The Double-Edged Sword of Recognition Rewards: Evidence from the Education Sector in Pakistan

Zahra Mansoor *

October 2021

Abstract

This paper presents experimental evidence on the impact of employer recognition on head teacher performance in a public in-service teacher training programme. I design and implement an experiment in collaboration with the teacher training academy in Punjab, Pakistan in which 3,394 head teachers attending a mandatory training are randomized into 4 different designs of recognition incentives that are tied to training performance, and a control group. The first recognition incentive makes peer-esteem in recognition salient, the second makes potential career benefits of recognition salient, and the third and fourth treatments combine the first two treatments with a motivational framing to improve the design of recognition rewards. I find two key results: first, employer recognition that makes potential career benefits salient improves training performance; second, adding a motivational framing improves teacher motivation but makes teachers overconfident reducing their performance in the training. These findings have implications for understanding how agents make decisions to exert effort in response to recognition rewards.

Acknowledgements: This research was made possible through numerous insights from partners at the Punjab Education Department. In particular, Mr. Waseem Shirazi, Mr. Nadeem Hussain, and Mr. Adnan Bashier at the Quaid-e-Azam Academy for Educational Development offered many helpful reflections. I thank Julien Labonne, Stefan Dercon, Clare Leaver, Christopher Woodruff, Simon Quinn, Michael Callen, Adnan Qadir, Martin Williams, and participants at CSAE workshops (2019, 2020), Nordic Development Conference (2021), Applied Development Economics Conference (2021), and OxDEV (2021) for helpful comments. I gratefully acknowledge funding from Economic Development Institutions (EDI), Center for the Study of African Economies (CSAE), and the Blavatnik School of Government (BSG). Usman Zahid offered excellent research support. The study is pre-registered at the AEA RCT registry (0004482). IRB approval was obtained at the University of Oxford. All errors are my own.

∗ Postdoctoral researcher, 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 a challenge in the education sector where low effort of public school teachers remains a critical concern across several developing country contexts (Bold et al., 2017). Public sector organizations have relied on financial incentives to address low effort of teachers with promising results (see Muralidharan and Sundararaman, 2011; Duflo et al., 2012; Mbiti et al., 2019; Leaver et al., 2021). But financial incentives, though effective when designed well, are expensive, and can often be distortionary via creating perverse incentives or crowding out intrinsic motivations (Titmuss, 1970; Holmstrom and Milgrom, 1991; Glewwe et al., 2010; Gneezy et al., 2011). Non-financial incentives such as recognition and status rewards, on the other hand, are financially more feasible and potentially effective in pro-social settings like education where frontline agents may put a lower weight on financial incentives (Besley and Ghatak, 2005). In this paper, I study the impact of non-financial incentives such as employer recognition on public school teacher effort.

Governments often design and roll-out employer recognition schemes, but evidence within the public sector with respect to the mechanisms through which these operate remains limited.[1] First, there is a question regarding the different channels of motivation for recognition. Theory highlights that agents may value employer recognition because it generates peer-esteem by being recognized in front of peers and colleagues (Besley and Ghatak, 2008; Frank, 1985) or because agents can use recognition to signal quality to supervisors to reap tangible career benefits in the future (Dewatripont et al., 1999; Besley and Ghatak, 2008). While existing evidence documents the impact of public versus private forms of employer recognition (as in Ashraf et al., 2014) or the impact of employer versus community recognition (as in Gauri et al., 2018), no existing study tries to unbundle the impact of the *peer-esteem* and *career-benefits* channels of employer recognition.

A second question relates to how such incentives are framed. This is relevant because theory by both psychologists, and more recently, economists argues that performance-contingent rewards

[1] 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).

can often dampen employee morale and beliefs in ability because they are perceived as controlling or as a negative signal of ability (Deci and Ryan, 1985; Bénabou and Tirole, 2003). These negative effects might especially be more pronounced in tournament-based rewards like recognition for low-ability individuals or individuals with low intrinsic motivation at baseline (see systemic review by Connelly et al., 2014; Ashraf et al., 2014, Mansoor, 2019). What this fundamentally implies is that equilibrium effort does not only depend not on the reward but also on the *information* contained in the reward (Gneezy et al., 2011), which makes the framing of rewards important. Despite this evidence, and calls for attention towards how performance-contingent rewards are framed (Frey, 1997), there is limited evidence on how to improve the framing of such incentives to improve their design.

In this paper, I aim to study these gaps and provide experimental evidence on the impact of employer recognition on head teacher effort. I ask the following two questions: First, could employer recognition elicit higher teacher effort? If yes, what is the impact of employer recognition that makes peer-esteem salient versus recognition that makes future tangible career benefits salient? Second, does framing recognition with a motivational framing that boosts morale and beliefs about ability improve their design and impact?[2]

To shed light on these questions, I collaborate with the Education Department in Punjab, Pakistan. Punjab is Pakistan's most populous province, employing a workforce of approximately 400,000 teachers spread across 52,000 schools.[3] Progress in learning outcomes has remained a challenge in Punjab, and teacher quality and effort is widely seen as one of the main explanations for low student learning outcomes in the province. 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 more than 70% of the teachers value non-financial drivers such as employer recognition, community recognition, and professional development opportunities.[4] This contextual setting creates direct relevance to study our questions of interest.

[2]In a pilot experiment preceding this paper (available as an online appendix), we show that more intrinsically motivated individuals at baseline respond more to recognition rewards. This provides us with additional supporting evidence to explore the second question.

[3]Annual School Census Data 2017.

[4]Based on author's own analysis of the Teacher Motivation survey data.

Within the Education department, 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). We embed the experimental design within a mandatory in-service training on school management and leadership offered to 3,394 head teachers with the recognition rewards being directly linked to *training performance* as measured through training test scores. This setting has several specific features and advantages which makes it relevant to study the questions of interest. First, the training is offered to head teachers who are critical agents for school performance in their capacity as school leaders (Hallinger and Heck, 1996; Silins and Mulford, 2015). Thus, understanding how to motivate them has direct implications for school performance. Second, attending the training is mandatory for all head teachers where exerting effort in the training is a function of the teacher's own motivation. This is arguably similar to many other non-student facing mandatory tasks that head teachers are required to do such as preparing workplans and budgets. Hence, understanding head teacher response to incentives in trainings allows us to understand broader head teacher preferences for these rewards. Third, we are able to learn about teacher preferences for these rewards by incentivizing training test scores which are one-dimensional in effort as opposed to incentivizing teacher performance in the classroom which is multi-faceted and complex. This allows us to avoid the standard issues of multi-tasking (Holmstrom and Milgrom, 1991).

As part of this experiment, the training academy randomly allocated 131 different training sessions (offered to 3,394 head teachers) across 4 different employer recognition schemes and a control group. Treatment 1 (Peer arm) makes 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' that will be shared with the departmental leadership and 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 or transfers to other lateral postings. Treatments 3 and 4 (Peer PLUS and Career PLUS) cross the first two

treatments with a motivational framing that is focused on boosting teacher morale and beliefs about ability to do well in the training and their jobs more broadly.

Our main treatment effects show two key results. First, we find that employer recognition can work when it is linked to tangible career benefits in the future. We find indicative evidence that the Career arm leads to a 0.25σ 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. This implies that head teachers value recognition rewards when they increase the likelihood of being able to access informal career incentives in the system. Second, we find that the positive effects of these incentives can backfire depending on how they are framed. The net impact of adding the motivational framing in Career PLUS is negative and significant – a reduction of 0.28σ in training test scores. The net impact of the framing in Peer PLUS is also negative but insignificant. Overall, the net impact of adding the motivational framing to the Peer and Career arms is to lower training test scores by 0.22σ.

To understand mechanisms underlying the treatment effects of the Career arm, we look at quantile treatment effects and find 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. This confirms that the observed treatment effects of the Career arm are not merely coming from the top of the distribution, and that they are stronger than the Peer arm. We also find that when the Career arm works, it works through individuals who are due for their next promotion sooner, have higher visibility to senior leadership such as Secretary Education and District Education Officers (DEOs), or are permanent employees. Given teachers have informal career incentives in the system such as through their preferred positions and postings, it is clear how an upcoming promotion can make such career-benefits more salient. Similarly, higher visibility to senior leadership implies greater access to supervisors for accessing informal career benefits which can explain why the Career arm could be more valuable for teachers who have higher visibility. These results lend further evidence in support of the career-benefits channel of recognition.

To understand the negative effects of the motivational framing, we test treatment effects on a teacher motivation index comprised of measures of self-efficacy, internal locus of control, and intrinsic motivation. We find that the impact of the PLUS arms on the motivation index is

positive and significant. This highlights that despite negative effects on training test scores, the PLUS treatments did bolster teacher motivation as expected.

Finally, our main hypothesis for explaining the negative effects of the PLUS treatments investigates if the motivational framing made teachers' over confident in their ability 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 framing 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.

These results add to the existing empirical literature on non-financial incentives in the public sector (Ashraf et al., 2014; Gauri et al., 2018; Ashraf, Bandiera, and Jack; 2014), highlighting one possible way to improve their effectiveness. This paper provides, to the best of our knowledge, the first empirical evidence that unbundles the peer-esteem and career-benefits channels of employer recognition. While existing theory argues that agents can have multiple motivations for employer recognition such as individual self-esteem, peer-esteem, or receiving tangible career benefits in the future (Dewatripont et al., 1999; Besley and Ghatak, 2008), these motivations are conceptually distinct which has implications for how such rewards are designed. This paper provides evidence that recognition rewards that are linked to tangible career benefits can improve the performance of frontline bureaucrats such as teachers.

Our results also directly link to the literature 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). Given all in-service trainings are not promotion-linked in the public sector, this paper shows that linking teacher performance to informal career incentives could be one possible way of feasibly incentivizing teachers in various types of in-service trainings.

This paper also adds to the literature on framing of incentives, with particular attention on how to improve the information contained in performance-contingent rewards to address negative effects on morale and beliefs about ability (as argued by Deci and Ryan, 1989; Connelly et al., 2014). Existing evidence on framing of incentives mainly focuses on the impact of framing incentives as losses or gains on employee performance (Goldsmith and Dhar, 2013; Lagarde and Blaauw, 2021). To the best of our knowledge, this is the first experimental study that offers performance contingent rewards (in this case recognition) in combination with a motivational framing. Our study highlights that framing non-financial incentives such as recognition with motivational framing focused on boosting morale and beliefs about ability can backfire by making teachers overconfident in their ability. This highlights that the framing of such incentives interacts with individual non-cognitive traits of the workforce in substantial ways. This can help explain why in some contexts such incentives work but in others they do not (as in Gauri et al., 2018). Overall, this highlights caution in the overall design, and in particular the framing of such rewards.

Finally, this paper contributes to the broader literature that focuses on improving public sector performance in developing countries. Implementing interventions on the selection margin by tweaking recruitment policies (as in Ashraf, Bandiera, and Jack, 2014; Deserranno, 2019) or on the performance margin by introducing formal pay-for-performance reforms (as in Khan et al., 2016) can often be hard to implement due to regulatory or financial constraints. While it is true that formal career incentives in public bureaucracies are negligible given promotions are determined by seniority (Finan et al., 2015), our results indicate that head teachers have career concerns through other informal mechanisms that may be easier to exercise in the system as compared to promotions. This paper highlights that understanding the sources of different agent 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 the challenge of weak 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 outlines the experimental context and Section 4 describes the experimental design, randomization, and data sources. Section 5 presents the empirical strategy and main results, and Section 6 presents mechanisms for understanding the

impact on our main outcome. Section 7 concludes.

2 Conceptual Framework

In this section, we present a simple conceptual framework to analyze how agents decide to exert effort in response to recognition rewards. A key aspect of our conceptual framework is that it captures the intuition that when agents decide to exert effort in response to performance-contingent rewards, the decision depends on the interaction between the reward and the *information* contained within the reward.

Assume that the principal cares about maximizing the value from the training (i.e. learning gains from training) Y, and employee effort e in the training produces value denoted by

$$Y = f(e) + \epsilon \tag{1}$$

Agents are risk neutral and care about maximizing expected utility. We assume that agents maximize utility over *extrinsic* payoffs from receiving the recognition reward and *intrinsic* payoffs based on the information contained within the reward. Also assume that the agent puts a weight of β_i over his/her extrinsic gain from the reward, and a weight $1 - \beta_i$ over the intrinsic payoff based on the information contained within the reward. In addition, there is a cost to effort given by $c(e)$, which is strictly increasing in effort.

The timing is as follows. The principal introduces a recognition scheme through which a worker i with effort e gets a recognition reward R based on the probability $p_i(e)$. The worker then chooses his/her effort levels e to maximize utility denoted by

$$\beta p_i(e)R + (1 - \beta)I(e)R - c(e). \tag{2}$$

We assume that effort is strictly positive and decreasing in the extrinsic payoff from the recog-

nition rewards (as is the case in standard utility maximising frameworks with incentives) such that $p_e>0$, $p_{ee}<0$.

Intrinsic payoffs depend on the way in which information contained in the reward is perceived by the agent. We assume this to be a function of a range of features such as agent's own ability, beliefs about ability, agent's own cognitive processes, and the outside signal given by the principal through the framing of the incentive. We capture this through the reduced form information function I(e) given in equation 2.[5] Frey (1997) and Deci et al. (1999) argue that the impact of rewards on performance ultimately depends on whether the rewards are administered in a motivational or controlling way, where the former can harness motivation but the latter can dampen it. To capture this intuition, we assume effort is strictly positive and decreasing in intrinsic payoffs if the information contained in the administration of the reward is motivational, such that $I_e>0$, $I_{ee}<0$. This would, for example, be in situations where recognition rewards are framed in a way that bolsters individual self-esteem and confidence, and hence increases intrinsic payoffs to effort. However, if the information contained in the reward's administration is controlling, we assume effort is decreasing in intrinsic payoffs such that $I_e<0$, $I_{ee}<0$. This difference in the functional form of I(e) depending on the framing of the reward captures the theoretical intuition that information could either harness or dampen motivation.

The first order condition that characterizes the employee's effort choice is as follows:

$$\beta\frac{\partial p_i(e)}{\partial e}R + (1-\beta)\frac{\partial I((e)}{\partial e}R - \frac{\partial c(e)}{\partial e} = 0 \tag{3}$$

This shows that if $1-\beta > 0$, then the decision to exert effort in response to the recognition reward will be determined both by the agent's extrinsic payoff but also the agent's intrinsic payoff which is determined by the information contained in the administration of the reward. The payoff will always be positive if the information is conveyed in a motivational way since $I_e > 0$. However, if the information is conveyed in a controlling way, the second argument will be negative. In this situation, the overall effect of the reward can be negative if the negative intrinsic payoffs

[5]Note that agent equilibrium utility also depends on non-effort dependent intrinsic payoff which captures motivations such as warm glow and personal satisfaction (Andreoni, 1989). We abstract away from this for simplicity as this does not determine effort choice.

outweigh the positive extrinsic payoffs.

Our experimental design relates to the first argument and the second argument in equation (3) above. We design and test the impact of recognition rewards leveraging the peer-esteem or the career-benefits channel to understand how to strengthen agent response to recognition rewards (i.e. the first argument). This is because agent's can have two key motivations for valuing employer recognition: first, the motivation for peer-esteem and approval as such awards can lead lead to respect, reputation, and peer-esteem (Frank, 1985). 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 et al., 2014).

We also combine these incentives with a motivational framing to understand whether intrinsic payoffs to the agents from these rewards can be improved by exogenously changing the framing of the rewards as performance-contingent rewards are often perceiving as controlling and can often dampen employee morale and beliefs in ability (Deci and Ryan, 1985; Bénabou and Tirole, 2003). (i.e. second argument).

Our hypotheses are as follows:

Hypothesis I: Employer recognition tools that leverage the peer-esteem channel improve training performance of teachers relative to the control group because agents care about recognition and 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.

Hypothesis III: Employer recognition tools that leverage the peer-esteem channel in combination with a motivational framing improve training performance more than the peer-esteem channel only.

Hypothesis IV: Employer recognition tools that leverage the career benefits channel in combination with a motivational framing 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. Hypotheses III and IV test whether the motivational framing of incentives improves the design of the recognition incentives.

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.[6] The public education system employs a workforce of approximately 450,000 teachers responsible for educating nearly 11 million children spread across 52,000 schools.[7] The School Education Department (SED) is the provincial public body which holds the mandate for all policy and implementation pertaining to primary and secondary education.[8]

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).[9] 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).

[6] Pakistan Population Census, 2017 (Pakistan Bureau of Statistics).

[7] Annual School Census Data 2017.

[8] Schools are further divided into primary (grades 1-5), elementary (grades 6-8), secondary (grades 9-10), and higher secondary (grades 11-12) schools.

[9] See I-SAP (2017).

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.[10] This contextual setting creates direct relevance and demand to generate rigorous causal evidence on the effectiveness of non-financial incentives such as employer recognition.

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, 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 into practices at the school and classroom level. 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.[11]

[10] Based on author's own analysis of the Teacher Motivation survey data.

[11] This is barring promotion-linked trainings which are offered to promotion eligible teachers. All participants

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.[12] 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 head teachers. Given capacity constraints at the district training center, the training was spread over 3 sequential rounds to accommodate all the sessions.[13]

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.[14]

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.

[12]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.

[13]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.

[14]we return to the discussion on the design of the tests and nature of question in Section 4.3.1.

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-esteem or career-benefits of recognition salient. The second two treatments cross the first two recognition incentives with a motivational framing that focuses on boosting morale and individual beliefs in their ability.

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 et al., 2014).

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.[15] This is followed by the scheduled training over the next four days. On the fourth and 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.

[15]To ensure quality and uniformity in the administration of the recognition incentives across all treatments, 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.

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-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' which would be shared with their district's leadership as well as the provincial leadership of the School Education Department, which could make them eligible for future career opportunities in the department. The script for T2 is exactly the same as T1 except for the difference in how career benefits as opposed to peer-esteem is 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 public school teachers in our context are limited since promotions are linked to seniority, three types of informal career incentives might be relevant for how head teachers can use the recognition reward to their advantage. 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.[16] 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 motivational framing to boost

[16]Employees who are eligible for promotion have to wait for their turn to get their promotion approved.

individual morale and beliefs in ability to do well in the training and the job more broadly. The motivational framing aims to improve the way information contained in the reward in perceived by employees. 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 motivational framing as in T3.

Motivational framing: The goal of this framing is to improve the information contained in the reward. This is because performance-contingent rewards can often negatively impact employee morale and beliefs in ability (Deci et al., 1989; Bénabou and Tirole, 2003), and these effects can be more pronounced for low-ability individuals in tournament-based rewards like recognition (Connelly et al., 2014).

To design this framing, we draw on the seminal work of Bandura (1986), who defines the concept of *self-efficacy* as the "perception of one's capability to accomplish a given level of performance", and the ability of vicarious experiences in boosting it.[17] The framing is structured as follows: teachers are first asked to reflect upon three key limitations and challenges in performing well in the training and their jobs more broadly. This is followed by the distribution of a one-pager with three inspirational stories of head teachers from Punjab that the trainees are asked to read. The stories are meant to serve as role models to bolster existing levels of belief in one's capability and ability (as in Beaman et al., 2012 and Tanguy et al., 2014 for example).[18] To create a final moment of reflection, after reading the stories trainees are asked to reflect on how they can address their own limitations.

[17]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. In designing the framing, we focus on teachers' domain specific self-efficacy

[18]We select these stories from a report on star teachers compiled by the Punjab School Education Department in 2017 to identify and record high performing teachers.

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 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.

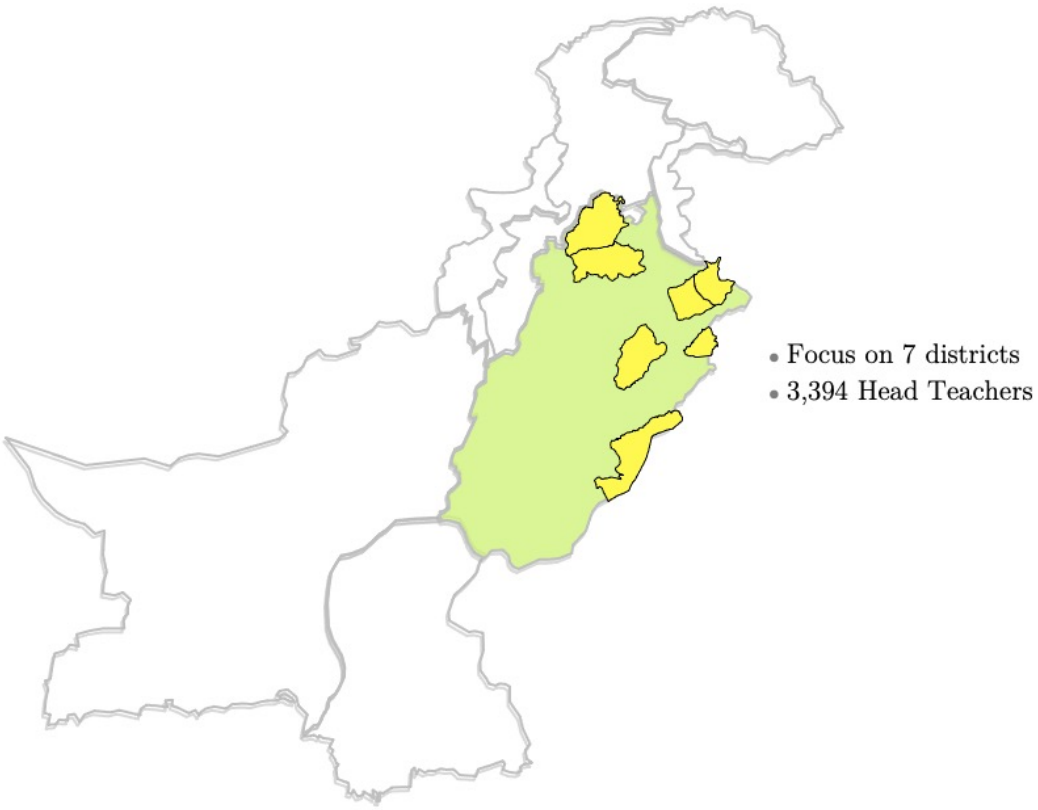


Figure 1: Districts included in the Sample

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 the tests. 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.[19]

We are also interested in understanding the extent to which teacher effort on the training test is driven 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 motivational framing versus the non-incentivesed dimension where the interaction effects are muted in comparison.

Teacher Surveys at endline. We collect a set of key variables through teacher surveys at endline to

[19]See Appendix D for a sample of the training tests.

understand mechanisms. These include attitudes and perceptions that can capture the effect of the recognition rewards and the motivational framing such as teacher self-efficacy, locus of control, beliefs about performance on the post-test, and intrinsic motivation. 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 motivation are captured using a tested battery of questions as in Amabile et al. (1995). Locus of control is captured through the RAND measure

*Teacher Surveys at baseline.*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; and non-cognitive traits such as personality type, intrinsic/extrinsic motivational orientation, pro-social motivation. 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).[20] We also measure self-efficacy, locus of control, and intrinsic motivation at baseline.

We also collect data on specific variables to test theory underlying how our treatment arms may operate. These include variables such as number of peers known, time till next expected promotion, contract type, and visibility to senior leadership.

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.

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

[20]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.

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.[21]

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.[22] 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 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.

[21] We also conduct joint F tests across all groups. All p-values for the joint test are greater than or equal to 0.10.

[22] This occurred either due to personal emergencies or teachers being absent without any officially sought leave.

5 Empirical Strategy and Results on Training Performance

5.1 Empirical Strategy

To identify the main treatment effects of our interventions on training test scores, we estimate the following:

$$y_{isd}^{Post} = \beta_0 + \rho . y_{isd}^{Pre} + \beta_1 PR_{isd} + + \beta_2 CR_{isd} + \beta_3 PRplus_{isd} + + \beta_4 CRplus_{isd} + \gamma X_{isd} + \mu_d + \alpha_r + \epsilon_{isd} \quad (4)$$

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 μ_d) to increase the efficiency of our estimate. We also control for training round effects, α_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). The β coefficients are the coefficients of interest. This specification allows us to test the first and second hypothesis ($\beta_1 > 0$ and $\beta_2 > 0$) as well as the third and the fourth hypotheses ($\beta_3 > \beta_1$ and ($\beta_4 > \beta_2$).

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.[23] Table A.5 shows that non-consent is not significantly related to any of the treatment groups.[24]

[23] This included 207 teachers which is roughly 6% of the sample.

[24] Our main treatment estimates remain the same with TOT estimation.

5.2 Main Effects on Training Performance

Treatment effects of Peer (T1) and Career (T2) Arms. Table 1 shows the treatment effects on standardized training test scores. We first focus on the treatment effects of the Peer and Career arms as shown in Column 2.[25] We find that teachers in the Career arm score 0.25σ higher training test scores as compared to the control group (significant at the 10% level). In comparison, the Peer arm has a coefficient of 0.04 and is insignificant. Kolmogorov–Smirnov tests of equality of distribution between the Peer and Career arm confirm that the distributions are significantly different from each other at the 1% level.[26]

Given the Career arm made potential career benefits salient, these results point towards the value of informal career benefits in the system that the teachers could have accessed through the recognition certificate (e.g. 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-esteem 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 strong informal career incentives in the system such as postings to their choice of school or other influential lateral appointments.

Treatment effects of Peer PLUS (T3) and Career PLUS (T4) Arms. Next, we focus on the treatment effects of the PLUS interventions with the motivational framing. Column 2 in Table 1 shows that the net impact of adding the motivational framing in Peer PLUS and Career PLUS is negative – a reduction of 0.15σ for Peer PLUS (not significant) and a reduction of 0.28σ 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

[25] Since we had imbalance on some teachers characteristics, we discuss the results where we include controls in column 2.

[26] The distributions are presented in Appendix B.3.

that the distributions are significantly different from each other (p-value <0.01 for both tests).[27] Since the net impact of adding the framing moves in a negative direction for both arms, we pool the PLUS treatments (i.e. those receiving the framing) and the non-PLUS treatments (i.e. those not receiving the framing) in Columns 3 and 4. Column 4 shows that the net impact of adding the motivational framing to either of the arms is to lower training test scores by 0.22σ (significant at the 5% level).

The negative effects of the PLUS interventions imply that adding the motivational framing to our recognition incentives resulted in teachers reducing effort on the test, which runs counter to our hypotheses in Section 2.[28] While our findings are inconsistent with the positive effects of similar motivational interventions (that focus on boosting aspirations and self-efficacy) on a range of outcomes such as job search and health-seeking behaviours (see Eden and Aviram, 1993; Haushofer, John, and Orkin, 2019), they are in line with arguments of skeptics who suggest that creating "positive illusions" of oneself can often have negative effects by leading to overconfidence (Baumeister, 1999; Swann, 1996).[29] They also relate to recent empirical evidence by McKenzie et al. (2021) on how such motivational interventions may backfire, by showing how raising financial aspirations of poor entrepreneurs can lead to frustration and lower economic investment if aspirations are unmet. Overall, this suggests caution in the design and administration of motivational interventions as they may interact with individual non-cognitive traits in nontrivial ways.

[27] The distributions are presented in Appendix B.4.

[28] Our analysis of the net impact of the framing assumes that the PLUS treatments are a linear combination of the incentive scheme and the motivational framing. However, if the two treatments bundle together in non-linear ways our point estimates of the net impact of the framing would not be accurate. Irrespective of this assumption, our results do show that the bundled treatment with the framing does worse than the recognition 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 framing on the effect of the incentives. This interpretation is equally relevant for the implications of these results.

[29] 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.

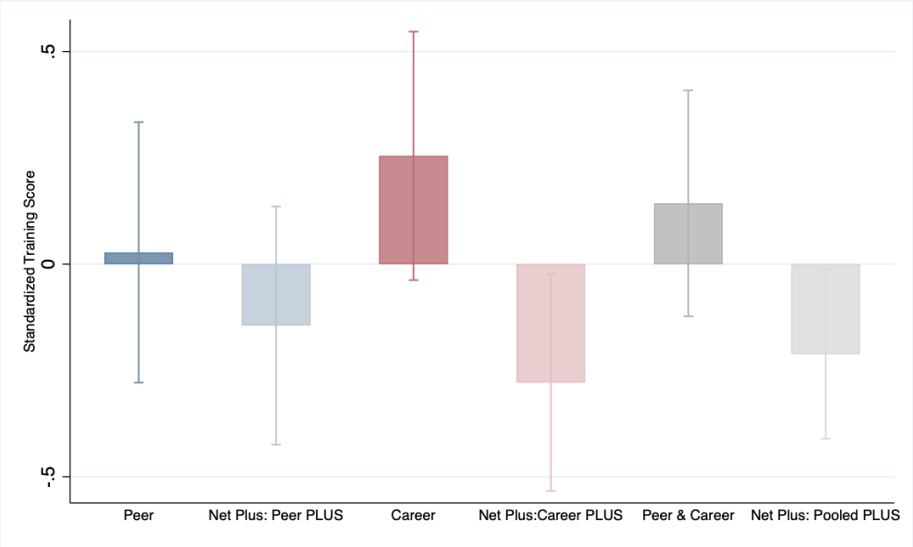


Figure 2: Main Treatment Effects

6 Mechanisms

Our main treatment effects highlight two main sets of results: first, that the career-benefits channel appears to improve training performance. Second, the motivational framing appears to have negative effects on training performance.

In this section, we follow these two sets of results one by one to understand mechanisms underlying these effects.

6.1 Mechanisms underlying treatment effects of the Career arm

Quantile Treatment Effects. While the Career arm appears to work, an important question is whether the effects are driven by the lower or upper tail of the distribution of training test scores. This is especially important in the context of tournament-based rewards (such as our recognition incentive) which often merely elicit effort from agents in the upper tail of the ability distribution

(Connelly et al., 2014). To understand the heterogeneity of our treatment across the distribution of training test scores, we estimate quantile treatment effects.

Table 2 shows quantile treatment effects at quantile $\tau \in$ (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9). It shows that the Career arm has a positive coefficient in the range of 0.20 - 0.37 across the entire distribution of the training test scores and that the coefficient is always higher than the Peer arm (see Figure 3). The coefficients on the Career arm are significant at the 5% level in the upper tail of the distribution, where they are also significantly different from the Peer arm. These trends confirm that the career-benefits channel encourages effort across the distribution of training test scores, instead of merely high ability individuals as proxied by training test scores. It also provides further evidence that the career-benefits channel is stronger than the peer-esteem channel in this context.

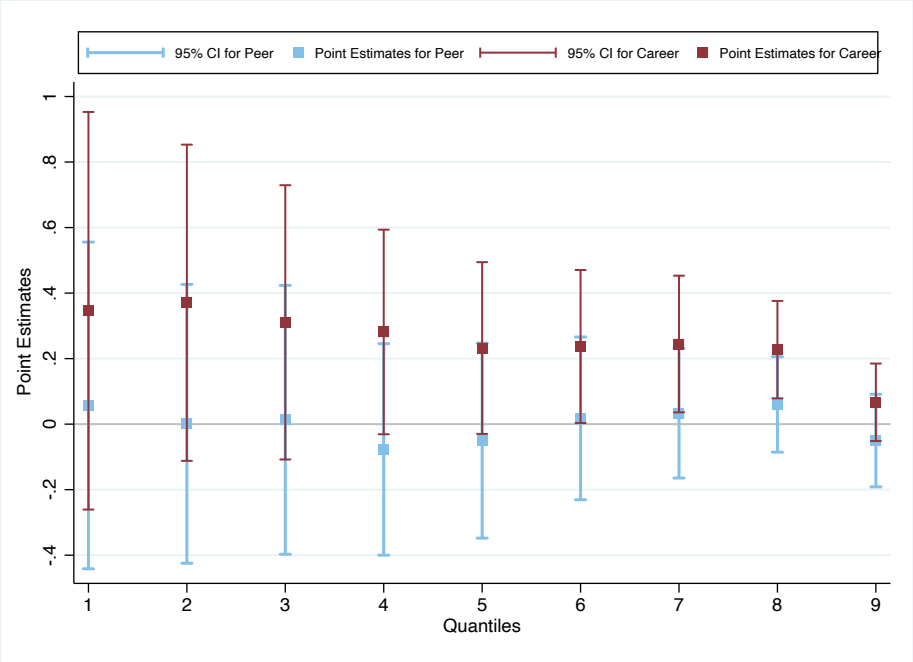


Figure 3: Quantile Treatment Effects

Heterogeneity by Key Moderators of the Career arm. Next, we investigate whether the Career arm works as hypothesized in theory, i.e. whether agents respond to the incentive because they believe it could result in tangible career benefits in the future.

To investigate this, we hypothesize that the Career arm should work better for teachers who have a higher likelihood of being able to use the recognition certificate to reap tangible career benefits in the future. We identify three such categories of teachers: teachers who are expecting a promotion soon, teachers who have higher visibility to senior leadership, and teachers who are permanent employees. These teachers are more likely to use the recognition incentive for career benefits because of the following reasons. First, given teachers in our context have informal career incentives of being posted to their postings of preference (e.g. a school of their liking or other lateral appointments), an upcoming promotion can make these incentives more salient. In our sample, the median time to next promotion is 5 years. We identify teachers as *more promotion eligible* as those who have an upcoming promotion within the next 5 years. Second, higher visibility to senior leadership can increase access to supervisors which can increase opportunities to reap informal career incentives in the system. We identify teachers where frequency of visits by senior leadership (such as Secretary and District Education Officers) to the teachers' districts is at least once in 3 months as *more visible to leadership*. Finally, 17% of our sample includes head teachers who are working on a contractual basis and do not have the same career incentives as permanent teachers. We identify teachers who are on permanent contracts as *permanent* employees.

Columns 1, 2, and 3 in Table 3 show that when the Career arm works, the treatment effect comes from individuals who are due for their next promotion sooner, have a higher frequency of visits by the Secretary, or are permanent employees. Of particular note is the fact that the difference in the treatment effect of the Career arm between teachers who have less time to promotion (0.33 σ) versus those who have more time to promotion (0.13 σ), and between permanent (0.28 σ) and temporary employees (0.02 σ) is statistically significant (see Figure 4).[30]

[30]We also hypothesize that the Peer arm should work better if teachers know their peers well in the training session. This rests on the assumption that the *peer-esteem* from the Peer arm would be stronger if a teacher knows his/her peers. To operationalize and test this theory, we also explore heterogeneous treatment effects by the number of peers each trainee knows in their group but find no significant effects (See column 4)

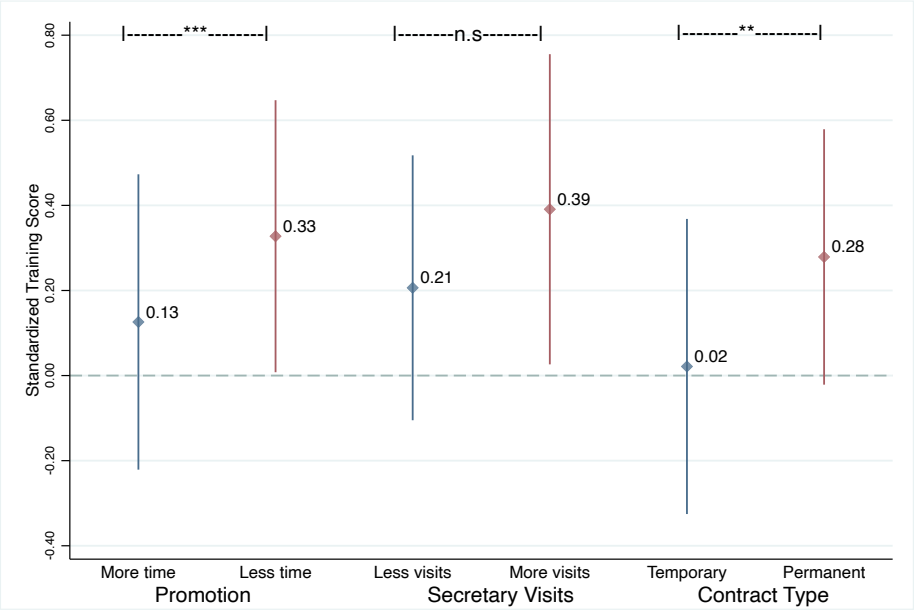


Figure 4: Career Arm Moderators

Treatment Effects on the Non-Incentivized Dimension. Finally, to confirm if the Career arm is working as an incentive, we can also observe the effects on the non-incentivized dimension of the test. To estimate these effects, we use the same specification as (4) but with the non-incentivized test scores as our outcome variable.

Column 6 in Table 4 shows that the treatment effect of the Career arm on non-incentivized training test scores is insignificant and much smaller as compared to the incentivized dimension (0.025 versus 0.255). This confirms that the Career arm is indeed working as an incentive, and in fact teachers respond to the treatment strategically by exerting more effort on the incentivized dimension to acquire the incentive.[31]

[31]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

Column 6 in Table 4 also shows that the net impact of the PLUS arms on non-incentivized training scores is negative as on incentivized test scores, with the results particularly negative and significant for the Career PLUS arm. These results also provide further evidence that teachers who received the motivational framing reduced effort in the training across the board.[32]

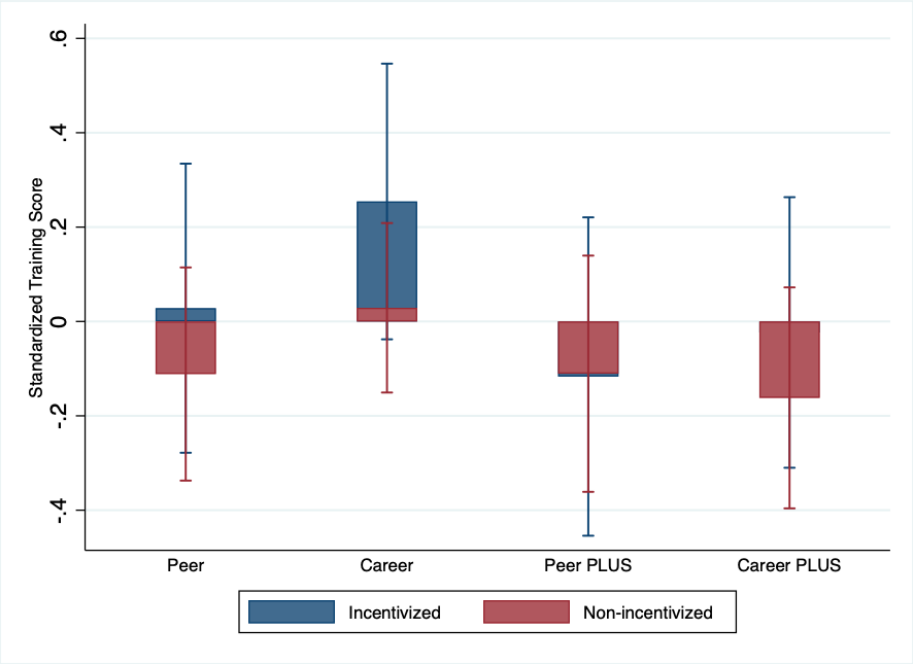


Figure 5: Treatment Effects on Non-Incentivized Dimension

more effort on the test as a result of the incentive.

[32]One might argue that the negative effects had something to do with the interaction effects between the incentive and the framing. This result rules out this argument

6.2 Mechanisms underlying treatment effects of the PLUS arms

Treatment effects on motivation index and beliefs about post-test performance. The main assumption behind the design of the PLUS treatments is that the motivational information should improve the framing of the recognition rewards by addressing potential negative effects on individual morale and beliefs about ability.

We investigate the impact of the PLUS treatments on different dimensions of teacher motivation such as intrinsic motivation (Deci and Ryan, 1985), self-efficacy (Bandura, 1986), and locus of control (Rotter, 1966). We measure these using pre-existing validated scales and normalize them by the mean and standard deviation of the control group. To avoid challenges of multiple hypothesis testing, we develop an overall index of teacher motivation as an average of these three measures. We also measure teacher beliefs about post-test performance to capture test specific teacher self-efficacy (i.e. teacher beliefs about their ability to perform well in the test) by asking teachers how much they believed they scored on the post-test on a scale of 1-100. To estimate our treatment effects on the motivation index and teacher beliefs about post-test performance, we run the same specification as (4) but with the teacher motivation index or beliefs about post-test performance as the outcome measure.

Panel B (column 1) in Table 5 shows that both the Peer PLUS and Career PLUS arms have positive and significant effects on the teacher motivation index, with an increase of 0.10σ and 0.12σ respectively. Column 2 in Table 5 shows that the net impact of the pooled PLUS arms on the motivation index is positive and significant, with an increase of 0.08σ. We observe that teachers receiving the motivational framing in the Peer PLUS arm report higher beliefs about post test performance (column 3) but do not find the same trend in the Career PLUS or the pooled PLUS arms.

These results confirm that while the motivational framing reduces training test scores, it does boost teacher motivation. Table A.7 shows these results by each dimension of teacher motivation - self-efficacy, locus of control, and intrinsic motivation - and confirms positive and significant effects of the PLUS arms on teacher self-efficacy, as well as positive coefficients on internal locus of control and intrinsic motivation.

6.3 Why did teachers reduce effort in PLUS arms?

In this subsection we investigate the mechanisms underlying the negative impact of the motivational framing on training test scores. We hypothesize that while the framing improved teacher motivation, 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 *over-correct* beliefs about ability leading to dangers of overconfidence (Swann, 1996; Baumeister, 1999; Bénabou and Tirole, 2002).

Treatment effects on 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. Typically, measures of overconfidence across Economics and Psychology are constructed by asking respondents a set of questions, along with their rate of confidence in the answers to each question. Overconfidence is then measured as the *positive bias*, when difference between average confidence level and the proportion of correct answers is greater than zero (Adams, 1957; Michailova, 2010). Since the fundamental idea in measuring overconfidence is to observe individual judgement and/or responses compared to a gold standard of truth (Baumann et al., 1991), our measure of overconfidence is an example of a direct measure of overconfidence and similar to measures used by others such as Glaser et al. (2005).[33]

To investigate treatment effects on overconfidence, we run the same specification as (4) but use teacher overconfidence as our outcome measure. Columns 1 and 2 in Table 6 show that the net impact of the motivational framing 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 non-PLUS arms and show that the net impact of adding the motivational framing is to make teachers 5.6% points more overconfident. Since our

[33]Typically studies try to measure innate overconfidence in individuals, and hence factors such as task complexity, subject's level of motivation, and skills of subject become important in being able to accurately asses the relationship between overconfidence and accuracy of judgement (Keasey and Watson, 1989). We are less concerned with these challenges given we are trying to assess the impact of overconfidence due to a specific intervention in a randomly assigned population. Randomization should ensure that individual level of motivation and skills are balanced between treatment and control groups

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.[34]

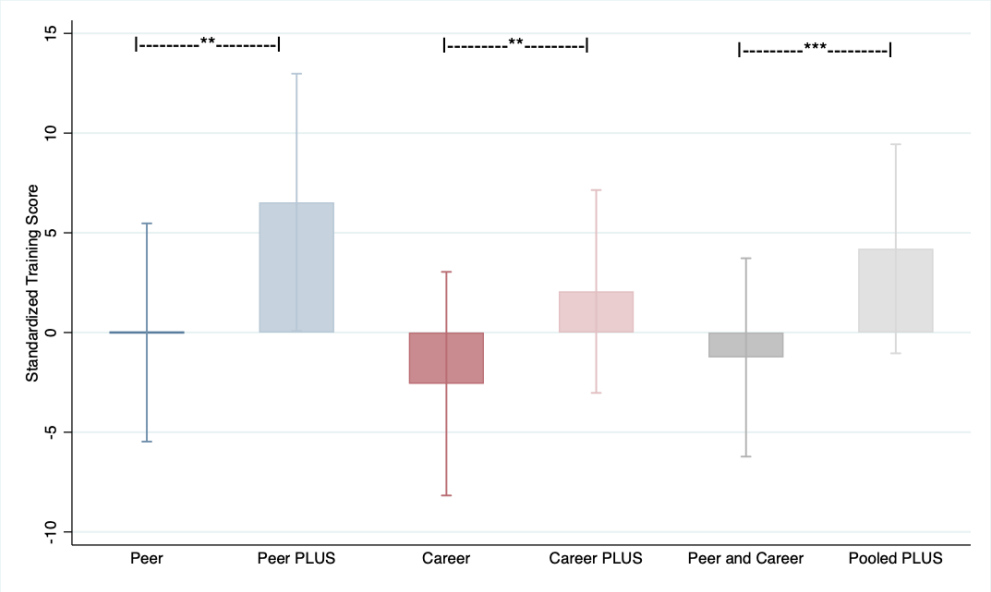


Figure 6: Treatment Effects on Overconfidence

Mediation Analysis for Overconfidence. We use mediation analysis to quantify the strength of the overconfidence channel in explaining the negative effects of the motivational framing.

[34] An alternate explanation could be that while the frames improved teacher self-efficacy, they simultaneously compromised their cognitive bandwidth (Mullainathan and Shafir, 2013). This could be because 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. 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.

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.[35] The identification of the 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.[36]

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)

[35] 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 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.

[36] 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.

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 7 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.

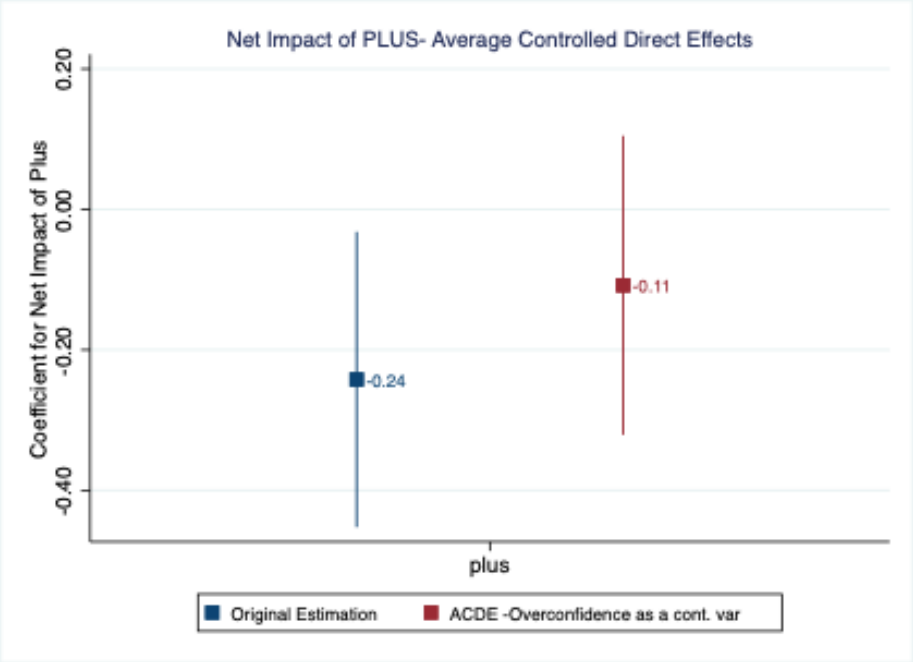


Figure 7: Average Controlled Direct Effect (De-mediated Test Scores)

7 Conclusion

We present experimental evidence on the impact of employer recognition on teacher training performance in mandatory government in-service trainings held by the Teacher Training Academy in 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 motivational framing focused on boosting morale and beliefs about ability 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 eliciting higher public school teacher effort more specifically, and public sector employee effort more broadly. The career-linked recognition incentive used in this experiment was fairly light touch, yet we find encouraging results which indicates the value of informal career benefits in the system. In our particular context, there are several 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, our results highlight the sensitivity of such incentives to framing effects and how they might interact with individual non-cognitive traits in nontrivial ways. This requires caution in how such incentives are designed across different contexts.

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, 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. Given implementation of high quality trainings to public sector employees and achieving their intended downstream effects is generally challenging (see Banerjee et al., 2016), future research on 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 would be useful. Finally, our experiment showed that creating exogenous variation in intrinsic motivation of public sector bureaucrats is possible through the use of our motivational framing. This opens up the possibility of additional research on how to create exogenous variation in intrinsic motivation and/or other non-cognitive traits in the workplace that can shape identities, norms, and culture.

Table 1: Treatment Effects on Training Performance (test scores at endline)

	(1)	(2)	(3)	(4)
Peer	0.014	0.038		
	(0.150)	(0.153)		
Career	0.234	0.251*		
	(0.151)	(0.148)		
Peer PLUS : Net Impact	-0.164	-0.163		
	(0.144)	(0.141)		
Career PLUS : Net Impact	-0.246*	-0.275**		
	(0.135)	(0.130)		
Peer and Career			0.125	0.145
			(0.133)	(0.133)
Pooled PLUS: Net Impact			-0.201*	-0.219**
			(0.103)	(0.100)
Peer PLUS*	-0.15	-0.12		
Career PLUS*	-0.01	-0.02		
PLUS*			-0.08	-0.07
Observations	3394	3392	3394	3392
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. 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	0.036	0.005	0.034	-0.070	-0.049	0.006	0.035	0.055	-0.056
	(0.299)	(0.213)	(0.224)	(0.161)	(0.146)	(0.131)	(0.102)	(0.075)	(0.076)
Career	0.299	0.367	0.334	0.274*	0.223	0.231*	0.235**	0.228***	0.068
	(0.373)	(0.249)	(0.243)	(0.165)	(0.136)	(0.120)	(0.102)	(0.073)	(0.064)
Peer PLUS : Net Impact	-0.164	-0.151	-0.129	-0.082	-0.067	-0.082	-0.107	-0.120*	-0.110*
	(0.235)	(0.193)	(0.205)	(0.161)	(0.151)	(0.112)	(0.095)	(0.071)	(0.066)
Career PLUS : Net Impact	-0.272	-0.303	-0.299	-0.285**	-0.256**	-0.259**	-0.212**	-0.187***	-0.143***
	(0.289)	(0.220)	(0.184)	(0.138)	(0.116)	(0.102)	(0.091)	(0.069)	(0.043)
Observations	3392	3392	3392	3392	3392	3392	3392	3392	3392
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)	(5)
Heterogeneous effects by:	Time till next promotion	Secretary visibility	Nature Contract	Peer known in class	Proportion peers known
Below Median (temp contract) x Peer	0.034	-0.029	-0.006	0.099	0.162
	(0.166)	(0.152)	(0.190)	(0.154)	(0.201)
Above Median (perm contract) x Peer	-0.041	0.243	0.032	-0.084	-0.197
	(0.160)	(0.216)	(0.152)	(0.168)	(0.209)
Below Median (temp contract) x Career	0.328**	0.206	0.021	0.214	0.099
	(0.162)	(0.157)	(0.175)	(0.160)	(0.216)
Above Median (perm contract) x Career	0.126	0.391**	0.279*	0.256	0.241
	(0.175)	(0.184)	(0.152)	(0.159)	(0.190)
Below Median (temp contract) x Peer PLUS	-0.225	-0.218	-0.191	-0.215	-0.015
	(0.187)	(0.166)	(0.189)	(0.172)	(0.237)
Above Median (perm contract) x Peer PLUS	-0.105	0.230	-0.135	-0.092	-0.338
	(0.165)	(0.214)	(0.171)	(0.185)	(0.215)
Below Median (temp contract) x Career PLUS	0.030	-0.016	-0.030	0.019	0.180
	(0.161)	(0.143)	(0.151)	(0.148)	(0.166)
Above Median (perm contract) x Career PLUS	-0.051	0.023	-0.013	-0.063	-0.300
	(0.156)	(0.207)	(0.150)	(0.160)	(0.237)
Observations	2181	3394	3394	3394	3394
PDS LASSO controls	No	No	No	No	No
District Fixed Effects	Yes	Yes	Yes	Yes	Yes

Notes: Errors clustered at the training session level which is the unit of randomization. All regressions include district FE. Each column represents heterogeneous treatment effects by a different moderator. Estimates are significant at the *10%, **5%, and ***1% level

Table 4: Treatment Effects on Non-Incentivised Training Scores

	Incentivised				Non Incentivised			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Peer	0.014	0.038			-0.095	-0.099		
	(0.150)	(0.153)			(0.106)	(0.108)		
Career	0.234	0.251*			0.045	0.025		
	(0.151)	(0.148)			(0.082)	(0.091)		
Peer PLUS : Net Impact	-0.164	-0.163			-0.017	-0.023		
	(0.144)	(0.141)			(0.125)	(0.116)		
Career PLUS : Net Impact	-0.246*	-0.275**			-0.171*	-0.193**		
	(0.135)	(0.130)			(0.095)	(0.092)		
Peer and Career			0.125	0.145			-0.023	-0.035
			(0.133)	(0.133)			(0.082)	(0.089)
Pooled PLUS: Net Impact			-0.201*	-0.219**			-0.095	-0.112
			(0.103)	(0.100)			(0.080)	(0.075)
Peer PLUS*	-0.15	-0.12			-0.11	-0.16		
Career PLUS*	-0.02	-0.03			-0.12	-0.17		
PLUS*			-0.08	-0.07			-0.11	-0.16
Observations	3394	3392	3394	3392	3394	3392	3394	3392
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 5: Treatment Effects on Motivation and Beliefs about Post Test Performance

	Motivation Index		Test Performance Beliefs (score out of 100)	
	(1)	(2)	(3)	(4)
Peer	0.051		-0.588	
	(0.034)		(1.102)	
Career	0.003		1.524	
	(0.034)		(1.220)	
Peer PLUS:Net Impact	0.042		2.370**	
	(0.037)		(1.007)	
Career PLUS: Net Impact	0.116***		-1.110	
	(0.039)		(1.396)	
Peer and Career		0.026		0.515
		(0.029)		(1.028)
Pooled PLUS: Net Impact		0.081***		0.539
		(0.029)		(0.961)
Peer PLUS*	0.10**		1.78	
Career PLUS*	0.12***		0.41	
PLUS*		0.11***		1.05
Observations	3373	3373	3072	3072
PDS LASSO controls	No	Yes	Yes	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 6: Treatment Effects on Overconfidence

	Overconfidence - continuous				Overconfidence - above median			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Peer	0.675	0.001			0.032	0.012		
	(2.824)	(2.765)			(0.067)	(0.062)		
Career	-2.211	-2.565			-0.065	-0.081		
	(2.898)	(2.834)			(0.072)	(0.066)		
Peer PLUS :								
Net Impact	6.170**	6.526**			0.124**	0.126**		
	(2.724)	(2.780)			(0.058)	(0.060)		
Career PLUS :								
Net Impact	5.308**	4.624**			0.115*	0.110*		
	(2.345)	(2.244)			(0.060)	(0.058)		
Peer and Career			-0.748	-1.073			-0.016	-0.034
			(2.578)	(2.647)			(0.063)	(0.057)
Pooled PLUS: Net Impact			5.629***	6.120***			0.117***	0.116**
			(1.878)	(2.034)			(0.044)	(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
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.

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.

Adams, J K. 1957. "A confidence scale defined in terms of expected percentages." *The American journal of psychology* 70 (3):432.

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.

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 et al. 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.

Baumann, Andrea O et al. 1991. "Overconfidence among physicians and nurses: The 'micro-certainty, macro-uncertainty' phenomenon." *Social science medicine (1982)* 32 (2):167–174.

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/`.

Connelly, Brian L et al. 2014. "Tournament Theory: Thirty Years of Contests and Competitions." *Journal of management* 40 (1):16–47.

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/`.

———. 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 L. and Richard M. Ryan. 1985. *Intrinsic Motivation and Self-Determination in Human Behavior [electronic resource].* Perspectives in social psychology. New York.

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.

Finan, Frederico et al. 2015. "The Personnel Economics of the State." *NBER Working Paper Series* :21825URL `http://search.proquest.com/docview/1752001802/`.

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/`.

Glaser, Markus et al. 2005. "Overconfidence of professionals and lay men : individual differences within and between tasks?"

Glewwe, Paul et al. 2010. "Teacher Incentives." *American Economic Journal: Applied Economics* 2 (3):205–227.

Gneezy, Uri et al. 2011. "When and Why Incentives (Don't) Work to Modify Behavior." *Journal of Economic Perspectives* 25 (4):191–210.

Goldsmith, Kelly and Ravi Dhar. 2013. "Negativity Bias and Task Motivation: Testing the Effectiveness of Positively Versus Negatively Framed Incentives." *Journal of experimental psychology. Applied* 19.

Hallinger, Philip and Ronald Heck. 1996. "Reassessing the Principal's Role in School Effectiveness: A Review of Empirical Research, 1980-1995." *Educational Administration Quarterly - EDUC ADMIN QUART* 32:5–44.

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`.

Keasey, Kevin and Robert Watson. 1989. "Consensus and accuracy in accounting studies of decision-making: A note on a new measure of consensus." *Accounting, organizations and society* 14 (4):337–345.

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, Susanne Neckermann, and Xiaolan Yang. 2014. "Knowing that You Matter, Matters! The Interplay of Meaning, Monetary Incentives, and Worker Recognition." *SSRN Electronic Journal* .

Lagarde, Mylène and Duane Blaauw. 2021. "Effects of incentive framing on performance and effort: evidence from a medically framed experiment." *Journal of the Economic Science Association* .

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." :22746URL `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.

McKenzie, David et al. 2021. "Aspirations and Financial Decisions: Experimental Evidence from the Philippines." .

Michailova, Julija. 2010. "Development of the overconfidence measurement instrument for the economic experiment." *IDEAS Working Paper Series from RePEc* URL `http://search.proquest.com/docview/1699082143/`.

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.

Rotter, J.B. 1966. "Generalized Expectancies for Internal versus External Control of Reinforcement." *Psychological Monographs* 33:300–303.

Silins, Halia and Bill Mulford. 2015. *Leadership and School Results*. 561–612.

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." .

Appendix

Appendix A: Tables

A. 1: Descriptive Statistics

	(1)	(2)	(3)	(4)	(5)
	Mean	Sd	p0.25	p0.50	p0.75
Basic teacher characteristics					
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
Basic job characteristics					
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				
Baseline Performance					
normalised					
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
Percent					
Pre Test Scores (Overall)	33.92	13.56	26.67	33.33	40.00
Non-cognitive traits					
Personality traits & Self-efficacy					
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
Motivational Orientation					
Extrinsic Motivation	0.25				
Intrinsic Motivation	0.41				
Pro-social Motivation	0.31				
Other intrinsic measures					
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 +	C-Peer	C-Career	C-Peer +	C-Career +	Peer- Career	Peer-Peer +	Peer- Car +	Career-Car +
Age	44.79	46.70	46.55	46.16	45.39	0.03**	0.03**	0.15	0.53	0.96	0.58	0.19	0.18
	(0.90)	(0.88)	(0.83)	(0.97)	(0.97)								
Gender (=1 if male)	0.42	0.57	0.50	0.56	0.45	0.03**	0.12	0.03**	0.67	0.26	0.94	0.05**	0.39
	(0.05)	(0.05)	(0.05)	(0.06)	(0.06)								
Salary	69009	74779	77471	74336	70669	0.07	0.02**	0.11	0.61	0.45	0.89	0.19	0.07
	(3578)	(3408)	(3670)	(3403)	(3403)								
Years of Education	15.66	15.72	15.73	15.72	15.73	0.30	0.17	0.29	0.19	0.74	0.95	0.74	0.99
	(0.06)	(0.06)	(0.06)	(0.06	(0.06)								
Married (=1 if married)	0.91	0.95	0.93	0.93	0.94	0.09	0.47	0.57	0.14	0.17	0.16	0.43	0.52
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)								
Basic job characteristics													
Time till next promotion (in yrs)	6.05	6.05	5.93	6.36	5.97	1.00	0.75	0.44	0.81	0.75	0.44	0.81	0.91
	(0.44)	(0.48)	(0.48)	(0.43)	(0.43)								
HT's school's enrollment capacity	237	267	330	256	246	0.32	0.04**	0.64	0.79	0.15	0.77	0.51	0.06
	(26.65)	(26.55)	(35.78)	(31.67)	(31.67)								
School Location of HT (=1 if urban)	0.11	0.15	0.17	0.07	0.08	0.35	0.19	0.30	0.39	0.68	0.08	0.11	0.07
	(0.03)	(0.04)	(0.04)	(0.03)	(0.03)								
Baseline Performance													
Pre Test (Incentivized)	-0.11	-0.13	-0.18	-0.28	-0.33	0.87	0.51	0.14	0.08	0.72	0.29	0.15	0.18
	(0.15)	(0.17)	(0.15)	(0.14)	(0.14)								
Pre Test (Overall)	-0.04	0.02	-0.05	-0.09	-0.14	0.70	0.84	0.64	0.33	0.58	0.45	0.24	0.40
	(0.13)	(0.15)	(0.13)	(0.12)	(0.14)								
Non-Cognitive Traits													
Overall BFI Index	0.09	0.05	0.06	0.08	0.03	0.21	0.34	0.69	0.07	0.78	0.46	0.57	0.39
	(0.04)	(0.03)	(0.04)	(0.04)	(0.04)								
Self-efficacy Index	-0.04	-0.03	-0.03	-0.04	-0.02	0.87	0.95	0.99	0.80	0.91	0.86	0.93	0.82
	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)								
Intrinsic Motivation	0.51	0.48	0.52	0.50	0.51	0.35	0.73	0.90	0.97	0.22	0.53	0.32	0.75
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)								
Extrinsic Motivation	0.20	0.21	0.17	0.20	0.21	0.61	0.26	0.80	0.51	0.06	0.80	0.88	0.06
	(0.03)	(0.03)	(0.03)	(0.03	(0.03)								
Pro-social Motivation	0.27	0.29	0.29	0.28	0.26	0.54	0.52	0.90	0.56	0.97	0.67	0.27	0.23
	(0.03)	(0.03)	(0.03)	(0.03)	0.03								
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) Attrited Sample	(2) Main Sample	(3) P-value difference
Basic teacher characteristics			
Age	47.04	45.85	0.40
	(1.51)	(0.69)	
Gender (=1 if male)	0.43	0.50	0.23
	(0.07)	(0.04)	
Salary	77039	73151	0.39
	(5158)	(2883)	
Years of Experience	21.08	20.10	0.52
	(1.69)	(0.90)	
Years of Education	15.75	15.72	0.75
	(0.11)	(0.05)	
Married (=1 if married)	0.92	0.93	0.66
	(0.03)	(0.04)	
Basic job characteristics			
Time till next promotion (in yrs)	4.64	6.10	0.04**
	(0.82)	(0.38)	
HT's school's enrollment capacity	237	267	0.32
	(26.65)	(26.55)	
School Location of HT (=1 if urban)	0.11	0.15	0.35
	(0.03)	(0.04)	
Baseline Performance			
Pre Test Scores (normalised)	-0.09	-0.20	0.39
	(0.18)	(0.14)	
Non-Cognitive Traits			
Overall BFI Index	0.01	0.06	0.50
	(0.07)	(0.03)	
Self-Efficacy Index	0.13	-0.03	0.15
	(0.12)	(0.06)	
Intrinsic Motivation	0.42	0.50	0.11
	(0.06)	(0.03)	
Extrinsic Motivation	0.23	0.19	0.55
	(0.06)	(0.02)	
Pro-social Motivation	0.34	0.28	0.28
	(0.06)	(0.02)	
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) Non-consenting Sample	(2) Consenting Sample	(3) P-value difference
Basic teacher characteristics			
Age	48.54	45.89	0.01***
	(1.15)	(0.69)	
Gender (=1 if male)	0.53	0.50	0.66
	(0.08)	(0.04)	
Salary	80358	73162	0.11
	(5034)	(2881)	
Years of Experience	23.16	20.12	0.00***
	(1.25)	(0.89)	
Years of Education	15.68	15.71	0.62
	(0.08)	(0.05)	
Married (=1 if married)	0.93	0.93	0.94
	(0.02)	(0.01)	
Basic job characteristics			
Time till next promotion (in yrs)	6.45	6.08	0.34
	(0.48)	(0.38)	
HT's school's enrollment capacity	237	267	0.32
	(26.65)	(26.55)	
School Location of HT (=1 if urban)	0.11	0.15	0.35
	(0.03)	(0.04)	
Baseline Performance			
Pre Test Scores (normalised)	-0.36	-0.21	0.05**
	(0.16)	(0.14)	
Non-Cognitive Traits			
Overall BFI Index	0.04	0.06	0.72
	(0.05)	(0.03)	
Self-Efficacy Index	-0.11	-0.02	0.26
	(0.09)	(0.06)	
Intrinsic Motivation	0.44	0.49	0.23
	(0.06)	(0.03)	
Extrinsic Motivation	0.20	0.21	0.72
	(0.04)	(0.03)	
Pro-social Motivation	0.33	0.27	0.08
	(0.04)	(0.02)	
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

A. 8: Motivation and Self Beliefs

	Self Efficacy	Internal Locus	Intrinsic Motivation
Peer	0.036	0.076	0.069
	(0.047)	(0.065)	(0.042)
Career	0.001	0.025	0.002
	(0.058)	(0.063)	(0.042)
Peer PLUS	0.115*	0.132*	0.046
	(0.061)	(0.072)	(0.046)
Career PLUS	0.132**	0.159**	0.094**
	(0.063)	(0.074)	(0.046)
Observations	3364	3364	3364

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. 9: 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. 10: Treatment Effects on Predicted Overconfidence

	(1)	(2)	(3)	(4)
Peer	0.794	-0.449		
	(0.694)	(0.387)		
Career	0.477	-0.213		
	(0.666)	(0.386)		
Peer PLUS : Net Impact of Frame	2.065***	1.503***		
	(0.660)	(0.368)		
Career PLUS : Net Impact of Frame	1.665**	0.920**		
	(0.669)	(0.386)		
Peer and Career			0.642	-0.090
			(0.588)	(0.486)
PLUS: Net Impact of Frame			1.844***	1.869***
			(0.492)	(0.395)
Observations	2963	2953	2963	2963
Adjusted R^2	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. 11: Mediation Analysis

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

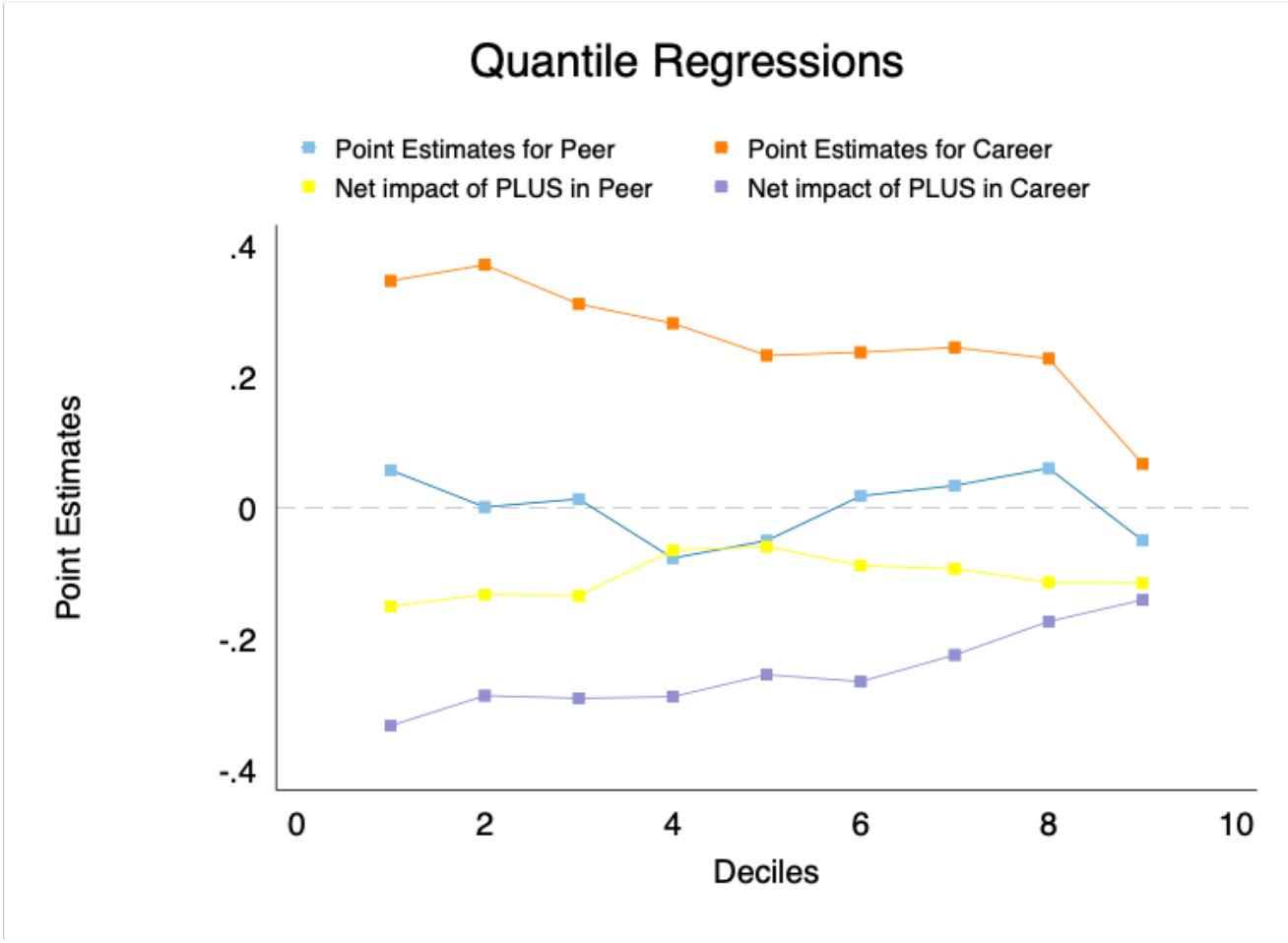
Notes: Errors are clustered at the training session level which is the unit of randomization. Both regressions are an ANCOVA estimation with the baseline value of the dependent variable and district FE. In column 1, the dependent variable is the post-test score. In column 2, the dependent variable is 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. 12: Heterogeneous Treatment Effects - Post Test Scores and Overconfidence

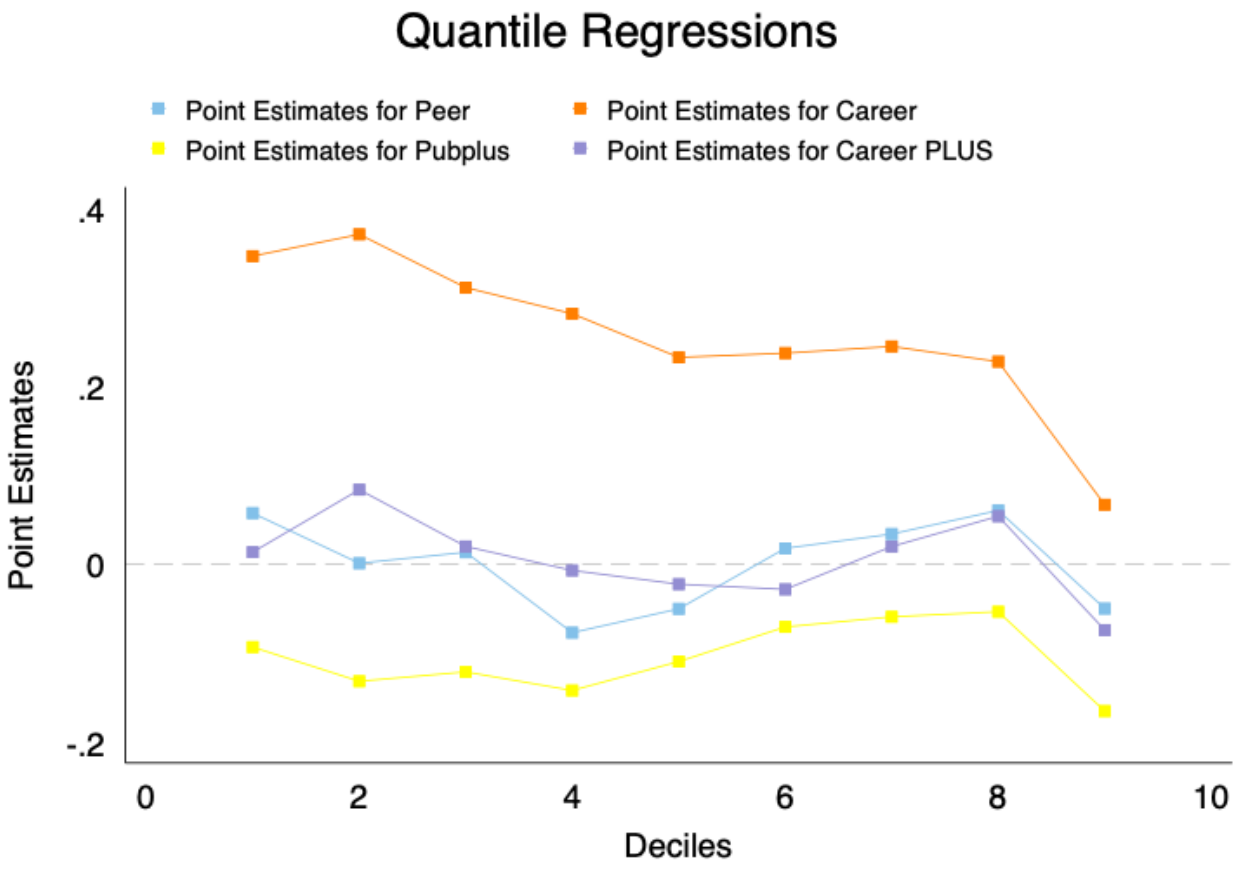
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Post Test	Overconf.	Post Test	Overconf.	Post Test	Overconf.	Post Test	Overconf.
Male x Peer + Career	0.056	2.115						
	(0.154)	(3.179)						
Female x Peer + Career	0.213	-3.447						
	(0.140)	(2.827)						
Male x Net PLUS	-0.302**	5.774**						
	(0.139)	(2.631)						
Female x Net PLUS	-0.141	4.855**						
	(0.099)	(2.011)						
Below Med. Ability x Peer + Career			0.215	-3.214				
			(0.173)	(3.064)				
Above Med. Ability x Peer + Career			0.013	1.775				
			(0.115)	(2.325)				
Below Med. Ability x PLUS			-0.226*	5.983***				
			(0.116)	(2.068)				
Above Med. Ability x PLUS			-0.175	4.059*				
			(0.110)	(2.072)				
Below Med. Experience x Peer + Career					0.070	0.225		
					(0.137)	(2.633)		
Above Med. Experience x Peer + Career					0.210	-2.626		
					(0.145)	(2.780)		
Below Med. Experience x PLUS					-0.164	3.735*		
					(0.111)	(1.972)		
Above Med. Experience x PLUS					-0.243**	6.494***		
					(0.112)	(2.124)		
Below Med. Personality x Peer + Career							0.099	0.426
							(0.140)	(2.872)
Above Med. Personality x Peer + Career							0.170	-2.278
							(0.137)	(2.365)
Below Med. Personality x PLUS							-0.216*	5.603***
							(0.116)	(2.089)
Above Med. Personality x PLUS							-0.198**	4.473**
							(0.100)	(1.882)
Observations	3394	3072	3394	3072	3394	3072	3382	3062
Adjusted R^2	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.

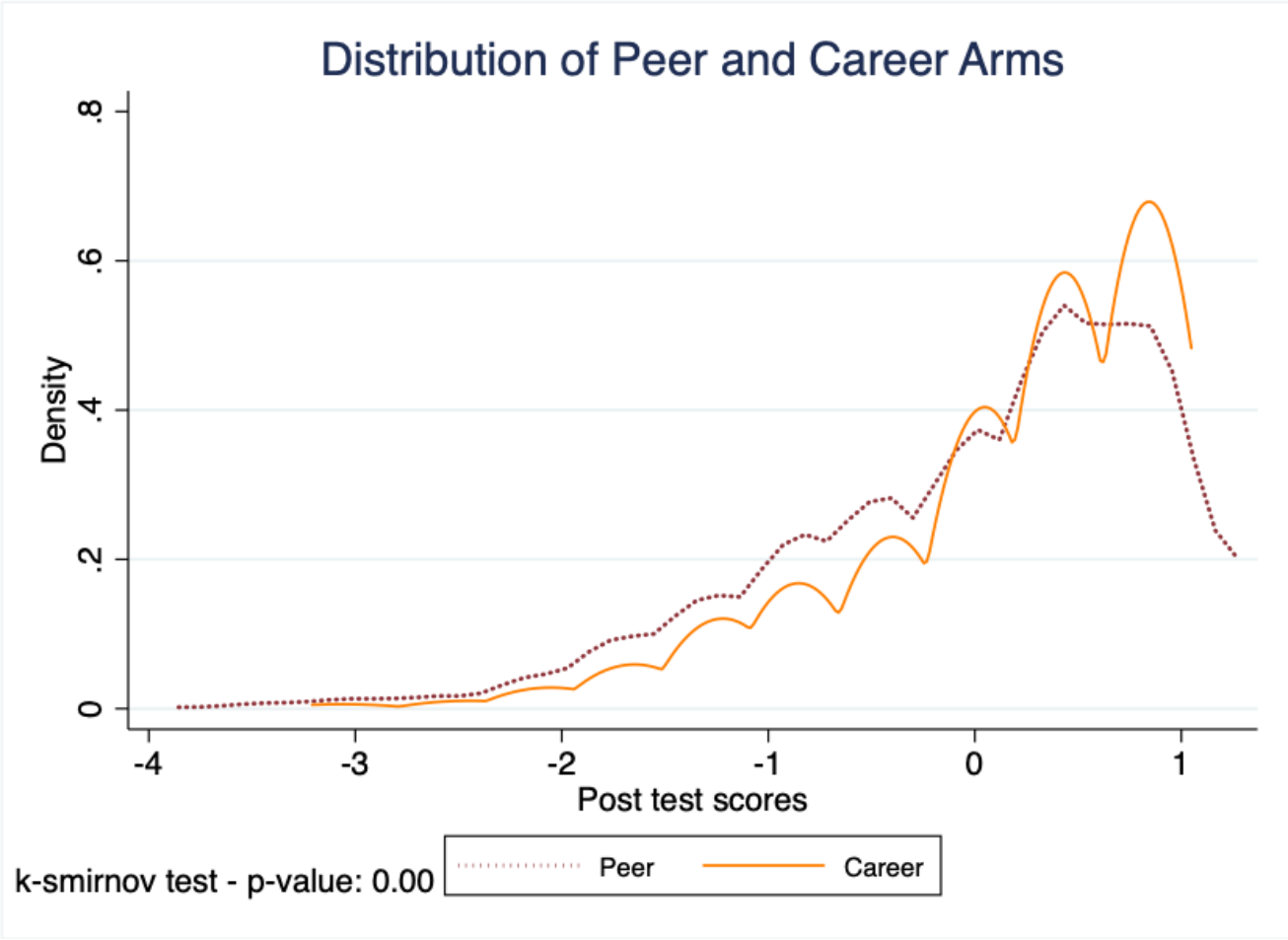
Appendix B: Figures



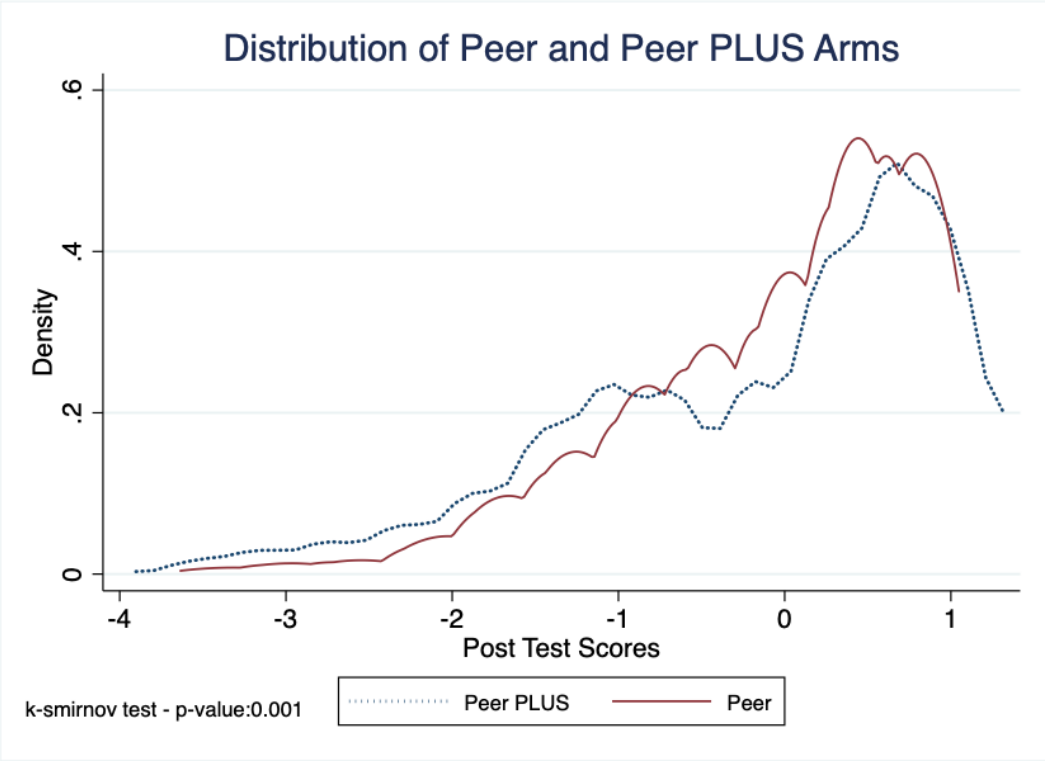
B. 1: Quantile Regression Estimates



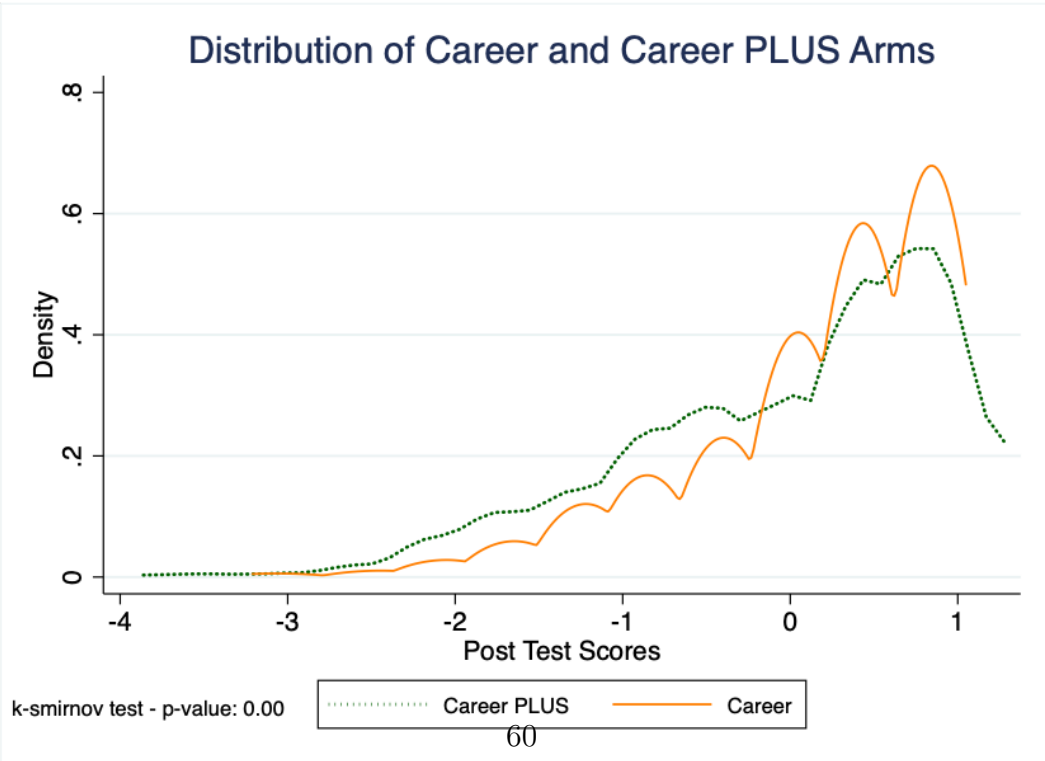
B. 2: Quantile Regression Estimates



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

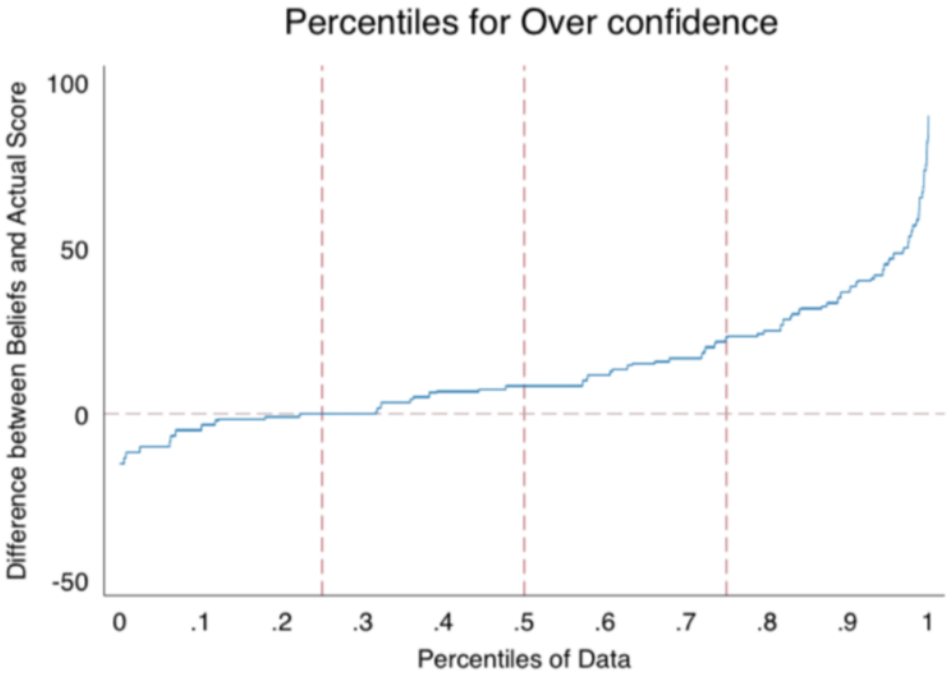


(a) Peer and Peer PLUS

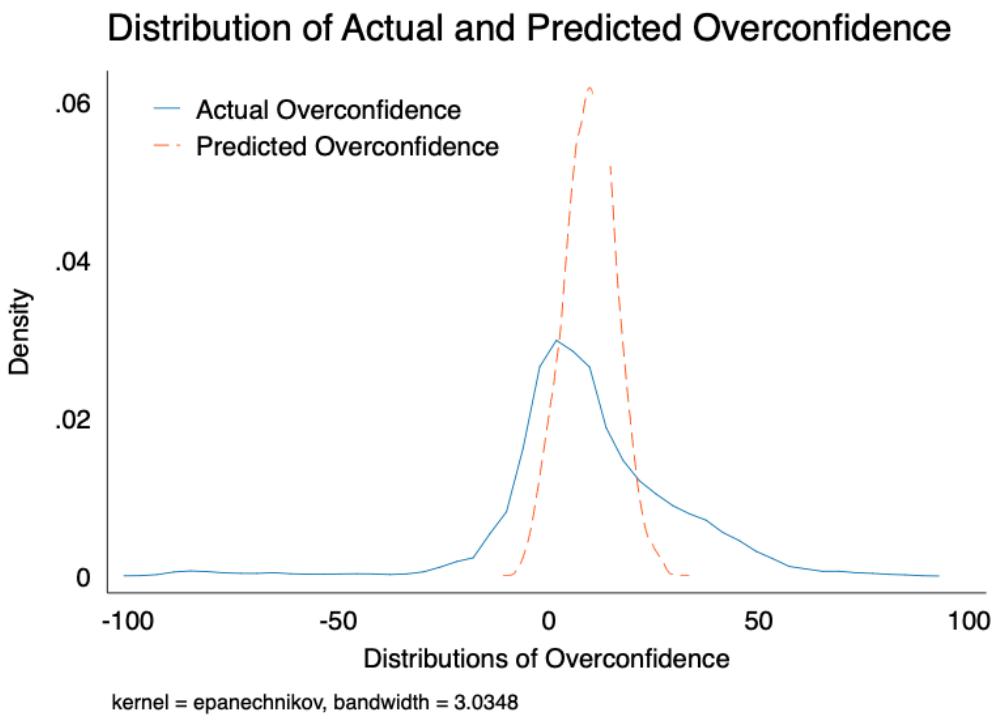


(b) Career and Career PLUS

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



B. 5: Qplot for Overconfidence



B. 6: Distribution of Actual and Predicted Overconfidence

Appendix C: Training Details

Training Topics

sr.no	Topic
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: ☐☐/☐☐/☐☐	School emiscode: ☐☐☐☐☐☐☐	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 b) Early Childhood Management c) Early Childhood Motivation 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

E. 2: Motivational Frame Hand-out

Appendix F: Snapshots of Training