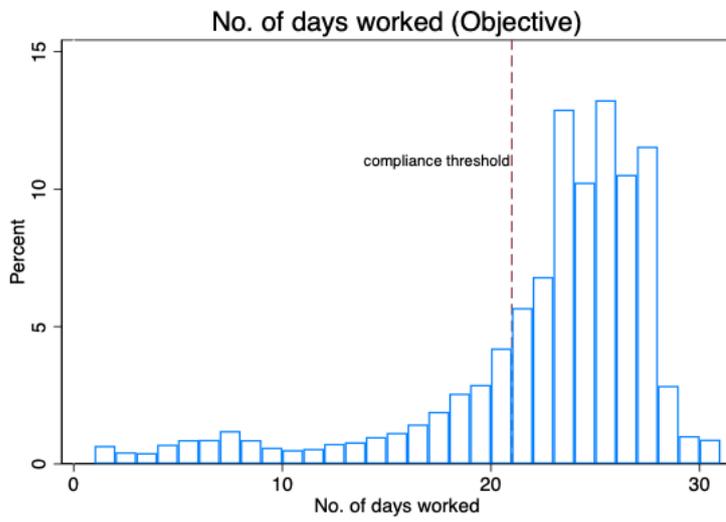
## Appendix B: Figures



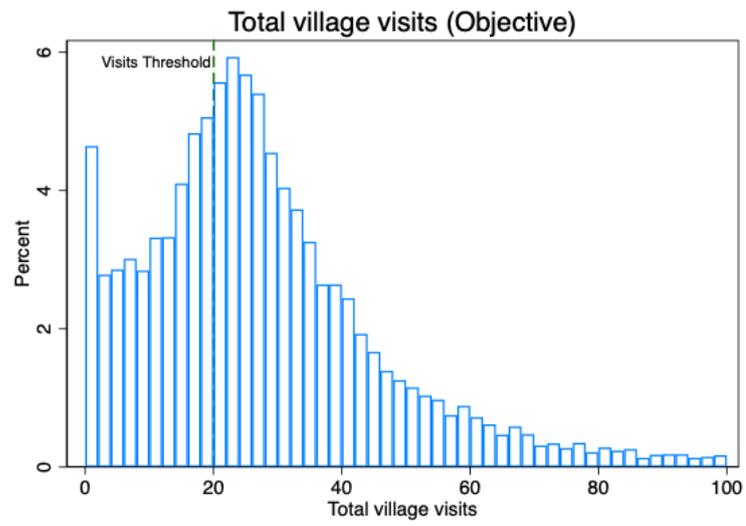
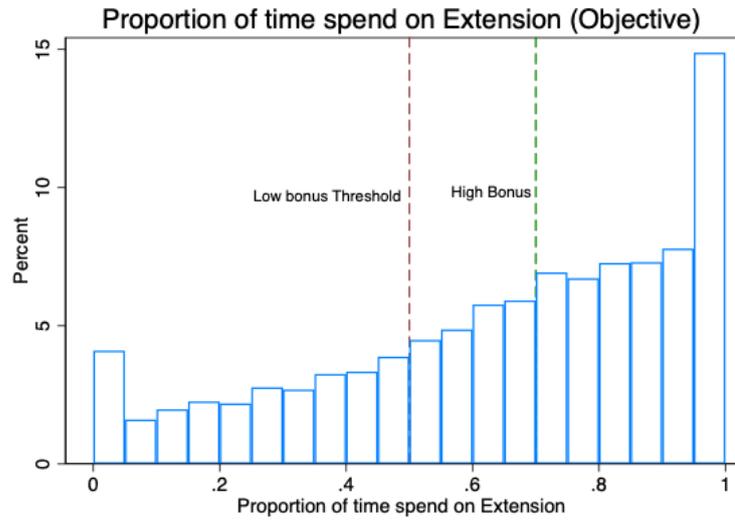
B. 1: Compliance over Time - Month by Month



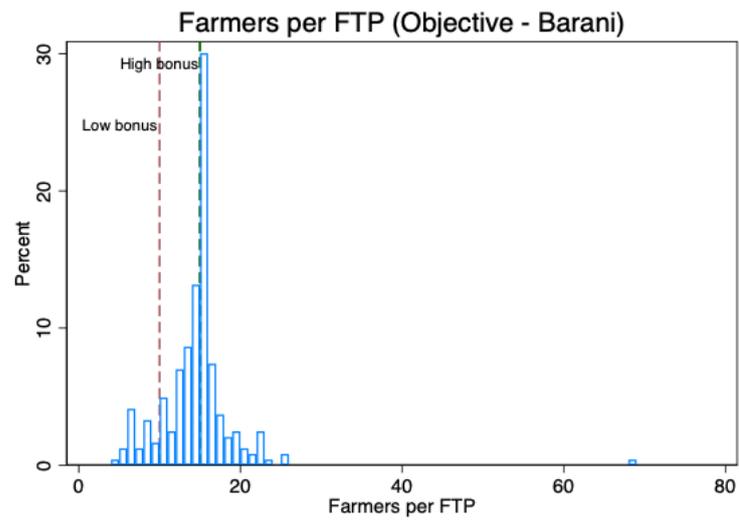
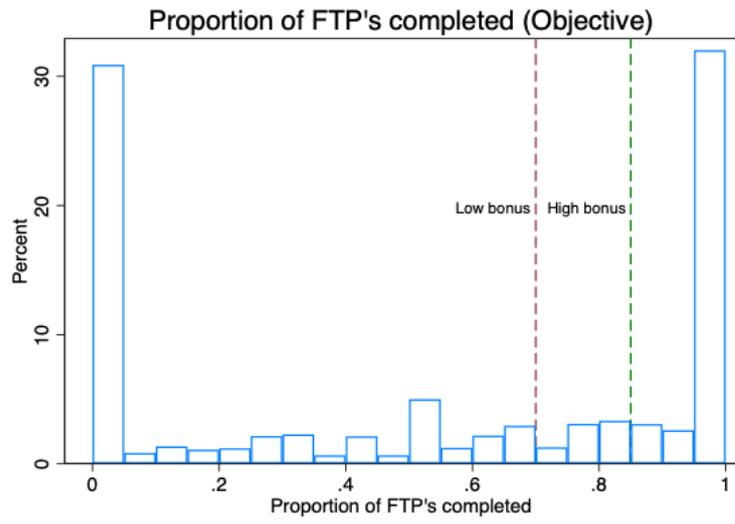
B. 2: Extension Visits over Time - Month by Month



B. 3: Bunching on Compliance Thresholds



B. 4: Bunching on Extension Tasks Thresholds



B. 5: Bunching on FTP-related Indicators

## Appendix C: Design Details

Designation	Payment Categories			
	No Payment	Base Payment*	Low Bonus	High Bonus
			(Base + Bonus)	(Base + Bonus)
FA	0	2500	4900 (2500 + <u>2400</u> )	7300 (2500 + <u>4800</u> )
AI	0	3000	6000 (3000 + <u>3000</u> )	9000 (3000 + <u>6000</u> )
AO	0	6000	10500 (6000 + <u>4500</u> )	15000 (6000 + <u>9000</u> )

Notes: All amounts are in PKR/month. The amounts mentioned here are the maximum possible amounts that can be received under Base Payment. The actual amount of Base Payment will be pro-rated based on the number of days worked in a month. The exact method of pro-rating is specified below.

### C. 1: BONUS PAYMENT CATEGORIES UNDER THE 3 INCENTIVE SCHEMES

<b>Indicator</b>	<b>No Payment</b>	<b>Base Payment*</b>	<b>Low Bonus</b>	<b>High Bonus</b>
<b>Number of days</b>	<22	22	22	22
<b>Hours per day</b>	<6	6	6	6
<b>Villages visited</b>			20	20
<b>% of time spent on Extension</b>			50%	65%
<b>% of scheduled FTP conducted</b>			70%	85%
<b>Farmers per FTP</b>			Irrigated areas: 20 Barani: 10	Irrigated areas: 25 Barani: 15
<b>Farmers visited</b>			48	72
<b>Small farmers visited</b>			24	36
<b>Distance travelled</b>			400	400

**Assumptions:**

**Farmers visited:** This assumes that out of a week of 6 days, an FA spends 1 day on FTP and 1 day on office activities. Out of the remaining 4 days, if soil sampling is being conducted on one of the days, it is not possible to visit more than 3 farmers. For the other 2 days, a bare minimum of 3 farmers must be visited. Based on this, a minimum threshold of 12 farmers per week is suggested which leads to around 48 farmers a month. For high bonus, the threshold increases by a factor of 1.5, i.e. 18 farmers are visited per week leading to the threshold of 72 farmers. Note that an FTP often takes a full work day because an FA may have to bring farmers to the training and drop them back after the training is completed.

**Village visited:** Out of a total of 5 working days (with 1 day reserved for potential office tasks), an FA is expected to visit a different village every day. To do his job well, an FA must spend a full day in a village trying to reach out to as many farmers as possible. This gives the threshold 20 villages/month. The village visited threshold for high bonus remains the same since higher than 20 villages per month can induce the FA in not maximizing his outreach within each village.

**Distance travelled:** To do his/her basic tasks, an FA needs to travel a minimum of 20 kms daily on average, which gives a threshold of 400 kms per month (assuming FA is in the field 5 days in a week). The threshold remains the same for high bonus since more distance travelled will not necessarily imply higher performance.

**C. 2: PAYMENT CATEGORY THRESHOLDS FOR FAS/AIS IN OBJECTIVE MEASURE SCHEME**

<b>Indicator</b>	<b>No Payment</b>	<b>Base Payment*</b>	<b>Low Bonus</b>	<b>High Bonus</b>
<b>Number of days</b>	<22	22	22	22
<b>Hours per day</b>	<6	6	6	6
<b>Villages visited</b>			20	20
<b>% of time spent on Extension</b>			50%	65%
<b>% of scheduled FTP conducted</b>			70%	85%
<b>Farmers per FTP</b>			Irrigated areas: 20 Barani: 10	Irrigated areas: 25 Barani: 15
<b>Farmers visited</b>			32	48
<b>Small farmers visited</b>			16	24
<b>Distance travelled</b>			640	640

**Assumptions:**

**Farmers visited:** This assumes that out of a week of 6 days, an AO spends at least 5 days doing FTPs, hence individual farmer outreach for AOs is lesser. It was proposed that a realistic low bonus threshold for AOs was around 8 farmers/week which leads to 32 farmers /month. The high bonus threshold is 1.5 times higher than this.

**Village visited:** Out of a total of 5 working days (with 1 day reserved for potential office tasks), an AO is expected to visit a different village every day. This gives the threshold 20 villages/month. The village visited threshold for high bonus remains the same since higher than 20 villages per month can induce the FA in not maximizing his outreach within each village.

**Distance travelled:** To do his/her basic tasks, an AO travels a minimum of 32 kms daily on average, which gives a threshold of 640 kms per month (assuming AOs are in the field 5 days a week). The threshold remains the same for the high bonus since more distance travelled will not necessarily imply higher performance.

**C. 3: PAYMENT CATEGORY THRESHOLDS FOR AOs IN OBJECTIVE MEASURE SCHEME**

#	MONTH	District	Tehsil	Officer Name	CNIC	IMEI	TA Category / Amount.	Status	Action
1	February 2020	KHANEWAL	MIANCHANNU	Ali Ahmad -FA	3650132629355	866991032962303	<a href="#">Assign Payment</a>	Pending	
2	February 2020	KHANEWAL	MIANCHANNU	Zafar Iqbal -FA	3610404908153	866991033001929	<a href="#">Assign Payment</a>	Pending	
3	February 2020	KHANEWAL	MIANCHANNU	Bashir Ahmad -FA	3610429528067	866991032954037	<a href="#">Assign Payment</a>	Pending	
4	February 2020	KHANEWAL	MIANCHANNU	Muhammad Akram -FA	3610370940913	866991033009302	<a href="#">Assign Payment</a>	Pending	
5	February 2020	KHANEWAL	MIANCHANNU	Muhammad Rizwan Najam -FA	3610472753307	866991033004360	<a href="#">Assign Payment</a>	Pending	
6	February 2020	KHANEWAL	MIANCHANNU	Faisal ur Rehman -FA	3610402975023	866991033007629	<a href="#">Assign Payment</a>	Pending	
7	February 2020	KHANEWAL	MIANCHANNU	Ijaz ul Haq -FA	3610475138243	866991032957121	<a href="#">Assign Payment</a>	Pending	
8	February 2020	KHANEWAL	MIANCHANNU	Ahmed Naeem -FA	3610404533501	866991033000616	<a href="#">Assign Payment</a>	Pending	
9	February 2020	KHANEWAL	MIANCHANNU	Khizar Hayat -FA	3610404912469	866991033011373	<a href="#">Assign Payment</a>	Pending	

C. 4: DASHBOARD PAYMENT ASSIGNMENT PAGE

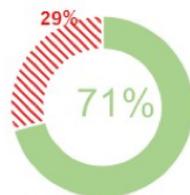
# PIR MAHAL

Monthly Tehsil Report

## AGRI-SMART PERFORMANCE

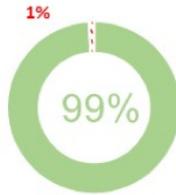
Score: **87%**

Compliance On Days & Hours Worked



Total Staff  
FA/AI: 7, AO:0

FTP Conducted



Total FTPs  
Scheduled: 144

Villages Covered



# of Extension  
Villages: 111

% Time Spent  
on Extension



## VERIFICATION OF AGRISMART DATA

Score: **57%**

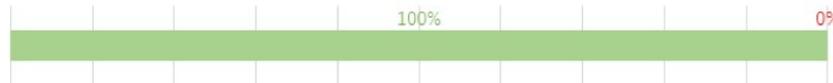
% Farmer Numbers Verified



## FARMER FEEDBACK

Score: **100%**

% Verified Farmers Who Found Extension Advice Useful



## VERIFICATION OF PERFORMANCE BONUS AND OUT-OF-STATION TA PAYMENTS

Score: **0%**

% Field Staff Paid Performance Bonuses



## ADs COMPLYING WITH SOPs

Score: **2/3**



Did AD Upload  
FTP's?



Did AD Process  
Leaves?



Did AD Assign  
Payments?

### C. 5: TEHSIL REPORT

## Appendix D: Sampling Strategy for Different Surveys

*Staff Surveys.* Since our treatment is stratified by tehsil, we select three field staff persons from 130 tehsils every month. This is equal to a 15% audit of all staff representative at the tehsil level.

For each tehsil, entire list of staff, randomly ordered, is shared with our survey firm (RCONS). Enumerators make 3 successful calls per tehsil following the order of staff provided in the sample. Enumerators go through the entire list until 3 respondents are successfully reached. If 3 respondents are not successfully reached, unsuccessful numbers are attempted again after some time lag until 3 numbers are successfully reached per tehsil. No number is tried more than 3 times. Since each tehsil has few AOs and many FAs/AIs, AOs are only be sampled in alternating months, in which the entire staff list for a tehsil is randomly ordered and shared with survey firm. Enumerators are instructed to ensure 3 staff are sampled per tehsil each month. Therefore, if an AO is not successfully reached after 3 attempts, the enumerators will replace with the next available FA or AI in the sample.

Based on data from each month's payment audit, people who retire, transfer out of extension, pass away, or go on long leave have been identified. These staff members are dropped from each month's payment audit sample.

*Agrismart Audit.* The Agrismart audit is conducted by randomly making one call against each employee every month.

For each employee, all relevant activities and their related farmer phone numbers are first fetched from the system. From these activities, two activities are randomly selected – one primary activity and a back-up activity in-case a one-time replacement is required (cases where there is only one reported activity, only one activity is fetched). In-case of an unsuccessful call, the hit-rate for a staff member is marked as 0. We also record employees, who have no valid activities in the survey period.

In terms of time-period from which activities are fetched, all activities for an employee are fetched

from a 2-3 week period prior to the date of the survey. More precisely, the activities are chosen as follows: first, we identify the week in which the survey is to be done for a particular employee, e.g. week 4 of March. We then skip the week before the survey week and fetch all the activities data submitted by each employee in the 2 weeks before this week i.e. for an employee for whom survey is to be done in week 4 of March, skip week 3, and fetch all activities submitted by the employee in week 1 and 2 of March.

All farmer numbers submitted by each employee are pooled for the relevant time-frame and the two activities are then randomly sampled from them. Group activities are sampled roughly 30% of the time based on this random sampling strategy. In the sample, the activity type for each phone number (individual or group) is identified and enumerators conduct the survey accordingly. For example, questions about attendance in an individual activity are only asked of farmers who are reported under individual activity.

*Farmer Call Back.* In the external farmer survey, we leverage an administrative dataset of 1.8 million farmers to make calls to farmers. In this document, we explain the data source in more detail, and outline of several balance checks that we conduct to ensure it is representative across treatment groups and across farmer households in Punjab, and finally an outline of our sampling strategy for the surveys.

Source Data: We use data from the farmer registration exercise by Agriculture Department as the sampling frame for our callback survey. The farmer data includes data for about 1.8 million unique farmers, covering all 36 districts of the Punjab Province.

Coverage of data based on our HR lists: We first check for the overall representation of villages in the farmer registration database. This is important because each employee is assigned a set of villages and missing villages could mean no external farmer data for a set of employees. We find good representation of employee villages in the farmer registration database with only 4% of employees whose villages are not in the overall data.

Overall HR representation

Employees in June 2019 HR data	2525	100%
Non-Traceable Employees in Farmer Data	108	4%
Employees we are able to draw sample against	2417	96%

The designation wise breakdown of the sample, in the table below shows that we have good overall representation for all designations, and the proportion of each designation in the sample looks very similar to the proportion in the original HR data.

Designation	Original HR	Sample	Representation
FA	83%	83%	95%
AO	10%	10%	99%
AI	7%	7%	99%

Farmer Representation: During the merging process, we lost about 340,000 farmers for whom we are unable to trace any employees. We conducted additional tests on these remaining 1.46 million farmers to ensure that the pool farmers we draw sample from continue to be balanced across treatment arms.

The results of balance tests of key farmer attributes as well as Tehsil level differences between the subset of farmer from the Agriculture Census 2010 is given in the balance tests below.

Balance Tests: To ensure that this data is representative for overall farm households in Punjab, as well as balanced across our treatment arms, we conduct the following set of tests:

1. Tests of balance across various treatment arms for key farmer attributes within the farmer data (see Table A)
2. We take the difference in the means of key variables that are reported in Agriculture Census 2010 and Farmer Registration Data. We then check if this difference is balanced across treatments (See Table B)

3. We also conduct a balance test on employee characteristics that are reflected in the FR data (96% of them) (See Table C).

Our tests show good balance across all of these 3 tests.

Sampling Strategy: We randomly sample one farmer + five backups against each employee every month. In terms of logistics, we follow the following steps.

1. Merge farmer registration data to the HR data

- Merge the HR data to the base data using employee ID's to obtain villages assigned to each employee
- Merge the first item to the farmer registration data using:
  - District + Tehsil + village names
  - District + village names
  - Village ID's

2. Sample farmers against each employee:

- For all months after the first month of survey, remove all farmer phone numbers for whom a survey phone call was already attempted
- Use the residual farmer pool and randomly sample six farmers (one principle + five backups) with seemingly valid phone numbers against each FA
- Remove the farmers already sampled against an FA and randomly sample six farmers seemingly valid phone numbers against each AO
- Remove the farmers already sampled against an FA or an AO and randomly

**Table A: Balance Test across treatments in FR (subset of farmers against whom we trace an Ag-Ext employee)**

	Mean				P-value differences					
	Control (I)	Outreach (II)	Subjective (III)	Subjective PLUS (IV)	Control - Outreach (I)	Control - Subjective (II)	Control - Plus (III)	Outreach- Subjective (IV)	Outreach- Subjective PLUS (V)	Subjective- Subjective PLUS (VI)
Owner (=1 if owner)	0.772 (.029)	0.772 (.027)	0.791 (.024)	0.748 (.042)	0.996	0.609	0.634	0.601	0.625	0.371
Owner c tenant (=1 if octenant)	0.077 (.021)	0.083 (.011)	0.077 (.015)	0.079 (.011)	0.79	0.988	0.92	0.747	0.797	0.914
Tenant (=1 if tenant)	0.151 (.02)	0.144 (.021)	0.131 (.019)	0.173 (.038)	0.826	0.479	0.609	0.648	0.514	0.327
Land holding in acres	5.532 (.246)	5.365 (.321)	5.163 (.239)	5.969 (.457)	0.68	0.283	0.401	0.615	0.282	0.12
Tractor (=1 if owned)	0.147 (.012)	0.163 (.014)	0.162 (.014)	0.17 (.012)	0.401	0.418	0.176	0.961	0.694	0.647
Tubewell (=1 if owned)	0.398 (.058)	0.471 (.053)	0.494 (.054)	0.531 (.057)	0.351	0.228	0.102	0.765	0.438	0.632
Canal (=1 if has access)	0.562 (.076)	0.638 (.071)	0.722 (.05)	0.703 (.054)	0.467	0.081*	0.133	0.333	0.466	0.795
Male dependents (number)	2.315 (.087)	2.314 (.164)	2.508 (.118)	2.437 (.143)	0.993	0.19	0.469	0.338	0.573	0.7
Female dependent (number)	2.306 (.077)	2.23 (.159)	2.487 (.114)	2.389 (.148)	0.666	0.19	0.618	0.191	0.463	0.601
Small animals (number)	1.883 (.165)	1.689 (.138)	1.882 (.145)	1.695 (.131)	0.37	0.995	0.374	0.337	0.975	0.34
Big animals (number)	2.764 (.196)	2.531 (.174)	2.905 (.15)	2.721 (.2)	0.377	0.568	0.879	0.106	0.476	0.463
Farmer Age (Completed years on 12/31/2018)	49.918 (.5)	50.698 (.514)	50.674 (.267)	50.457 (.45)	0.278	0.184	0.424	0.967	0.725	0.679
Farmer Education (years)	6.561 (.202)	6.525 (.189)	6.294 (.21)	6.205 (.222)	0.898	0.361	0.239	0.414	0.275	0.774
Farmer Gender (=1 if female)	0.038 (.004)	0.042 (.005)	0.034 (.003)	0.057 (.013)	0.541	0.514	0.167	0.186	0.276	0.093*
<b>Joint F</b>					0.11	0.399	0.24	0.272	0.555	0.884

**Table B: Balance Test across treatments for differences in means of Ag census and FR (subset of farmers against whom we trace an Ag-Ext employee)**

	Mean				P-value differences					
	Control	Outreach	Subjective	Subjective PLUS	Control - Outreach	Control - Subjective	Control - Subjective Plus	Outreach- Subjective	Outreach- Subjective PLUS	Subjective- Subjective PLUS
	(I)	(II)	(III)	(IV)	(I)	(II)	(III)	(IV)	(V)	(VI)
Owner (=1 if owner)	-0.026 (.032)	-0.038 (.027)	-0.072 (.032)	-0.028 (.025)	0.773	0.305	0.95	0.408	0.795	0.279
Owner c tenant (=1 if octenant)	-0.061 (.021)	-0.044 (.016)	-0.012 (.023)	-0.039 (.013)	0.51	0.119	0.368	0.263	0.815	0.313
Tenant (=1 if tenant)	0.089 (.025)	0.082 (.019)	0.086 (.02)	0.068 (.016)	0.818	0.929	0.478	0.877	0.572	0.483
Land holding (total land owned in acres. Only farm HHs)	-0.001 (.269)	0.009 (.324)	-0.39 (.271)	0.225 (.384)	0.98	0.311	0.631	0.346	0.669	0.193
Tractor (=1 if owned)	0.047 (.012)	0.053 (.013)	0.041 (.011)	0.049 (.008)	0.748	0.683	0.897	0.471	0.804	0.537
Tubewell (=1 if owned)	0.279 (.043)	0.297 (.038)	0.287 (.033)	0.288 (.04)	0.757	0.893	0.883	0.836	0.87	0.98
Male dependents (number)	0.953 (.123)	0.838 (.126)	1.016 (.139)	1.124 (.112)	0.514	0.737	0.308	0.345	0.093*	0.546
Female dependent (number)	0.269 (.144)	-0.021 (.15)	0.146 (.144)	0.331 (.135)	0.165	0.546	0.754	0.424	0.083*	0.35
Small animals (number)	-0.753 (.24)	-0.686 (.28)	-0.278 (.204)	-0.751 (.32)	0.856	0.134	0.996	0.241	0.879	0.215
Big animals (number)	0.213 (.158)	-0.224 (.171)	-0.043 (.266)	0.314 (.193)	0.063*	0.411	0.687	0.567	0.039**	0.281
<b>Joint F</b>					0.513	0.379	0.678	0.398	0.818	0.624

**Table C: Balance Test across employee attributes and smartphone indicators (subset of employees who are in sample)**

	Mean				P-value differences					
	Control (I)	Outreach (II)	Subjective (III)	Subjective PLUS (IV)	Control - Outreach (I)	Control - Subjective (II)	Control -	Outreach- Subjective (IV)	Outreach- Subjective PLUS (V)	Subjective- Subjective PLUS (VI)
							Subjective Plus (III)			
Age	45.581 (.948)	45.211 (.501)	45.311 (.722)	45.186 (.891)	0.731	0.821	0.762	0.91	0.98	0.913
Salary	30016.248 (1636.347)	30427.814 (1039.978)	30826.272 (689.454)	30548.853 (1633.306)	0.832	0.649	0.818	0.75	0.95	0.876
Years in service	17.712 (.998)	17.497 (1.059)	17.455 (.861)	17.55 (.686)	0.883	0.846	0.894	0.976	0.967	0.932
Days worked (at onset of intervention)	17.903 (.439)	18.593 (.244)	18.331 (.278)	17.949 (.342)	0.173	0.412	0.935	0.481	0.128	0.387
Hours worked in a day (at onset of intervention)	7.15 (.171)	7.518 (.131)	7.352 (.115)	7.279 (.2)	0.089*	0.328	0.625	0.343	0.318	0.75
Proportion of time spent on extension (at onset of intervention)	0.543 (.018)	0.586 (.022)	0.595 (.026)	0.56 (.019)	0.127	0.106	0.505	0.796	0.381	0.296
Individual farmers visited (at onset of intervention)	10.391 (1.2)	11.021 (.988)	13.675 (1.309)	11.692 (1.287)	0.686	0.067*	0.461	0.108	0.679	0.282
Farmers per group activity (FTP) (at onset of intervention)	21.794 (.721)	23.331 (1.152)	23.917 (.778)	30.719 (5.08)	0.261	0.049**	0.086*	0.674	0.16	0.189
<b>Joint F</b>					0.446	0.182	0.851	0.35	0.452	0.787

# Appendix E: TA DA Approval of Bonus Head

**"SAVE TREES & PROTECT ENVIRONMENT"**



**GOVERNMENT OF THE PUNJAB  
FINANCE DEPARTMENT**

Diary No. 25837  
Date 17/12/2019  
Agriculture Deptt.  
Civil Secretariate Lhr.

**Subject: MONTHLY PERFORMANCE BASED BONUS FOR AGRICULTURE EXTENSION FIELD STAFF (FIELD ASSISTANTS, AGRICULTURE OFFICERS, AGRICULTURE INSPECTORS)**

SO (B&A) AGRIC. DEPTT. 8785  
Diary No. 8785  
Dated 18-12-19

Will the Section Officer (B&A) Government of the Punjab, Agriculture Department, kindly refer to his U.O. No. SO(B&A)8-9/2017-18-Ext.-Vol-I dated 04.11.2019 on the subject noted above?

2. Administrative Department is advised to use existing Travelling Allowance object for claims of the Field Staff. Administrative Department is further advised to use Object Code "A06104-Bonus" for any amount to be disbursed /sanctioned from competent authority as "Bonus".

<b>DGA (EXT. &amp; AR)</b>	
DA (EXT) HQRS	✓
DAC (FT & AR)	
DA (HOV) TP.	✓
DA (P.M.E)	
DA (I.P.M)	
DA (I.M)	
PS-DGA (EXT)	

The Section Officer (B&A)  
Government of the Punjab,  
Agriculture Department.  
U.O. No. 6/26-Agri (FD)/19  
28.12.19

*Hasan Rauf*  
**(HASAN RAUF)**  
SECTION OFFICER (F&C/AGRI)

Dated Lahore, the 16.12.2019

**GOVERNMENT OF THE PUNJAB  
AGRICULTURE DEPARTMENT**

Dated Lahore, the: 26.12.2019

SECTION OFFICER (B&A)  
Ph#99210511  
26.12.19.

Endst. No. SO (B&A) 8-9/2017-18/Ext-Vol-I

A copy is forwarded for information and necessary action to the Director General of Agriculture (Ext. & A.R), Punjab, Lahore in reference to his office letter No. 17583-84/Acs; date: 05.05.2019.

*Very Impth.*  
*Direct.*  
*DD 2.0.*  
*Mr. Usama Zahedi*  
*AGRI DEPT.*

*29/12*

## Appendix F: Misreporting or Gaming in AgriSmart

Despite several built-in checks in the smartphone data to ensure data veracity, treated staff may still have an incentive to: a) misreport data; b) staff in the objective arm may try to game the system by putting in just enough effort to meet thresholds for payment; c) report “better” as opposed actually improving outreach.

In this appendix, we rule out the first two problems.

*Improved outreach or misreporting?* We use two sources of data to verify whether the extension outreach effects are real or due to extension staff attempting to game the system – the smartphone administrative data and the Agrismart audit survey data.

The smartphone data includes phone numbers of all farmers that are visited for extension activities over the 2-year period under study.<sup>11</sup> We use this to construct a farmer phone number database to construct several measures that allow us to test whether extension outreach increased to a unique set of farmers. Table 5 presented first in Section 7 confirms that while all treated groups report more farmer numbers, extension staff in the Objective and Subjective plus arms also reach out to more unique farmers (columns 2 and 3).

Next, we use the Agrismart audit survey data to further assess the veracity of the reported smartphone data. Given the survey is administered to farmers reported in the AgriSmart data, the proportion of successful phone connections (after following the set protocols for phone surveying) allows us to study the quality of reported numbers across treatments. Where other audit indicators are concerned such as whether the farmer knows the AO/FA and/or whether the farmer attended the reported extension activity, ideally we expect to observe no significant treatment effects. This is because negative treatment effects would indicate that the extension worker never visited these farmers despite having reported to do so. At the same time, large positive effects can indicate possibilities of collusion.

---

<sup>11</sup>For group-level farmer activities, the app collects phone numbers for up to 6 farmers. Each group activity has around 20 farmers on average.

Columns 1 and 2 in Table A.11 show that treatment groups do not have a significantly lower ‘survey hit rate’ – i.e. proportion of successful connections. In fact, the hit rate is 2% higher for farmer phone numbers reported in the Subjective Plus arm. Columns 3-6 show that there are no alarming treatment effects on whether the farmer knows the extension agent and on whether the farmer attended the reported activity.

Overall, the analysis based on the farmer phone numbers and the audit data does not point towards strong evidence that extension workers were gaming the system across all treatments.

*Improved outreach or gaming of thresholds in Objective arm?* Given the Objective arm incentivized extension workers against low and high bonus thresholds for specific indicators as measured through AgriSmart, agents may have tried to game the system by misreporting data to meet the threshold for bonus. We use the AgriSmart data to test this by plotting the frequency distribution of each indicators on which the bonuses were given and observe if there is any visible evidence of bunching around the low and high bonus thresholds of the indicators.

Figures B.3 - B.5 show that there is no obvious bunching for any of the variables (days worked, hours worked, village visits, farmer training programmes conducted) except the number of farmers per FTP. This is interesting since the number of farmers reported per FTP is an entirely self-reported measure with no objective details about each farmer that was present. In comparison, the remaining indicators either include automated time-calculation or geo-tagged activities which are harder to misreport against.



## Conclusion

## Conclusion

This thesis explored the question of how to improve the performance of frontline public service agents in the Punjab Education and Agriculture Department in Pakistan through the use of non-financial and financial incentives.

In Chapter 1, I present experimental evidence from a pilot randomized controlled trial that is implemented in collaboration with the Punjab Teacher Training Academy - Quaid-e-Azam Academy of Educational Development (QAED). We embed three different recognition schemes (and a control group) within a routine professional development training of primary and secondary school teachers that incentivize teachers to perform well in the training by tying the recognition reward to training test scores - a private recognition scheme that rewards high performers in teacher training privately; peer recognition that rewards high performers in a public ceremony attended by peers and colleagues; and career recognition that rewards high performers in the teacher training privately but makes career benefits of recognition salient. Each treatment leverages a different underlying theory for why agents may value employer recognition.

The experiment shows null average treatment effects of the recognition schemes on training performance of teachers as measured via training test scores. However, we find heterogeneous treatment effects by the teachers' source of motivation - teachers who report 'intrinsic' reasons for entering the service such as 'interested in the profession of teaching' exhibit a positive treatment response and those who report extrinsic reasons such as 'salary' show a negative treatment effect, with the two effects being significantly different from each other. The treatment effects for each of the individual recognition schemes follow the same trend for intrinsic and extrinsic teachers, however both the positive effect for intrinsic teachers and the negative effect for extrinsic teachers primarily comes from the more outward-facing Peer and Career arms. The stronger positive effect for the more outward (or public) versus inward (or private) facing arms is inline with the existing literature which highlights that more public versus private recognition rewards appear to work better (Markham et al., 2002; Ashraf, Bandiera, and Lee, 2014). Further analysis shows positive treatment effects on endline self-efficacy for intrinsic but negative effects for extrinsic teachers, with the effects being significantly different from each other. This indicates that self-efficacy may

be an important mechanism for how such incentives work.

Chapter 2 builds upon the findings of the pilot presented as Chapter 1. I present experimental evidence on the impact of employer recognition on teacher training performance (as measured via training test scores) in routine in-service trainings for head teachers held by QAED. We embed four different recognition schemes and a control group within this training - Treatment 1 (Peer arm) leverages the peer/collegial approval channel of recognition; Treatment 2 (Career arm) leverages the career-benefits channel of recognition; Treatments 3 and 4 (Public PLUS and Career PLUS) cross the first two treatments with a self-efficacy enhancing frame that aims to bolster teacher perceptions in their ability to do well in the training and their jobs more broadly. The study shows that employer recognition can improve teacher performance in trainings if it is linked to tangible career benefits in the future. Despite these positive results, we find that these effects can backfire depending on how such incentives are framed. In particular, we find that adding a self-efficacy enhancing frame to our recognition treatments “over corrects” teacher beliefs about ability to do well in the training leading to overconfidence and reduced effort.

In Chapter 3, we evaluate the impact of a large scale randomized controlled trial in Punjab, Pakistan that aims to measure the impact of three different pay-for-performance schemes on extension outreach and quality. Leveraging the extension department’s comprehensive digital performance management system called AgriSmart, the incentive schemes link incentives to objective metrics on AgriSmart (Objective arm), supervisors’ own subjective evaluation (Subjective arm), and supervisors’ own subjective evaluation with an element of top-down monitoring to align supervisors’ incentives with the objectives of the principal (Subjective Plus). Our results show that while all treatments improve performance on AgriSmart measures of extension outreach, Subjective Plus also improves performance on farmer reported measures of extension outreach and quality, and farmer-level outcomes. However, these effects are not observed in the Objective or the Subjective arm. In addition, supervisors (ADs) in Subjective Plus also show positive and significant effects on their managerial task of scheduling farmer training programmes (FTPs), which is an essential input into the extension outreach of frontline staff. The positive treatment effects in the Subjective Plus arm across different measures of extension outreach and quality are not entirely conclusive as to whether the positive treatment effects are driven by checking

bias of supervisors (ADs) in their bonus assignments or via improving supervisors' (ADs) own performance and management practices. For example, if the impact observed in Subjective Plus is primarily driven through the AD performance and practices channel instead of incentives for frontline workers, this could have significant cost implications for the government.

*Policy Implications and Future Research - Chapter 1 and Chapter 2.* The results presented in Chapter 1 and Chapter 2 have several policy implications for how to design effective non-financial incentives for teacher in-service trainings, and also (more broadly) for eliciting higher teacher effort in other job-related tasks. First, the career-linked recognition incentive used in this experiment was fairly light touch which simply made future tangible career benefits of recognition salient. Yet, we find encouraging results on teacher performance in in-service trainings which indicates towards the value of informal career benefits in the system. In our particular context, there are several such informal career benefits such as getting a transfer to a school of liking, getting laterally appointed to an influential position (such as Project Director) of a large donor-funded program, or getting appointed to a higher grade position (with the same pay and civil service grade) if a vacancy arises. Our results highlight that there is potential to design even sharper non-financial incentives that can make the link to such informal career benefits sharper. In the public sector, where formal incentive-based reforms are often hard to implement, designing “soft” incentives that can leverage such informal career incentives can address part of the inefficiency in incentive systems.

Second, the sensitivity of our results to framing effects and non-cognitive traits points towards the external validity of studies that evaluate the impact of such incentives. For example, while the peer recognition/approval channel appears to work for more junior primary and secondary school teachers in Chapter 1, it is less effective for head teachers in Chapter 2. Similarly, while we observe negative effects on individual self-efficacy for certain groups of teachers in Chapter 1, we do not observe similar effects for head teachers in Chapter 2. These results across the pilot and the main experiment indicate that recognition incentives are highly contextual where the effects can vary a lot depending on cadres and non-cognitive traits of those cadres. This requires caution in how such incentives are designed across different contexts and also indicates the importance of piloting before scaling up across the board.

Several additional questions remain open to inquiry which could be explored further through ongoing engagement with QAED. Our experiment was only able to offer the recognition incentive for a single time. Future work could look at the decay rate in the impact of such incentives in trainings, and circumstances under which the effects are sustained. Recognition has been often modelled in standard principal-agent frameworks, but clarity around the weight placed on such incentives in comparison to financial incentives would be useful in calibrating their value and assessing the cost effectiveness of such incentives more explicitly. The null treatment effect of our incentives on school-level outcomes raise important questions regarding the extent to which incentives in trainings can improve the intended downstream impacts of in-service trainings at the school and classroom level. Given implementation of high quality trainings and achieving their intended downstream effects is generally challenging, this calls for further academic inquiry into the extent to which incentives in trainings can encourage such downstream implementation, and whether certain types of incentives are more effective than others in achieving that. Finally, our experiment showed that creating exogenous variation in self-efficacy beliefs of public sector bureaucrats is possible. This opens up the possibility of additional research on how to create exogenous variation in intrinsic motivation and/or other non-cognitive traits, and their impact on training performance, implementation of training at the school and classroom level, and other job-related tasks.

*Policy Implications and Future Research - Chapter 3.* The findings of our study will have direct relevance for the Punjab Agriculture Extension Department. The current findings indicate that while a purely Objective arm does increase extension outreach on several outreach metrics as measured via AgriSmart, Subjective Plus improves extension outreach on metrics of AgriSmart and also farmer reported measures of extension outreach and quality. At the same time, the Subjective Plus is also twice as costly as the Objective arm. Ultimately, the government's choice of the incentive scheme will depend on the kinds of trade-offs each scheme induces and the related costs and benefits. The impact evaluation will directly support this analysis.

Our results contribute to the literature on how to design effective financial incentives in public services where tasks have both easily measurable and hard to measure dimensions and supervisors hold important local information. Our results will also contribute to the literature on rules-

based bureaucracy versus discretion. In particular, it will show circumstances under which subjective evaluations are effective or ineffective, how supervisors use their private sources of information versus objective sources of information, what type of supervisors use their discretion more effectively, and the kind of trade-offs induced by purely objective metrics versus subjective metrics.

## References

- Ashraf, Nava, Oriana Bandiera, and Scott Lee. 2014. "Awards unbundled: evidence from a natural field experiment." *Journal of economic behavior and organization* 100:44–63. URL <http://search.proquest.com/docview/1523807233/>.
- Markham, Steven E. et al. 2002. "Recognizing Good Attendance: A Longitudinal, Quasi-Experimental Field Study." *Personnel Psychology* 55 (3):639–660.