Personalities and Public Sector Performance:

Evidence from a Health Experiment in Pakistan

Michael Callen[*], Saad Gulzar[†], Ali Hasanain[‡], Muhammad Yasir Khan[§], Arman Rezaee[¶]

June 4, 2024

Abstract

This paper presents evidence that selecting better people to work in government and improving their incentives are complements at improving government effectiveness. To do so, this paper combines a policy that improved incentives for health service delivery in Punjab, Pakistan, with data on health worker personalities. We present three key results. First, government doctors with higher personality scores perform better, even under status quo incentives. Second, health inspectors with higher personality scores exhibit larger treatment responses when incentives are reformed. Last, senior health officials with higher personality scores respond more to data on staff absence by compelling better subsequent attendance.

JEL codes: C93, D02, D73, O31, H1, HI1, HI18

Keywords: State capabilities, public sector absence, public health, personality psychology, Big Five personality traits, public sector personnel, bureaucracy, public service motivation, information communication technology, monitoring, information management

[*]London School of Economics. Email: m.j.callen@lse.ac.uk
[†]Princeton University. Email: gulzar@princeton.edu
[‡]Lahore University of Management Sciences. Email: hasanain@lums.edu.pk
[§]University of Pittsburgh. Email: myk17@pitt.edu
[¶]University of California, Davis. Email: abrezaee@ucdavis.edu

Authors' Note: We thank Farasat Iqbal (Punjab Health Sector Reforms Project) for championing and implementing the project and Asim Fayaz and Zubair Bhatti (World Bank) for designing the smartphone monitoring program. Support is provided by the International Growth Centre (IGC) state capabilities program, the IGC Pakistan country office, the University of California Office of the President Lab Fees Research Program Grant #235855, and grant #FA9550-09-1-0314 from the Air Force Office of Scientific Research. We thank Tahir Andrabi, Eli Berman, Ali Cheema, Julie Cullen, Clark Gibson, Naved Hamid, Gordon Hanson, Asim Khwaja, Jennifer Lerner, Jane Mansbridge, Edward Miguel, Craig McIntosh, Ijaz Nabi, Rohini Pande, Christopher Woodruff, and seminar participants at UC Berkeley, UC Los Angeles Anderson, UC San Diego, Paris School of Economics, New York University, University of Washington, Harvard Kennedy School, and participants at the IGC Political Economy Group, Development and Conflict Research (DACOR), Pacific Development (PacDev), New England Universities Development Consortium (NEUDC), Southern California Conference in Applied Microeconomics (SoCCAM), Bay Area Behavioral and Experimental Economics Workshop (BABEEW), Symposium on Economic Experiments in Developing Countries (SEEDEEC), and the Bureau for Research and Economic Analysis of Development (BREAD) conferences for insightful comments. Excellent research assistance was provided by Muhammad Zia Mehmood. The research reported in this paper was approved by the University of California San Diego Human Research and Protections Program (Project # 111442).

1 Introduction

Two fundamental means of raising government effectiveness are selecting better people to work in government and improving incentives (Hamilton and Jay, 1788; Besley, 2006). Accordingly, substantial bodies of research examine the benefits to government effectiveness of improving selection (Dal Bó et al., 2013; Klinger et al., 2013; Rasul and Rogger, 2018; Ashraf et al., 2020, 2014; Finan et al., 2015; Deserranno, 2016; Grossman and Slough, 2022) and of strengthening incentives (World Bank, 2004; Reinikka and Svensson, 2004; Chaudhury et al., 2006; Banerjee et al., 2008; Bandiera et al., 2009; Olken and Pande, 2012; Wild et al., 2012; Finan et al., 2015; Dhaliwal and Hanna, 2017; Aman-Rana, 2020). However, there is less evidence regarding whether selection and incentives are complements or substitutes, and, more generally, how they interact. Especially in resource-poor settings like Pakistan, understanding these interactions could carry valuable lessons for how to target resources.

This paper reports results from a field experiment in Pakistan designed to understand how the quality of government workers interacts with efforts to improve incentives.[1] Three key elements comprise the study. First, the government of Punjab introduced a province-wide monitoring effort for its health workers, the main impacts of which are reported by Callen et al. (2020).[2] Punjab has an estimated population of over 110 million, 90 percent of which rely on these government workers for their healthcare (National Institute of Population Studies, 2013). Rates of absence for this group are exceptionally high, even relative to those recorded in other low and middle income countries—two thirds of Punjab's 2,496 doctors serving rural areas were absent from work during random audits, for example. The monitoring intervention was aimed at addressing this.

Second, we collect data on the Big Five personality characteristics and Perry Public

[1] We did not pre-register this experiment following today's best practices. We began conducting the experiment in 2011 before pre-registration was common. There is clear evidence that we did pre-commit to conducting the analysis in this paper, however—we devoted substantial resources tracking down and measuring personality traits through our surveys with health workers.

[2] This paper departs from Callen et al. (2020) in several ways. First and foremost, the focus of this paper is descriptive rather than experimental. Second, the empirical specifications and sample used for analysis is varied in several ways. We report robustness to these choices when relevant.

Sector Motivation (PSM) of all of the workers affected by the reform, including both frontline workers like these doctors and very senior bureaucrats in the Health Department. The Big Five and PSM measures were developed by psychologists in the 1980s and remain two of the most widely used measures in personality psychology (John et al., 2008; Borghans et al., 2008; Perry and Wise, 1990; Perry, 1996; Petrovsky, 2009). Despite their very high rates of absence, we managed to track down and survey a representative sample of 389 doctors across Punjab. We also surveyed the universe of health inspectors (who are above doctors in the chain of command and are most directly targeted by the monitoring intervention) and senior health officials (the senior-most health bureaucrats in each district), that is, 102 inspectors and 33 senior health officials. This second element allows us to study heterogeneity in who reacts to improving incentives, yielding insights about the interaction between changes to incentives and the stock of government employees. As the characteristics of this stock of employees is determined by selection, these insights allow us to investigate the interaction between selection and incentives.

Third, the monitoring system introduced in Punjab funneled information on performance to senior health officials through an online dashboard. This third element allows us to extend our notion of performance and our associated focus on selection and incentives to very senior bureaucrats. Concretely, we can study whether personality characteristics predict who among the senior management cadre react when they are informed that their subordinates are absent.

The exercise yields three key results. First, personality and motivation measures correlate with several measures of performance, ranging from simple attendance to efforts to undermine the reform. This is true for both frontline workers, where doctors who exhibit normatively better personality traits and more motivation, for example those who are more conscientious, are absent less often, and for middle-managers in the health bureaucracy, where inspectors who exhibit these same traits are found to collude less to falsify reports.

Second, workers with better personality traits and more motivation also respond more to

treatment reform. This positive interaction suggests the possibility that improving selection and incentives in tandem can drive larger performance improvements than reforms that target either margin individually.[3] Importantly, these two results validate the idea the Big Five personality and Public Sector motivation measures in studies of selection such as Dal Bó et al. (2013).

Third, senior managers with these same qualities exhibit larger responses to information (i.e. they perform better on one important duty—ensuring that subordinate doctors show up to work). This links our exercise to papers focused on whether and how policy makers use data in policy formulation. Personality traits are useful in identifying which policy makers will react to data. The push to encourage governments to adopt policies based on evidence is focusing the attention of researchers on whether and how policy actors assimilate this knowledge (DellaVigna et al., 2019; Hjort et al., 2021).

In addition to speaking directly to, and between, the literatures on selection and on incentives for public servants in developing countries, this paper contributes to an active literature examining the role of non-cognitive traits in performance in developed contexts, including both individuals and firms in the United States (Borghans et al., 2008; Almlund et al., 2011; Heckman, 2011). This paper is also in agreement with psychology literature that documents these measured personality traits are more than situational specific, and thus are worthwhile to use for explanatory purposes as we do in this paper (Roberts, 2009).

While our data allow us to relate personalities to performance, they also face some limitations. First, and perhaps most importantly, piloting revealed that our respondents could react negatively to exercises that measure cognitive ability, such as Raven's matrices, or their innate honesty. We therefore are unable to directly compare the relevance of cognitive and non-cognitive attributes, as well as honesty, for service delivery. Second, no component of the personality traits we measure is easy to manipulate experimentally, limiting our ability to identify the causal relationship between personalities and performance (Deaton, 2010).

[3]For specific conditions under which this would be the case, see our Framework in Section 2.2.

To address this, in our information experiment with senior officials, we aimed to manipulate a factor affecting performance—information about the performance of their subordinates—that could plausibly mediated through the mechanism of personalities. That is to say, for information to have an effect it must first be taken in by senior officials and they then must process it in order to make a decision to act on it (the information itself did not require an action). Personality measures such as conscientiousness or civic duty could plausibly affect either how information is taken in (i.e. how careful the senior official is when looking at the dashboard) and/or how the information is processed (i.e. whether senior official see it as their duty to act on bad performance).

The paper proceeds as follows: Section 2 provides the institutional details necessary to understand our results. Section 2.2 provides a framework. Section 3 outlines our research design and reports results. Section 4 concludes.

2 Public Health Services in Punjab

This section describes the main institutional details relevant to our experiment and our empirical results.

In Punjab, the provision of health care services is managed by the Department of Health. Authority in the department is highly centralized in the upper echelons of the bureaucratic hierarchy. Senior actors described in this section play a central role in determining the quality of delivery. They are also responsible for a substantial number of facilities spread, in many cases, across vast geographic distances. This presents a major challenge for monitoring that we aim to address with our smartphone monitoring system.

The main performance outcomes in this paper are measured at primary front-line public health clinics, called Basic Health Units (BHUs). BHUs are designed to be the first stop for rural patients seeking medical treatment in government facilities, providing mainly primary services, including out-patient services, neo-natal and reproductive health care, and

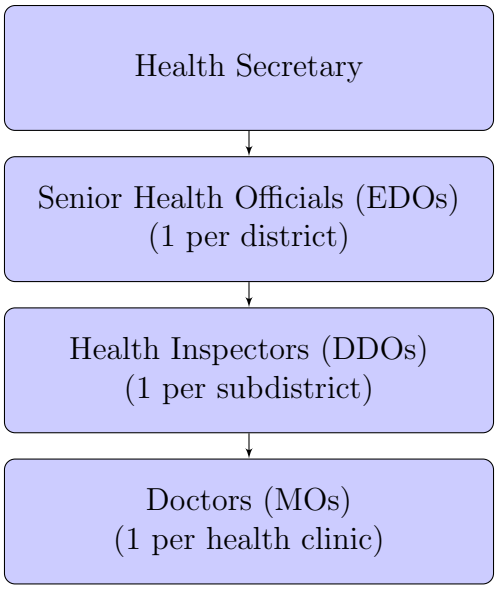

Figure 1: Health Sector Administration in Punjab

vaccinations against diseases. Hereafter in this paper, we use the word 'clinic' interchangeably to describe BHUs. There are 2,496 BHUs in Punjab. Each Basic Health Unit serves approximately one Union Council (Union Councils are smallest administrative units in Pakistan). Almost all BHUs are located in rural and peri-urban areas. Each facility is headed by a doctor, known as the Medical Officer, who is supported by a Dispenser, a Lady Health Visitor, a School Health and Nutrition Supervisor, a Health/Medical Technician, a Mid-wife and other ancillary staff. Officially, clinics are open, and all staff are supposed to be present, from 8AM to 2PM and patients seen in these clinics are required to pay a nominal fee of around $0.01 USD per visit.

2.1 Health Sector Administration

Figure 1 depicts a simplified version of the health administration hierarchy in Punjab. District governments are responsible for managing local health facilities. Each District Department of Health is headed by an Executive District Officer (EDO) who reports both to the official in charge of the district (District Coordination Officer) and to two provincial health

officials (Secretary of Health and Director General of Health Services). EDOs are directly supported by several Deputy District Officers (DDOs). DDOs primarily inspect and manage health facilities in their area of jurisdiction, a *Tehsil*, the largest administrative unit within a district.[4] DDOs are required to inspect every clinic in their jurisdiction at least once a month and record information collected during the visit on a standard form. DDOs have the authority to punish a clinic's absent staff by issuing a formal reprimand, suspending staff, and/or withholding pay (in the case of contract staff). Each Medical Officer is similarly responsible for their own clinic, with proportional duties. Throughout the paper, we will refer to Executive District Officers as senior health officials, Deputy District Officers as inspectors, and Medical Officers as doctors, focusing on their role rather than their title.

As is true in many developing countries, low health worker attendance is a major issue in Punjab. From unannounced visits to clinics in 2011, we find that only 56 percent of clinics were inspected in the prior two months, and that doctors were only present 43 percent of the time when one was posted. Doctors were not posted at 35 percent of clinics, which means unconditional doctor presence was only 32 percent. This points to a lack of enforcement that allows health inspectors and doctors to shirk.

2.2 Framework

In this section, we provide a framework to help us understand the primary three research questions considered in this paper—do personality measures (i) predict performance under status quo incentives, (ii) predict responses to a reform that increases the probability of detecting shirking, and (iii) predict responses to information on the performance of subordinates?[5]

We assume that the personality score of a health worker is part of their utility function,

[4]While inspections are the primary official functions of the DDO, our time use data indicate that, on average, DDOs spend 38.9 percent of their time on inspections and management, with the remainder of their time principally spent managing immunization drives. For full details please see Callen et al. (2020).

[5]A number of papers incorporate personality traits into standard economic models such as the Roy Model (Almlund et al., 2011) or the principle-agent framework (Besley and Ghatak, 2005; Benabou and Tirole, 2003).

such that a higher personality score increases the benefit of exerting effort at work. Better personality types draw more utility from performing their duties, either through intrinsic motivation if the task results in a prosocial outcome, or through extrinsic motivation such as social image. At the same time, effort at work is costly. And given that effort is hard to observe directly (i.e. the probability of detecting shirking is low), wages do not depend on this effort. This set of assumptions suggests that, in the status quo, personality will positively predict effort. Following this logic, if we were to change the incentives of health workers such that shirking is more likely to be detected, (i) less workers will shirk, and (ii) those workers who switch from shirking to not will be of higher personality type from among the pool of previously shirking workers.

Applying this framework to our specific context, it suggests a positive correlation between personality measures and performance for doctors under status quo incentives. For health inspectors the status quo implications are more subtle. If inspectors believe that their inspections cannot improve a prosocial outcome (quality health service delivery) because no one will read their reports and/or because doctors will be absent regardless, then better personality types may actually conduct less inspections and instead spend their time on other activities that may improve a prosocial outcome.

This framework does, however, suggest a positive correlation between personality measures and health inspectors' response to an increased detection of shirking via our monitoring experiment, since this will be holding constant the prosocial benefit of their inspections. As for senior health officials, once doctor and health inspector performance data is available on the web dashboard, those with high personality should be more likely to exert effort to act on this data.

More generally, this framework suggests a specific, non-linear relationship between selection and incentives. Health workers with the worst personality traits will never respond to incentives while those with the best personality traits will not need incentives as they will always comply, so it is specifically those in the middle of the distribution for which incen-

tives can be effective. This is supported by our non-parametric correlations presented below in Figure 7. It follows that for reforms that target selection and incentives in tandem to drive larger performance improvements than reforms that target either margin individually it must be that the reform of selection is focused on replacing the worst types with those in the middle of the distribution.

3 Results

In this section, we present three sets of results. First, we study correlations between the measured personality traits of doctors and health inspectors, their job performance (attendance and inspections respectively), and their propensity to collude with one another. Second, we use these measures to predict health inspectors' response to an experimental intervention which increases the probability of detecting shirking. Finally, we examine whether traits identify which senior health officials react to information about the absence of their subordinates. This analysis relies on manipulating the information provided to senior officials about the absence of their subordinates.

3.1 Do personality measures predict performance under status quo incentives?

We first examine whether personality measures predict bureaucratic performance under status quo incentives, for doctors and then for health inspectors. We measured personality for doctors in Punjab posted to a representative sample of 850 of the 2,496 rural health clinics in the province. Of the 850 facilities in this sample, 306 facilities had no doctor posted. We omit these clinics from our analysis of doctor performance. To reach the remaining doctors, we interviewed doctors in two unannounced independent inspections, and then followed up with pre-scheduled interviews. Doctors were strongly encouraged to attend the pre-scheduled interviews by the Department of Health. This process resulted in interviews of 389 out of

544 posted doctors, or 72 percent of our sample population.

We recognize that these doctors may be potentially unrepresentative of the overall sample of posted doctors. However, we believe that this select sample is highly relevant for two reasons. First, there are very likely a number of ghost workers—names on government payrolls that do not correspond to an actual person, allowing other corrupt actors to capture their salary. In this setting, there is no way for us to know how many of the doctors we did not reach actually exist. Given the substantial lengths we went to, including involving the active collaboration of the Department of Health in scheduling interviews, it is possible that many of them are indeed ghost workers and so are not part of the relevant sample of interest. Second, our pre-scheduled interviews were facilitated by doctors' supervisors via multiple phone calls and clear orders. If a doctor is not at work when we visit twice independently and refuses direct orders from their superior, clearly the doctor is underperforming. We are less interested in understanding how the individual characteristics of such intractably resistant individuals relate to performance.

We also measured personality for the universe of health inspectors and senior health officials in Punjab, or a total of 102 inspectors and 33 senior health officials. We interviewed inspectors and officials through pre-arranged office visits.

For all 850 clinics in our sample, we also measured attendance during unannounced visits in November 2011, June 2012, and October 2012.

3.1.1 Measuring Personality

The personality measurement batteries in this paper are from personality psychology and are used broadly, including recently in economics. We use two measures: the Big Five personality traits and the Perry Public Service Motivation (PSM) traits.

Developed by psychologists in the 1980s, the Five Factor Model is now one of the most widely used personality taxonomies in the field.[6] We measure the Big Five traits using a

[6]See John et al. (2008) for a summary of the measure and its history. Borghans et al. (2008) provide a summary of empirical results in psychology and economics. Additionally, see (Johnson et al., 1985; Barrick

60 question survey developed specifically in Urdu and validated for use in Pakistan by the National Institute of Psychology at Quaid-i-Azam University, Islamabad. Each of the 60 questions offers the respondent a statement such as "I see myself as someone who does a thorough job" and asks them to agree or disagree with the statement on a five point Likert scale (Disagree strongly, Disagree a little, Neutral, Agree a little, or Agree strongly).[7]

In addition to measuring Big Five traits separately as the mean response to twelve questions (where disagree strongly is assigned a 1, disagree a little a 2, etc.), all traits are normalized into z-scores and averaged to form a single Big Five index. This approach is consistent with research in psychology that finds high degrees of correlation between the five personality traits in many different studies and suggests that the traits can be collapsed into a General Factor of Personality, which can be interpreted "as a basic personality disposition that integrates the most general non-cognitive dimensions of personality. It is associated with social desirability, emotionality, motivation, well-being, satisfaction with life, and self-esteem. It also may have deep biological roots, evolutionary, genetic, and neurophysiological" Musek (2007, pg. 1213).[8] We also document a high degree of correlation between Big Five traits in four different populations in Pakistan in Appendix Figure A.1.

The Perry Public Service Motivation (PSM) battery is designed to measure intrinsic motivation for public service. Also developed in the 1980s, it comprises a total of 40 questions measuring six traits—attraction to policymaking, commitment to policymaking, social justice, civic duty, compassion, and self-sacrifice (Perry and Wise, 1990; Perry, 1996; Petrovsky, 2009). We reproduce both the Big Five and PSM batteries we used to interview the doctors in the appendix. We used the same instrument for heath inspectors and senior health officials.

Table 1 reports summary statistics for these measures separately for doctors and health

and Mount, 1991; Kaplan and Saccuzzo, 1997; Salgado, 1997; Schmidt and Hunter, 1998; Bowles et al., 2001; Bertrand and Schoar, 2003; Hogan and Holland, 2003; Nyhus and Pons, 2005; Heckman et al., 2006; Groth-Marnat, 2009; Gatewood et al., 2010; Bazerman and Moore, 2012; Nyhus and Pons, 2005).

[7]John et al. (2008) provide the mapping between questions and traits.

[8]See Digman (1997) and Van der Linden et al. (2010) for two additional meta-analyses with similar results.

inspectors in treatment and control districts in our randomized control evaluation of a new monitoring technology. There is substantial variation in personality traits across individuals consistent with the original intention of the battery: to capture substantial and important differences in personality types. It is this heterogeneity that allows for the possibility of linking differences in personality to variation in performance. The full distributions for these measures are reported in Table A.1. Summary statistics for senior health officials are reported in Table A.2.

We capture these measures after treatment is administered. This raises the possibility that treatment could impact traits, confounding our analysis. However, if treatment impacted traits then there would be differences between treatment and control workers in personality measures. We find no evidence that treatment affected personality traits. This increases our confidence that they are stable over the horizon of the study. This is consistent with previously cited literature that suggests personality traits are stable over the years (Cobb-Clark and Schurer, 2012), and malleability only arises over the course of years, not months (Roberts et al., 2006), or given intense cognitive-behavioral therapy (Kautz et al., 2014; Blattman et al., 2015).

3.1.2 Measuring Doctor Performance

To obtain measures of performance, we collected primary data on a representative sample of 850 of the 2,496 clinics or Basic Health Units in Punjab. Clinics were selected randomly using an Equal Probability of Selection design, stratified on district and distance between the district headquarters and the clinic. Our estimates of absence are, therefore, self-weighting and require no sampling correction. All districts in Punjab except Khanewal—the technology pilot district—are represented in our data. Figure 2 provides a map of clinics in our experimental sample along with the district boundaries in Punjab.

Information on staff absence, health inspections, and facility usage was collected through three independent and unannounced visits of these clinics. These visits were done by our

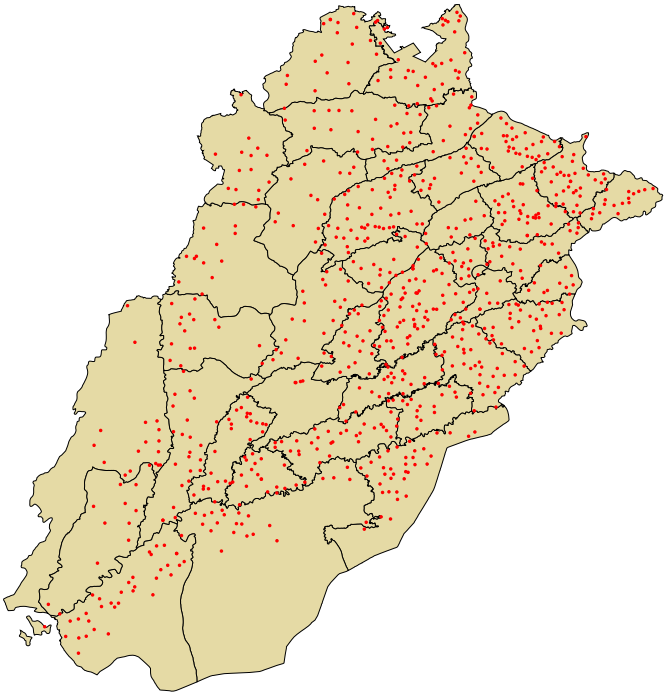

Figure 2: Locations of Clinics (Basic Health Units) in the Experimental Sample

survey teams hired and trained at regional hubs. Our teams visited each facility three times: November 2011, June 2012, and October 2012. Our survey team interviewed and physically verified the presence of the Medical Officer, or doctor. In addition, the attendance of Dispensers, Health/Medical Technicians, Lady Health Visitors, Midwives, and School Health and Nutrition Specialists were also recorded. The attendance sheet for the staff was filled out at the end of the interviews and in private. Inspectors record inspection visits by signing paper registers maintained at the health facility. We measure whether an inspection occurred by interviewing facility staff and verifying the register record.

We have two measures of doctor job performance: (i) whether doctors were present during our unannounced visits, and (ii) a proxy measure of collusion between doctors and health inspectors to falsify reports. We define collusion as a dummy variable coded as one when the doctor is absent during both of our post-treatment unannounced visits and is marked present during every single health inspection during the treatment period. The median number of health inspections for each facility in our treatment sample is 12, with a max of 50. The

collusion we have in mind occurs when a health inspector calls a doctor before an inspection to alert him to be in attendance. Then, after the health inspector records his presence, the doctor is under very little pressure to attend until he gets another similar phone call from the inspector.[9] We believe this is a relevant and distinct measure of performance that amounts to falsifying data. Collusion of this form will lead official health inspections to show doctors as being present 100% of the time when in reality they are present much less. As we will show in Section 3.3.2 below, senior health officials look at and respond to this information. Collusion of this form therefore directly limits the ability of higher bureaucrats to respond to and improve health performance while benefiting shirking doctors.

We find doctors to be present during forty three percent of the unannounced visits and predict collusion with health inspectors thirteen percent of the time. These baseline performance measures for doctors are reported in Table A.1.

3.1.3 Personality and Doctor Performance

Figure 3, Panel A shows that doctors that score one standard deviation above the mean on the Big Five measure of conscientiousness are about five percentage points more likely to be present at work during an unannounced visit. Similarly, self-sacrifice, a PSM measure, is also significantly predictive, and the aggregate PSM index is nearly significantly predictive at 95%. Finally, all but one coefficient are positively correlated with doctor attendance. In Panel B, we find that doctor personality measures are even stronger predictors of collusion between health inspectors and doctors. Doctors who score one standard deviation higher on measured civic duty, for example, are about 6 percentage points more likely to be identified as potentially colluding. Both the Big Five and PSM indices and ten out of eleven Big Five

[9] Of course, such patterns in the data could arise by chance, though the chance decreases with the number of inspections. As such, we have run all of our collusion analysis using weighted least squares and we find results very similar to those OLS results presented below. Results provided upon request. The strong correlation we find between these measures and personality types also suggests that the proxy is successfully capturing malfeasance. An immediate problem with this proxy is that it partly reflects attendance. We deal with this by also reporting p-values adjusted to reflect multiple hypotheses.

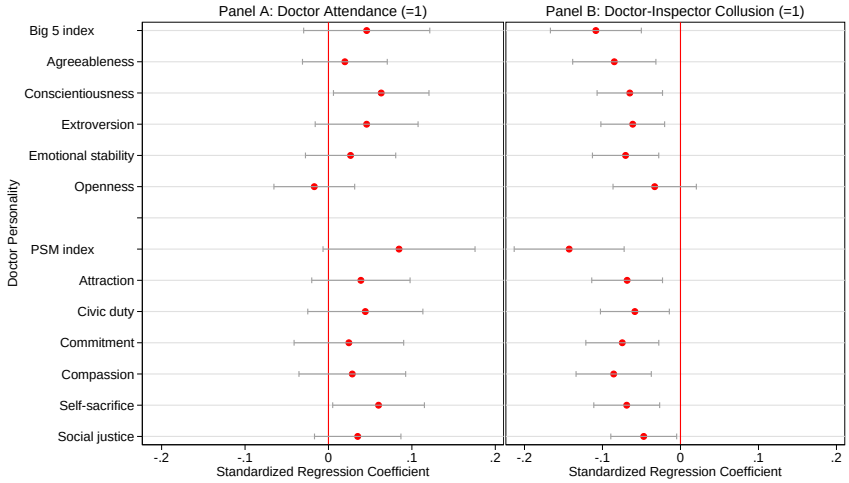

Figure 3: Personality and Performance: Doctors

Notes: Each regression coefficient reported comes from a separate regression of the performance measure, Doctor Attendance in Panel A and Doctor-Inspector Collusion in Panel B, on the indicated doctor personality measure. Error bars represent 95 percent confidence intervals, with standard errors clustered at the clinic level. All regressions include tehsil (sub-district) and survey wave fixed effects. In all cases, personality measures are normalized to have mean zero and standard deviation of one in the sample, and thus the regression coefficients reported can be interpreted as the impact of a one standard deviation increase in a given personality trait or aggregate measure. The sample for Panel A is restricted to control district clinics for which doctor personality data are available and a doctor is posted (479 observations across 190 doctors). The sample for Panel B is restricted to treatment district clinics for which doctor personality data are available and a doctor is posted (273 observations across 273 doctors). Regressions corresponding to the figure are reported in Appendix Tables A.3 and A.4.

and PSM traits are highly predictive of collusion, with negative signs.[10]

We draw three lessons from this exercise. First, in Appendix Table A.5, we find that personality is a stronger predictor for doctors than three other plausibly important observables—doctor tenure in the department of health, doctor tenure at the specific health clinic at which the doctor worked at the time of the survey, and the distance from this clinic to the doctor's home in Pakistan (in KM). Though we have only a limited number of covariates for this exercise, they are potentially correlated with a wide number of factors relevant to the relationship between personality and performance. Overall tenure, for example, will be correlated with age, experience, and the number of relationships with others in the health department. Tenure at a specific facility will be correlated with how much influence a doctor has in the Department of Health as transfers are frequent and often undesirable. Distance to home might proxy for the desirability of a posting as in interviews doctors frequently expressed a strong desire to work near their home and family.

Second, the degree of the estimated coefficients is meaningful. While ideally we would have measures of health outcomes to correlate with doctor performance, we are able to correlate this performance with the number of out-patients seen at a clinic in a given month. We document a strong positive correlation between doctor presence at their clinic during one of our unannounced visits and reported out-patients seen at that clinic in Appendix Table A.6. It is worth noting, as well, that the confidence intervals on our estimates are fairly narrow. For conscientiousness and doctor attendance, for example, the 95 percent confidence interval lies between 0.01 and 0.11, and so we can reject both negative and large positive effects. The confidence intervals are similarly narrow for our collusion correlations.

Third, importantly, not every personality measure is significantly predictive of performance for doctors. While conscientiousness is significantly predictive, agreeableness is not. Five of the six PSM measures are not significantly predictive as well. While it is tempting to read into these differences as pointing towards which personality characteristics are more

[10]See Appendix Tables A.3 and A.4 for point-estimates.

important in this setting, because we did not randomize doctors' personalities, we refrain from such speculation. Given we find some significant and some insignificant correlations, however, we do take very seriously that our significant correlations could simply be due to chance as we are testing multiple hypotheses. In section 3.4 we adjust the p-values in Figure 3, as well as those from all of our following analysis, for multiple hypothesis testing and discuss the broad patterns we find in our data.

3.1.4 Monitoring Intervention

We collected personality data during a larger experimental policy reform that considered audits by government monitors as a solution to the problem of bureaucratic absence. The "Monitoring the Monitors" program replaced the traditional paper-based monitoring system for clinic utilization, resource availability, and worker absence with an android-based smartphone application. In the new system, data generated by health inspections are transmitted to a central database using General Packet Radio Service (GPRS). Data are then aggregated and summary statistics, charts, and graphs are presented in a format designed in collaboration with senior health officials to effectively communicate information on health facility performance. These data are also: (i) geo-tagged, time-stamped, and complemented with facility staff photos to check for reliability; and (ii) available in real time to district and provincial officials through an online dashboard. The objective of this monitoring system is to make the activities of health inspectors available to their senior officials in real time. Figure 4 shows one view of the online dashboard.[11]

We can think of this monitoring system as increasing the probability that a health inspector will be caught if he is failing to do his inspections. Prior to Monitoring the Monitors, and in control districts, the paper-based monitoring system severely limits a senior officials ability to monitor inspectors. In treatment districts, on the other hand, reports are imme-

[11] Application development started in August 2011. After developing the application and linking it to a beta version of the online dashboard, the system was piloted in the district of Khanewal. We remove Khanewal district from the experimental sample. Health administration staff were provided with smartphones and trained to use the application.

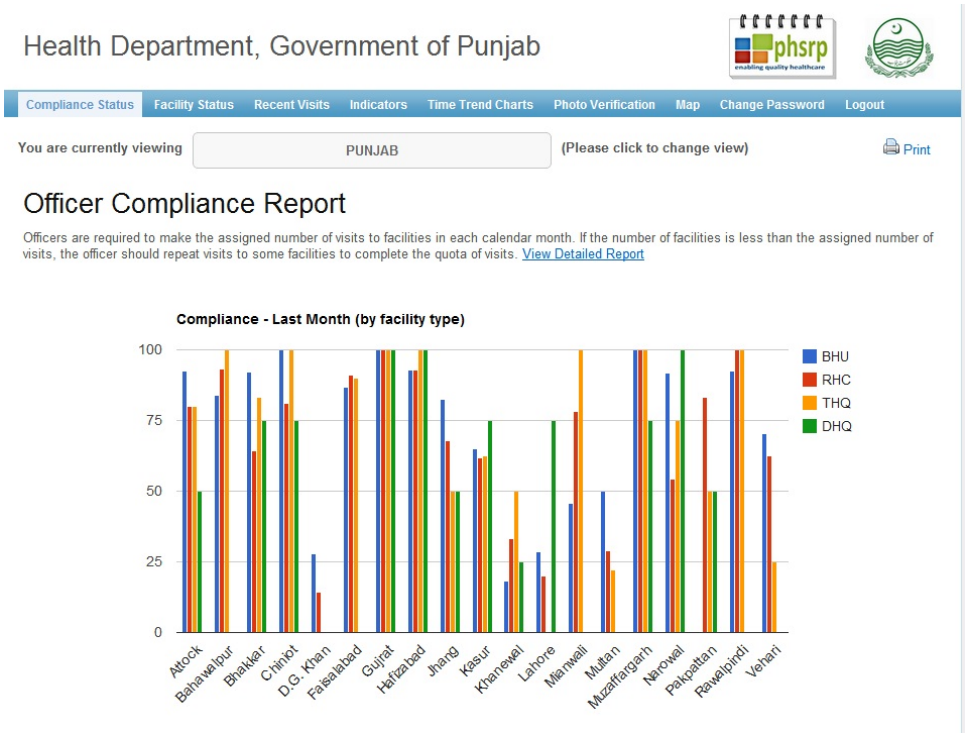

Figure 4: Online Dashboard - Summary of Inspection Compliance by District

diately and automatically sent up the chain of command, and the required geo-tags, time stamps, and photos serve as instant verification that the inspector and all reported staff are present at the clinic being inspected.[12]

3.1.5 Measuring Inspector Performance

We have two measures of job performance for health inspectors: (i) a dummy equal to one if the facility records an inspection in the two months prior to an unannounced visit; and (ii) the same proxy measure of collusion between doctors and health inspectors to falsify reports as described in Section 3.1.2. These measures were obtained during the same three independent and unannounced inspections of health clinics described in Section 3.1.2. Baseline performance measures for health inspectors are reported in Table A.1.

[12]See Callen et al. (2020) for the core results from the broad Monitoring the Monitors experiment.

3.1.6 Personality and Inspector Performance

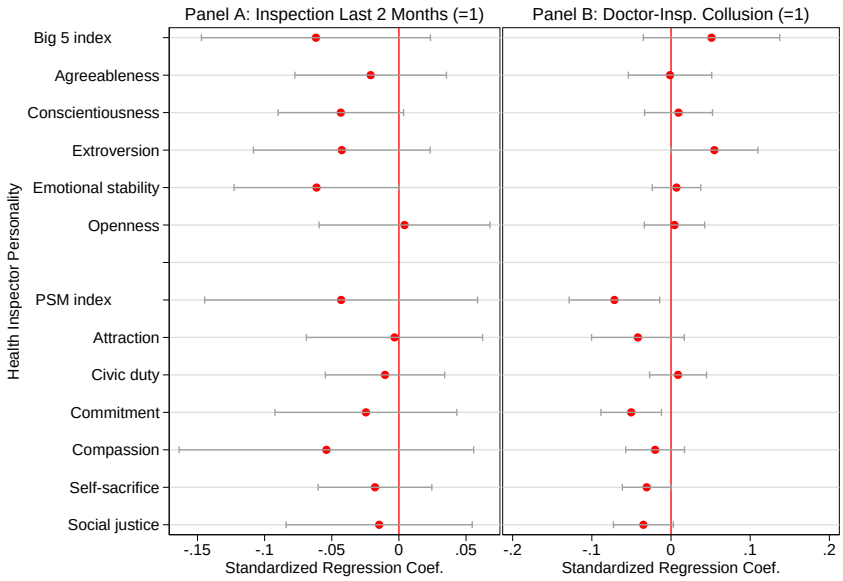

Figure 5: Personality and Performance: Health Inspectors

Notes: Each regression coefficient reported comes from a separate regression of the displayed performance measure on the indicated standardized health inspector personality measure. Error bars represent 95 percent confidence intervals. Standard errors are clustered at the clinic level. All regressions include tehsil (sub-district) and survey wave fixed effects. In all cases, personality measures are normalized to have mean zero and standard deviation of one in the sample, and thus the regression coefficients reported can be interpreted as the impact of a one standard deviation increase in a given personality trait or aggregate measure. The sample for Panel A is restricted to control district clinics for which health inspector personality data are available and a doctor is posted (467 observations across 46 inspectors). The sample for Panel B is restricted to treatment district clinics for which health inspector personality data are available and a doctor is posted (292 observations across 48 inspectors). Appendix Tables A.7 and A.8 provide corresponding regression tables.

We examine how much the personalities of health inspectors predict their job performance in control districts (i.e., those under status quo incentives) in Figure 5. In Panel A, we consider the relation between personalities and whether an inspection was carried out in the last two months. Appendix Table A.7 provides complete details of the results summarized here. We find a negative relationship between both conscientiousness and emotional stability and the number of inspections. We do not find any other statistically significant relationships between individual personality traits or the Big 5 or PSM index and inspections. In Panel B, we see that three of the six PSM traits are associated with less collusion, enough to distinguish the coefficient on the aggregate index from zero. In this case, health inspectors that score one standard deviation higher on aggregate PSM are about seven percentage

points less likely to be identified as potentially colluding. Appendix Table A.8 provides complete results of this analysis. As with our doctor correlations, despite the less clear story in Panel A, our 95 percent confidence intervals are fairly narrow, ruling out large positive or negative effects while allowing for a range of meaningful effect sizes both positive and negative. And our 95 percent confidence intervals in Panel B are substantially more narrow in standard deviation units. On the whole, compared to the other outcomes in this paper, we find the weakest evidence for the role of personality measures in predicting health inspectors performance (see Section 3.4).

The negative correlation between two personality measures and health inspections is worth discussion. Anecdotally this may be driven by "better" health inspectors understanding health inspections are not accomplishing anything under the status-quo and so choosing not to waste their time. This is consistent with the discussion in our framework that inspectors who derive pro-social utility from improving outcomes would not be motivated to conduct inspections if they see them as not leading to better outcomes. Of course, this is a very nuanced understanding of performance that our data cannot capture.

Since our collusion outcome is defined at the doctor-inspector level, we can also examine how doctor and inspector traits simultaneously predict collusion. We find no evidence that these traits interact when predicting collusion. In specifications which include doctor personality, inspector personality, and their interaction, only the coefficients on doctor personality predict collusion in twelve of thirteen cases. See Appendix Table A.9 for these results. While many stories could explain this pattern, it is consistent with the fact that inspectors may not see calling doctors ahead of a visit as a courtesy while doctors choosing to only come in when called is clearly shirking.

In Appendix Table A.10, we examine how health inspector personality predicts job performance relative to six other plausibly important observables—age, whether the inspector has completed higher education, the inspector's tenure in the department of health, the inspector's tenure as an inspector, the distance from the inspector's office to his hometown (in

KM), and a dummy for whether the inspector reports liking his current post. We do not find that any of these six observables are systematically better predictors than personality. In fact, the PSM index is clearly the strongest predictor in this exercise.

3.2 Do personality measures predict responses to a reform that changes incentives?

We now consider whether personality traits, including the tendency to procrastinate, predict health inspectors' response to a reform that increased incentives to complete inspections. In other words, does the stock of workers that has been selected to work for Punjab's Health Department interact with an effort to improve incentives?

3.2.1 Evaluating the Smartphone Monitoring

Our experimental sample comprises all health facilities in the district of Punjab, which has a population of at least 85 million citizens. Tens of millions of public sector health users therefore were potentially affected by the program. As described above, we monitored a subsample of 850 clinics, drawn to be representative of facilities in the province, using independent and unannounced inspections. We randomly implemented the program in 18 of the 35 districts in our experimental sample. In assigning treatment, we stratified on baseline attendance and the number of clinics in a district to ensure a roughly even number of treatments and controls. Figure 6 depicts control and treatment districts.

3.2.2 Personality and Treatment Response

We investigate whether impacts of the monitoring program are heterogeneous by the personality type of the inspector. Table 1 presents personality measures by treatment status for doctors and health inspectors. There is one significant difference in the balance table—treated health inspectors have slightly lower civic duty scores than those in control groups

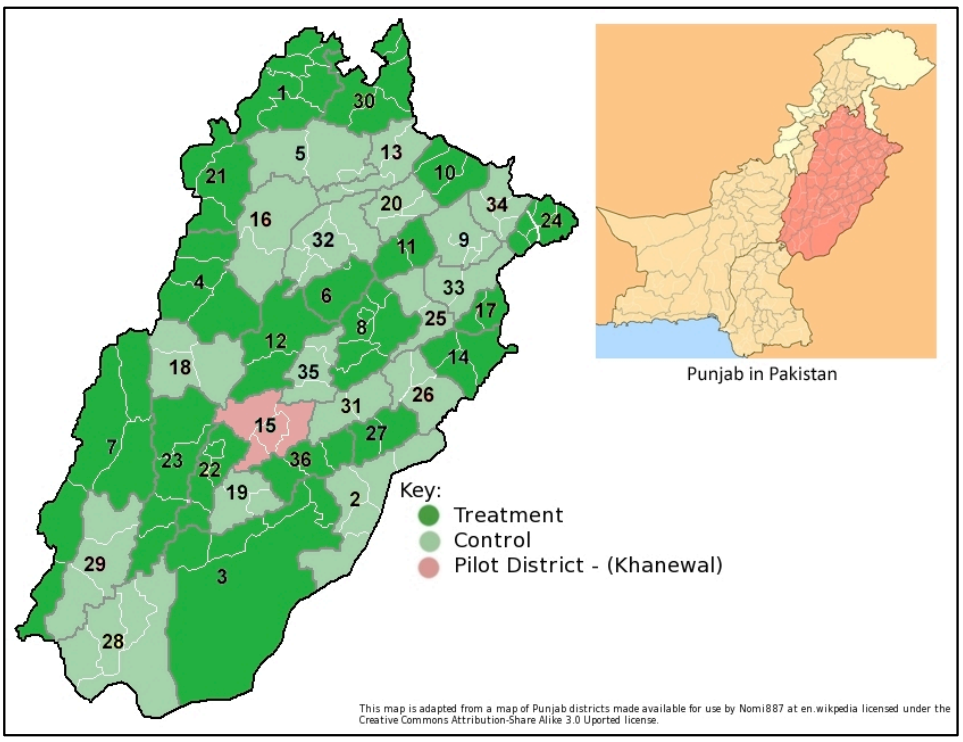

Figure 6: Treatment and Control Districts

on average. This is plausibly due to sampling fluctuation as it is a fairly small difference and the only one among the 27 differences estimated.

We consider the effects of an increase in health inspector monitoring on their performance by inspector personality. Results are presented in Table 2.[13] We estimate regressions using the difference-in-difference specification

$$Y_{dit} = \beta_0 + \beta_1 Trait_{di} + \beta_2 Treatment_{dit} + \beta_3 Treatment_{dit} \cdot Trait_i + \delta_t + \lambda_i + \varepsilon_{dit} \quad (1)$$

where Y_{dit} is a dummy equal to one if a facility records an inspection in the prior two months, $Treatment_{dit}$ is a variable equal to one for treated districts during the post-treatment periods (waves two and three), where i refers to the clinic, d refers to the district, and t to the survey

[13]Our other previous measure of performance, collusion between inspectors and doctors, cannot be studied in this context because the construction of collusion relies on data from our treatment districts' smartphone app. We have no information on health inspector-reported doctor attendance in the control districts of the Monitoring the Monitors experiment.

Table 1: Treatment Balance on Doctor and Health Inspector Personality

	Big Five Personality Traits							
	Doctor Personality Traits				Inspector Personality Traits			
	Treatment	Control	Difference	P-value	Treatment	Control	Difference	P-value
Big Five Index	-0.058	0.042	-0.100	0.295	-0.017	0.018	-0.035	0.801
	[0.713]	[0.820]	(0.095)		[0.637	[0.738]	(0.138)	
Agreeableness	3.498	3.577	-0.079	0.309	3.783	3.666	0.117	0.253
	[0.622]	[0.678]	(0.077)		[0.477]	[0.537]	(0.102)	
Conscientiousness	3.958	3.996	-0.037	0.605	4.159	4.113	0.046	0.646
	[0.548]	[0.570]	(0.072)		[0.452]	[0.531]	(0.099)	
Extroversion	3.624	3.686	-0.062	0.277	3.703	3.724	-0.021	0.830
	[0.464]	[0.501]	(0.057)		[0.525]	[0.459]	(0.099)	
Emotional Stability	-2.647	-2.536	-0.111	0.180	-2.461	-2.343	-0.119	0.322
	[0.641]	[0.702]	(0.082)		[0.571]	[0.618]	(0.119)	
Openness	2.926	2.932	-0.006	0.907	3.020	3.123	-0.103	0.218
	[0.372]	[0.451]	(0.050)		[0.471]	[0.353]	(0.083)	
	Perry Public Service Motivation							
	Doctor Personality Traits				Inspector Personality Traits			
	Treatment	Control	Difference	P-value	Treatment	Control	Difference	P-value
PSM Index	-0.017	-0.018	0.001	0.989	-0.061	0.064	-0.125	0.309
	[0.695]	[0.691]	(0.079)		[0.621]	[0.610]	(0.122)	
Attraction	3.481	3.442	0.039	0.581	3.552	3.585	-0.033	0.764
	[0.630]	[0.610]	(0.070)		[0.532]	[0.575]	(0.110)	
Civic duty	4.182	4.184	-0.002	0.969	4.255	4.421	-0.165	0.051
	[0.594]	[0.526]	(0.059)		[0.415]	[0.432]	(0.084)	
Commitment	3.773	3.774	-0.001	0.982	3.915	3.956	-0.040	0.628
	[0.511]	[0.463]	(0.050)		[0.458]	[0.379]	(0.083)	
Compassion	3.493	3.546	-0.053	0.432	3.743	3.663	0.080	0.400
	[0.515]	[0.516]	(0.067)		[0.475]	[0.484]	(0.095)	
Self Sacrifice	4.065	4.080	-0.015	0.820	4.316	4.392	-0.077	0.409
	[0.563]	[0.574]	(0.065)		[0.482]	[0.450]	(0.092)	
Social Justice	3.950	3.906	0.044	0.464	4.098	4.196	-0.098	0.284
	[0.571]	[0.619]	(0.060)		[0.490]	[0.427]	(0.091)	
# Health Workers	242	147			52	50		

Notes: Variable standard deviations reported in brackets. Standard errors clustered at the district level reported in parentheses. The doctor sample is limited to clinics where a doctor is posted at baseline. The Big Five and PSM traits are each mean responses to statements that represent the trait on a five point Likert scale, in which 1 corresponds to disagree strongly, 2 to disagree a little, 3 to neutral, 4 to agree a little, and 5 to agree strongly. Likert responses are given the same direction (5 always being more agreeable, for example, never less). The Big Five and PSM indices are z-score averages of the five and six traits within the Big Five and PSM respectively. Actual observations for each regression vary by a small amount based on no responses.

wave, and $Trait_i$ is a personality trait of the inspector overseeing facility i. δ_t and λ_i are survey wave and clinic fixed effects, respectively. We cluster all standard errors at the district level. For each regression, we present both standard, asymptotic p-values for the hypothesis test that $\beta_3 = 0$ based on these clustered standard errors and adjusted p-values. These adjusted p-values correct for multiple hypothesis testing. Our procedure to do so is explained at length in Section 3.4.

For health inspectors, there are heterogeneous effects of our experiment on the rate of health inspections according to personality traits. Health inspectors with a Big Five index

Table 2: Testing for Heterogeneous Impacts of Monitoring by Personality Type

	Health Inspection in Last Two Months (=1)								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
PANEL A: Big Five Personality Traits									
Monitoring (=1)		0.178	0.022	-0.006	0.010	0.003	0.030	-0.033	0.023
		(0.154)	(0.129)	(0.114)	(0.109)	(0.115)	(0.124)	(0.118)	(0.129)
Monitoring x Big Five Index				0.351**					
				(0.133)					
Monitoring x Agreeableness					0.170*				
					(0.094)				
Monitoring x Conscientiousness						0.186*			
						(0.102)			
Monitoring x Extroversion							0.116		
							(0.098)		
Monitoring x Emotional Stability								0.210**	
								(0.083)	
Monitoring x Openness									0.195
									(0.126)
Mean of dependent variable		0.641	0.655	0.655	0.655	0.655	0.655	0.655	0.655
# Observations		1332	1146	1146	1146	1146	1146	1146	1146
# Clinics		645	548	548	548	548	548	548	548
R-Squared		0.048	0.048	0.069	0.069	0.062	0.053	0.064	0.063
P-value		0.256	0.867	0.013	0.078	0.078	0.245	0.017	0.133
Adjusted P-value				0.083	0.214	0.214	0.274	0.101	0.249
PANEL B: Public Service Motivation									
Monitoring (=1)	0.178	0.022	0.023	0.026	0.039	0.024	0.012	0.041	0.021
	(0.154)	(0.129)	(0.120)	(0.111)	(0.127)	(0.111)	(0.119)	(0.130)	(0.122)
Monitoring x PSM Index			0.202						
			(0.140)						
Monitoring x Attraction				0.211**					
				(0.078)					
Monitoring x Civic Duty					-0.029				
					(0.066)				
Monitoring x Commitment						0.103			
						(0.082)			
Monitoring x Compassion							0.184		
							(0.115)		
Monitoring x Self Sacrifice								0.016	
								(0.090)	
Monitoring x Social Justice									0.014
									(0.102)
Mean of dependent variable	0.641	0.655	0.648	0.648	0.648	0.648	0.648	0.648	0.648
# Observations	1332	1146	1165	1165	1165	1165	1165	1165	1165
# Clinics	645	548	556	556	556	556	556	556	556
R-Squared	0.048	0.048	0.057	0.076	0.051	0.062	0.062	0.054	0.053
P-value	0.256	0.867	0.159	0.011	0.661	0.218	0.119	0.863	0.892
Adjusted P-value			0.250	0.101	0.508	0.274	0.249	0.508	0.508

Notes: This table reports heterogeneous impacts of our smartphone monitoring treatment by personality type. Column (1) reports average treatment effects on treatment and control district clinics. Columns (2) - (10) are limited to clinics in tehsils for which health inspector personality data is available. The difference in observations between Panels A and B is due to one inspector answering the PSM but not the Big Five survey. The Big Five and PSM traits are each mean responses to statements that represent the trait on a five point Likert scale, in which 1 corresponds to disagree strongly, 2 to disagree a little, 3 to neutral, 4 to agree a little, and 5 to agree strongly. Likert responses are given the same direction (5 always being more agreeable, for example, never less). All personality traits are then normalized across inspectors. The Big Five and PSM indices are z-score averages of the five and six traits within the Big Five and PSM respectively. P-values reported are from a two-sided hypothesis test that the null effect is zero. Adjusted P-values are corrected for multiple hypothesis testing. One correction is done across the Big Five and PSM indices P-values using the Family-Wise Error Rate procedure. A second is done across the eleven Big Five and PSM traits using False Discover Rate procedure. Both procedures are reported in Anderson (2008). *Levels of Significance:* $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

one standard deviation above the mean, for example, exhibit a 35 percentage point higher treatment effect in terms of health inspections. With an unconditional mean inspection rate of 66 percent, inspectors with a z-score one standard deviation above the mean come very close to completing all of their inspections as a result of treatment. We decompose this effect in columns (5)-(9) and find that that it is being driven most strongly by emotional stability—the trait of being able to capably respond to new stressors and demands. Besides openness, all Big Five traits have positive and large coefficients. We also see some positive and similarly large effects of the PSM index, attraction, and compassion within the PSM traits, though only attraction is significant.[14,15]

Figure 7 presents nonparametric treatment effects of health inspector Big Five index across the distribution of inspectors according to the Big Five index. We can see that the effect in Table 2 is primarily being driven by those health inspectors in the middle of the Big Five distribution. This fits the framework presented in Section 2.2 in which it is plausible that the effects of this intervention are localized to those inspectors in the middle of the distribution. See Appendix Figures A.2 and A.3 for nonparametric treatment effects trait-by-trait. While the location of the treatment effect peaks varies by trait, the overall shape is similar for specific traits.[16] Note in this figure we see a negative correlation between health inspector Big Five and health inspections in the control group. The slope coefficient of this

[14]See Appendix Table A.11 for heterogeneous treatment effects with whether a facility was inspected in the last month as the outcome (as opposed to the last two months reported here). Our effects are not robust to this different outcome. Control inspection rates led us to select the two month indicator as our preferred outcome in this paper: whereas control facilities are inspected 23 percent of the time in the last month before our surveys on average, they are inspected 61 percent of the time in the last two months. The fact that few inspections are happening in a one month horizon suggests it may be a more intractable outcome. Callen et al. (2020), used the one month outcome because the aim there was to evaluate the policy at achieving its stated goals which was monthly visits.

[15]Note that to test for robustness in our effects to the small number of district clusters in our analysis, we have conducted Fisher exact tests (randomization inference) for all heterogeneous treatment results as a separate exercise to adjusting for multiple hypothesis testing. In all cases, the estimated p-value is as at least as significant as from un-adjusted OLS. We have also separated the differential effects into our two post-treatment survey waves and find that the results sustain over time for as long as we were able to follow health clinics (roughly one year after treatment began). This is important because Callen et al. (2020), documents that the overall treatment effects on health inspections do in fact fade by the second survey wave. Results available upon request.

[16]Note that the point estimates in Figure 7 do not match those from Table 2. This is due to the fact that the regressions in the table include survey wave and clinic fixed effects.

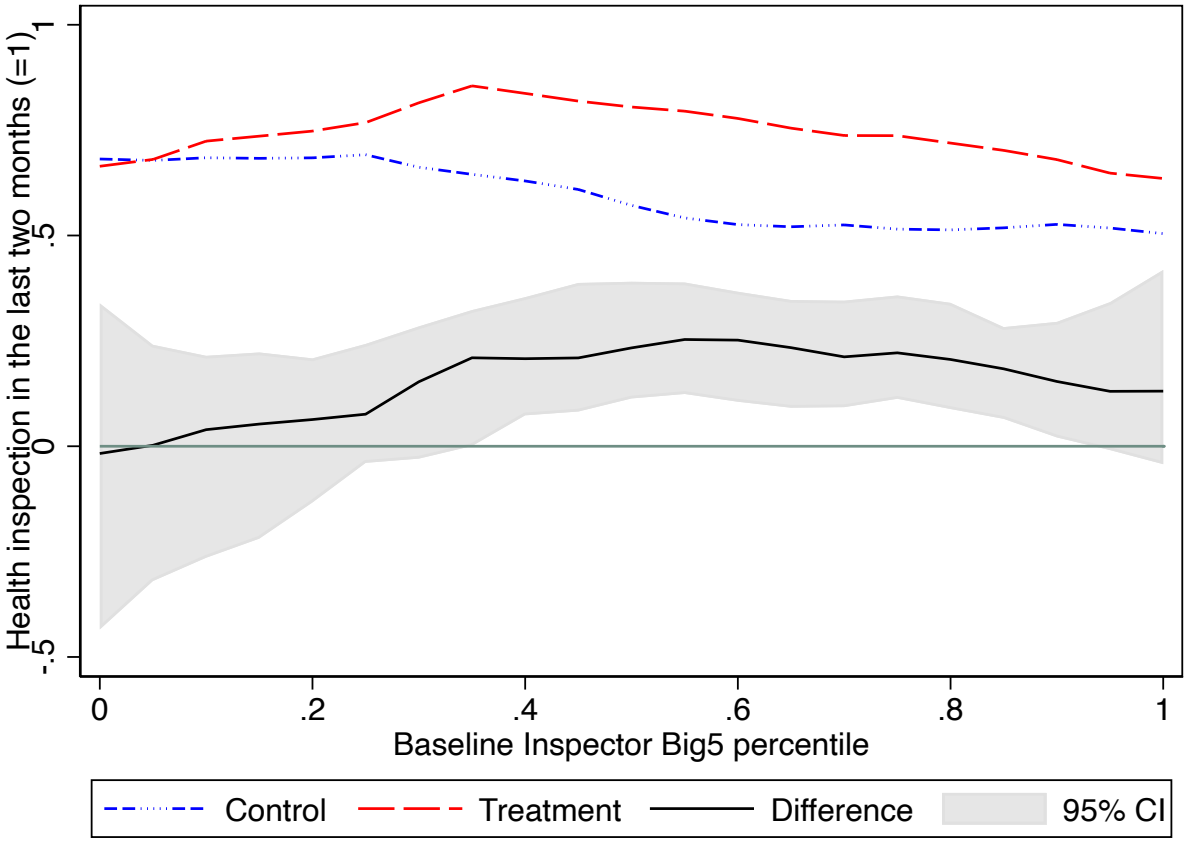

Figure 7: Nonparametric treatment effects

Notes: This figure plots a kernel-weighted local polynomial regression of whether a clinic had a health inspection in the last two months on every 5th percentile of baseline Big Five index separately for treatment and control districts, as well as the difference at each 5th percentile of baseline scores. The confidence intervals of the treatment effects are constructed by drawing 1,000 bootstrap samples of data that preserve the within-district correlation structure in the original data and plotting the 95 percent range for the treatment effect at each 5th percentile of baseline scores. Data from 794 observations across 93 health inspectors over two post-treatment survey waves.

line is -0.07 (s.e. 0.04, p-value 0.07). This is consistent with our correlations in Figure 5. As we say above, while our data is insufficient to prove it, this is consistent with the discussion in our framework that inspectors who derive pro-social utility from improving outcomes would not be motivated to conduct inspections if they see them as not leading to better outcomes.

There are two more points to make about these experimental results. First, while treatment is randomized, the personality characteristics of inspectors likely correlate with other measures. In Appendix tables A.12 we report results interacting treatment with both personality measures and the available set of observables. Estimates of the main coefficients

of interest remain stable. Second, the coefficients on the interaction terms are large, and the associated 95 percent confidence intervals include correspondingly large effect sizes. For example, effect sizes up to 0.61 standard deviations lie within our 95 percent confidence interval for the interaction of treatment with inspectors' Big Five personality scores. Increased inspections may not lead to an overall increase in doctor attendance, but they generate information that is helpful in the case that a health inspector or more likely a senior health official *is* interested in enforcing attendance. We will see this directly in the next subsection.

3.3 Do personality measures predict who will respond to salient information on subordinate absence?

In this section, we examine whether personality identifies the senior health officials who will react to information about the absence of their subordinates. To do this we study the response of senior officials, as measured by doctor absenteeism in clinics under their supervision, to a second policy intervention in which we manipulated the presentation of information to these officials.

3.3.1 Information Experiment

The Monitoring the Monitors system aggregates data from health inspections and presents them to senior health officials in each district of Punjab on an online dashboard. This dashboard is only visible to these senior health officials as well as to the Secretary of Health for Punjab and the Director General of Health for Punjab. Figure 8 provides an example of a dashboard view visible to senior health officials.

To test whether senior health officials react to information about the absence of their subordinates, we directly manipulated the data on the dashboard to make certain facilities with high staff absence salient. This was achieved by highlighting in red, or "flagging" reports by inspectors that found three or more staff absent at a clinic.[17] This cutoff of three

[17]Callen et al. (2020) examines at length whether this manipulation affects subsequent doctor absence,

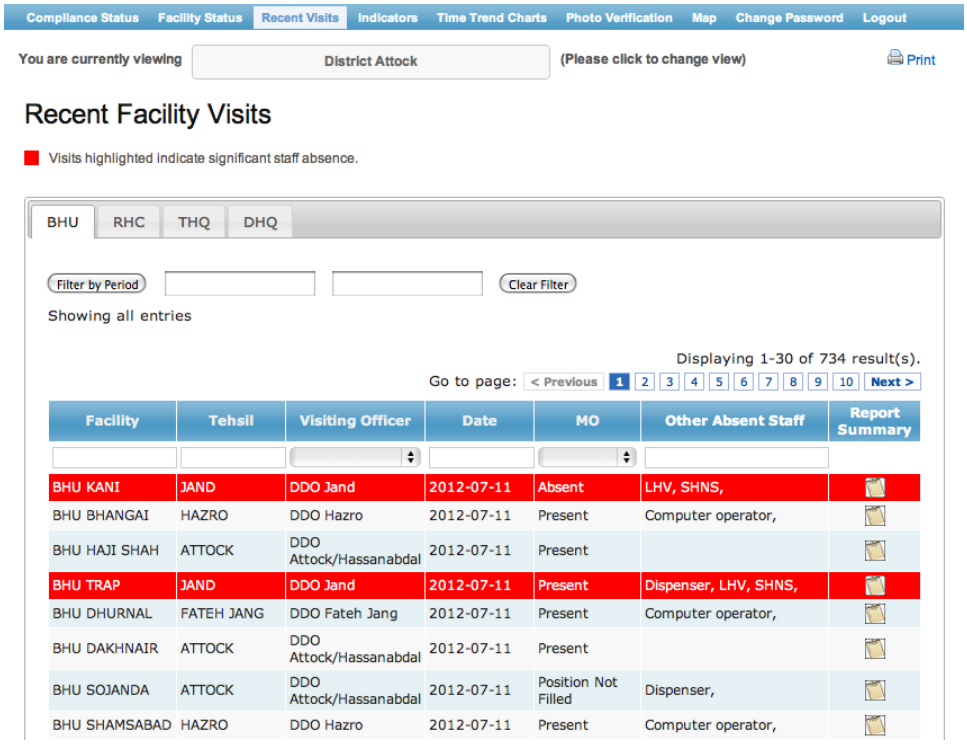

Figure 8: Highlighting Underperforming Facilities to Test Mechanisms

or more staff absences was set by our research team and was not communicated to any of the doctors, health inspectors, or senior health officials. We selected this cut-off based on the distribution of staff absence from baseline data. The peak of the distribution lies at two or three absent staff, suggesting that a cut-off at the center of this peak would yield the highest power to detect an effect of flagging in red.

Though the cutoff was purposefully arbitrary, our motivation for making absence data salient was not. Senior health officials in Punjab are in charge of health service provision in their district. These officials are constantly receiving information from facilities, staff, and citizens. Given the volume of information available to these officials, we designed the intervention to test whether making information salient could catalyze action by senior health officers.

finding consistent evidence that flagging facilities leads to decreased subsequent doctor absence.

3.3.2 Personality Predicts Response to Information

Appendix Table A.2 presents summary statistics for senior health officials in Punjab, which are similar in magnitude to summary statistics of both doctors and health inspectors. We examine whether manipulating attendance information affects subsequent doctor absence with the following specification

$$Absent\ Survey_{it} = \psi_0 + \psi_1 Trait_i + \psi_2 Flagged_{it-1} + \psi_3 Trait_i * Flagged_{it-1} + \delta_t + \eta_{it} \quad (2)$$

where $Absent\ Survey_{jt}$ is equal to one if the doctor posted to facility i was absent during our unannounced visit in wave t, $Flagged_{it-1}$ is a dummy equal to one if the facility was flagged in red on the dashboard prior to survey wave t, $Trait_i$ is a personality measure for the senior official in charge of facility i, and δ_t are survey wave fixed effects. For each regression, we present both standard, asymptotic p-values for the hypothesis test that $\psi_3 = 0$ based on clustered standard errors and adjusted p-values. These adjusted p-values correct for multiple hypothesis testing. Our procedure to do so is explained at length in Section 3.4.

Facilities are flagged only if three or more staff members are absent. Consequently, if we restrict our sample to only facilities where, in the month prior to our unannounced visit, only two or three staff were absent, we can estimate the effect of flagging on a sample where the only difference might plausibly be whether the facility was flagged.[18]

Table 3 reports results from this test, limiting the sample to facilities with two or three staff absent during an inspection. Facilities flagged for absence to a senior official with a Big Five index one standard deviation above the mean subsequently experience an increase in doctor attendance that is 40 percentage points greater than a facility flagged to a senior official at the mean Big Five index.[19] The 95 percent confidence interval for this heterogeneous

[18]In Appendix Table A.13 we verify the drop in absence for people who score higher on the Big Five index is limited to right around the discontinuity, with a waning, though significant, effect in a slightly larger window.

[19]Note that in Table 3 we cannot reject the null hypothesis that the interaction term on the Big Five index is different than the uninteracted flagging effect. In Appendix Table A.14, we show that when senior health officials' are split into quartiles by Big Five index, we can significantly reject that those in the bottom

Table 3: Tests of Heterogeneity in the Information Treatment by Senior Official Personality

	Doctor Present (=1)								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
PANEL A: Big Five Personality Traits									
Clinic Flagged as Underperforming on Dashboard		-0.161*	-0.146	0.159	0.140	0.144	0.132	0.154	0.163
		(0.095)	(0.103)	(0.098)	(0.103)	(0.100)	(0.105)	(0.100)	(0.110)
Flagged x Big Five Index				0.402**					
				(0.200)					
Flagged x Agreeableness					0.086				
					(0.144)				
Flagged x Conscientiousness						0.172*			
						(0.097)			
Flagged x Extroversion							0.097		
							(0.096)		
Flagged x Emotional Stability								0.185*	
								(0.105)	
Flagged x Openness									0.051
									(0.106)
Mean of dependent variable		0.563	0.520	0.520	0.520	0.520	0.520	0.520	0.520
# Observations		142	123	123	123	123	123	123	123
# Clinics		122	106	106	106	106	106	106	106
R-Squared		0.226	0.204	0.231	0.206	0.227	0.211	0.219	0.205
P-value		0.092	0.160	0.047	0.551	0.078	0.313	0.081	0.630
Adjusted P-value				0.000	1.000	0.747	0.781	0.747	1.000
PANEL B: Public Service Motivation									
Clinic Flagged as Underperforming on Dashboard	-0.161*	-0.146	0.165	0.146	0.155	0.254**	0.153	0.146	0.201*
	(0.095)	(0.103)	(0.105)	(0.103)	(0.104)	(0.121)	(0.110)	(0.103)	(0.108)
Flagged x PSM Index			0.124						
			(0.169)						
Flagged x Attraction				0.072					
				(0.102)					
Flagged x Civic Duty					0.027				
					(0.089)				
Flagged x Commitment						0.231			
						(0.148)			
Flagged x Compassion							-0.028		
							(0.114)		
Flagged x Self Sacrifice								-0.032	
								(0.100)	
Flagged x Social Justice									0.139
									(0.097)
Mean of dependent variable	0.563	0.520	0.520	0.520	0.520	0.520	0.520	0.520	0.520
# Observations	142	123	123	123	123	123	123	123	123
# Clinics	122	106	106	106	106	106	106	106	106
R-Squared	0.226	0.204	0.208	0.207	0.204	0.217	0.204	0.204	0.219
P-value	0.092	0.160	0.464	0.481	0.761	0.123	0.809	0.749	0.155
Adjusted P-value			1.000	1.000	1.000	0.747	1.000	1.000	0.747

Notes: This table tests for heterogeneity in the impact of providing information about clinic staff absence to senior officials by the personality types of the senior officials. Clinics were flagged in red on an online dashboard if three or more of the seven staff were absent in one or more health inspections of the clinic fifteen to forty-five days prior to an unannounced visit by our survey enumerators. All columns restrict the sample to those clinics where only two or three staff were absent (up to seven staff can be marked absent). In addition, the sample is limited to Monitoring the Monitor treatment districts due to the necessity of the web dashboard for flagging clinics. Column (1) reports un-interacted impacts of flagging. Columns (2) - (10) are further limited to clinics in districts for which senior health official personality data is available. The Big Five and PSM traits are each mean responses to statements that represent the trait on a five point Likert scale, in which 1 corresponds to disagree strongly, 2 to disagree a little, 3 to neutral, 4 to agree a little, and 5 to agree strongly. Likert responses are given the same direction (5 always being more agreeable, for example, never less). All personality traits are then normalized across inspectors. The Big Five and PSM indices are z-score averages of the five and six traits within the Big Five and PSM respectively. Standard errors clustered at the clinic level reported in parentheses. All regressions include district and survey wave fixed effects and condition on a doctor being posted. P-values reported are from a two-sided hypothesis test that the null effect is zero. Adjusted P-values are corrected for multiple hypothesis testing. One correction is done across the Big Five and PSM indices P-values using the Family-Wise Error Rate procedure. A second is done across the eleven Big Five and PSM traits using False Discover Rate procedure. Both procedures are reported in Anderson (2008). *Levels of Significance:* $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

flagging effect is from 0.01 to 0.79, covering a wide positive range.

There are several ways through which the above effect may have operated. For instance, the health officials could have taken formal action against delinquent workers, or they could simply have censured the officers informally. While we are unable to discern this effect given our data, anecdotally, we have learned that the second channel is more likely to work, given limited powers for hiring and firing people.

Appendix Table A.15 provides suggestive evidence that senior health officials with higher personality types stepped up the share of their time spent monitoring health facilities in response to dashboard flags. You can see senior health officials with a one standard deviation higher Big Five index increased the share of their time spent monitoring health facilities by 3.1 percentage points for each facility that was flagged in their district in the window prior to our collection of their time use information (wave three). The mean number of flags per district in this time-frame was 7.88, which translates to large increases in time spent monitoring by better personality types in response to flags. Although, this evidence is at best suggestive because it is based on 17 observations.[20]

The worry with the above results is that senior health officials might be substituting other work with increased monitoring of health facilities. The data suggest that senior health officials may have decreased their share of time spent on the lunch prayer break, on work related to monthly polio vaccination drives, and on 'other work' in response to flags.

and top quartile have the same flagging effect (with a substantial differential effect). We define the window during which a clinic can be flagged in red prior to one of our unannounced visits as 15 to 45 days before our visit. Senior health officials only looked at the web dashboard every week or two, so we would not expect an immediate response from flagging. However, if the window is made too long, virtually every facility will become flagged and we will lose variation. The p-values of the significance of the coefficient on the Big Five index and PSM index for a wide range of windows are reported in Appendix Figures A.4 and A.5. These figures also indicate that we have not selected the window most favorable for our result.

[20]Time use information was collected through a written module provided in the same visit in which personality measures were collected in which officials were asked to account for all work activities in each half-hour block between 8:30am and 8:30pm from the last two regular work days. Officials could choose from fourteen categories, including Monitoring Visits to the BHUs, Management of BHUs done in the office, Meetings with BHU staff in office, Monitoring visits to RHCs, Management of RHCs done in the office, Monitoring visits to THQ & DHQ, Management of THQ & DHQ done in the office, Lunch/Prayer break, Tea Break, Meeting with General Public, Meeting with other Govt. Official, EPI and Polio, Other Official activities, and Other.

These effects are not significant individually.[21]

As with the correlational and experimental results above, we show that personality is a better predictor of the response to information than other important covariates for senior health officials. See Appendix Table A.16 for these results.

The results presented in this section provide another validation of personality measures in predicting performance, this time in the case of senior health officials. Personality measures predict which senior health officials will react to information about the absence of their subordinates with large magnitudes. Simply flagging high absence clinics in red essentially eliminates doctor absence in clinics overseen by senior health officials one standard deviation above the mean in terms of their Big Five index. These results also speak to potential mechanisms. It seems plausible that the same information treatment provided to individuals in highly comparable positions results in different real world impacts because different personality types take different action in response to information.

3.4 Summary of Results and Multiple Hypothesis Testing

Consistent with a growing emphasis in economics on accounting for potential overrejection of the null hypothesis of no effect that may result from multiple inference, we present multiple inference adjusted p-values for all of our primary analysis (Anderson, 2008; Miguel et al., 2014; Bidwell et al., 2016; Casey et al., 2012). This primary analysis measures the association between two different personality measures and six objective performance measures for public health workers at three different levels of the bureaucracy in Punjab, Pakistan. As explained in Section 3.1.1, we primarily consider a single index each as the measures of the Big Five and Perry Public Service Motivation personality traits. Creating an index to collapse multiple hypothesis tests into one is a common means of accounting for multiple inference (Kling et al., 2007). However, as we are still testing two null hypotheses for each of our performance measures—that the Big Five index is not associated with differential performance and that

[21] Category-by-category time use tables available by request.

the PSM index is not—we adjust p-values across these two indices for each outcome.

Specifically, for correlations between personality measures and doctor and inspector performance under status quo incentives, we apply false discovery rate (FDR) adjustments at the personality measure level. When testing for heterogeneous treatment effects, we apply family wise error rate (FWER) corrections at the personality measure level. In both cases we use the procedure outlined in Anderson (2008). While our preference would be to follow Anderson in applying the more conservative FWER corrections for all of our non-exploratory analysis, the FWER correction requires drawing placebo treatment assignments which is not possible for the status quo correlations. Thus we use the FDR correction.

For our exploratory, trait-by-trait analysis, we apply false discovery rate (FDR) adjustments at the personality trait level, adjusting for each of the eleven tests (pooling Big Five and PSM traits) we are conducting for each outcome. This is consistent with Anderson (2008), Bidwell et al. (2016), and Casey et al. (2012).

Table 4 presents a summary of p-values for rejecting the null hypothesis for each of our primary results with and without multiple inference corrections. Focusing on the indices, we reject the null of no association between personality and performance for six of twelve tests at the five percent level before we adjust for multiple inference. After adjusting, we reject the null for four of twelve tests at the five percent level and for six of twelve tests at the ten percent level. That is to say that adjusting our p-values causes two cases in which a coefficient previously significant at five percent slips to ten percent. We take this as encouraging for our argument that personality measures predict performance.

Adjusting for multiple inference has more of an impact on our exploratory, trait-by-trait analysis. We reject the null hypothesis of no relationship for twenty six of 66 tests at the ten percent level or below with unadjusted p-values. Once we adjust them for multiple inference, we reject the null only thirteen times at the ten percent level or below, and eleven of these thirteen are for one outcome—doctor collusion. Note however that an additional eleven adjusted p-values are between 0.1 and .25. Given how conservative these adjustments

Table 4: Results Summary

Alternative Hypothesis:	Personality Predicts Performance				Personality Predicts Monitoring Treatment Heterogeneity	Personality Predicts Information Treatment Heterogeneity
Public Actor:	Doctor		Inspector			Administrator
Performance Measure:	Attendance	Collusion	Inspections	Collusion	Inspections	Doctor Attendance
			Panel A: Un-adjusted P-Values			
Big 5 Index	+ (0.22)	- (0.00)	- (0.16)	+ (0.25)	+ (0.01)	+ (0.05)
Agreeableness	+ (0.73)	- (0.00)	- (0.47)	- (0.96)	+ (0.08)	+ (0.55)
Conscientiousness	+ (0.03)	- (0.01)	- (0.08)	+ (0.67)	+ (0.08)	+ (0.08)
Extroversion	+ (0.07)	- (0.01)	- (0.21)	+ (0.06)	+ (0.24)	+ (0.31)
Emotional Stability	+ (0.22)	- (0.00)	- (0.06)	+ (0.66)	+ (0.02)	+ (0.08)
Openness	- (0.52)	- (0.62)	+ (0.90)	+ (0.82)	+ (0.13)	+ (0.63)
PSM Index	+ (0.03)	- (0.00)	- (0.41)	- (0.02)	+ (0.16)	+ (0.46)
Attraction	+ (0.24)	- (0.02)	- (0.92)	- (0.17)	+ (0.01)	+ (0.48)
Civic Duty	+ (0.02)	- (0.02)	- (0.65)	+ (0.63)	+ (0.66)	+ (0.76)
Commitment	+ (0.21)	- (0.00)	- (0.48)	- (0.01)	+ (0.22)	+ (0.12)
Compassion	+ (0.70)	- (0.00)	- (0.34)	- (0.30)	+ (0.12)	- (0.81)
Self Sacrifice	+ (0.03)	- (0.00)	- (0.41)	- (0.06)	+ (0.86)	- (0.75)
Social Justice	+ (0.20)	- (0.02)	- (0.68)	- (0.08)	+ (0.89)	+ (0.16)
			Panel B: P-Values Adjusted for Multiple Hypothesis Testing			
Big 5 Index	+ (0.12)	- (0.00)	- (0.48)	+ (0.14)	+ (0.08)	+ (0.00)
Agreeableness	+ (0.50)	- (0.01)	- (1.00)	- (1.00)	+ (0.21)	+ (1.00)
Conscientiousness	+ (0.12)	- (0.01)	- (0.73)	+ (0.80)	+ (0.21)	+ (0.75)
Extroversion	+ (0.15)	- (0.01)	- (1.00)	+ (0.23)	+ (0.27)	+ (0.78)
Emotional Stability	+ (0.27)	- (0.01)	- (0.73)	+ (0.80)	+ (0.10)	+ (0.75)
Openness	- (0.50)	- (0.06)	+ (1.00)	+ (0.97)	+ (0.25)	+ (1.00)
PSM Index	+ (0.07)	- (0.00)	- (0.48)	- (0.04)	+ (0.25)	+ (1.00)
Attraction	+ (0.27)	- (0.01)	- (1.00)	- (0.31)	+ (0.10)	+ (1.00)
Civic Duty	+ (0.12)	- (0.01)	- (1.00)	+ (0.80)	+ (0.51)	+ (1.00)
Commitment	+ (0.27)	- (0.01)	- (1.00)	- (0.17)	+ (0.27)	+ (0.75)
Compassion	+ (0.50)	- (0.01)	- (1.00)	- (0.53)	+ (0.25)	- (1.00)
Self Sacrifice	+ (0.12)	- (0.01)	- (1.00)	- (0.23)	+ (0.51)	- (1.00)
Social Justice	+ (0.27)	- (0.01)	- (1.00)	- (0.24)	+ (0.51)	+ (0.75)

Notes: This table provides a summary of coefficient direction and P-values (in parentheses) for the primary hypothesis tested in each of the regressions available in Figures 3 and 5 and Tables 2 and 3.Coefficient directions are indicated by either + (positive) or - (negative). P-values are in parentheses. Un-adjusted P-values reported are from a two-sided hypothesis test that the null effect is zero. Adjusted P-values are corrected for multiple hypothesis testing. One correction is done across the Big Five and PSM indices P-values using the Family-Wise Error Rate procedure. A second is done across the eleven Big Five and PSM traits using False Discover Rate procedure. Both procedures are reported in Anderson (2008).

are (they are more conservative than adjusting across outcomes within each trait or than adjusting within each personality measure separately, and we are using two-sided tests when one-sided could be more appropriate), we take these results to be a strong caveat against interpreting trait-by-trait results but one that does not change the underlying picture.

Note that we are correcting for multiple inference across personality measures within an outcome rather than across outcomes within a measure, as is more traditional in the literature. This is consistent with how we are interpreting our analysis outcome-by-outcome. However, as a robustness check Appendix Table A.17 presents multiple hypothesis corrections across outcomes for each personality measure. Note in this case we cannot use FWER corrections as we cannot draw placebo treatment assignments for status quo correlations, so we use FDR corrections across each of the six outcomes for a given trait. While there are some changes in significance levels, the results are, if anything, stronger with this approach to multiple hypothesis testing. 15 of 78 tests remain significant at the five percent level or lower, 27 are significant at the ten percent level, and an additional 17 have adjusted p-values between 0.1 and .25.

Putting all of our results together in one table also demonstrates some patterns in which traits are more often significantly predictive of performance. In Table 4, Panel A, amongst the Big 5 traits, conscientiousness is a significant predictor at ten percent or better in five of six tests, emotional stability in four, extroversion in three, agreeableness in two, and openness in zero. At the same time, amongst the PSM traits, self-sacrifice is significant in three of six tests, attraction, civic duty, commitment, and social justice in two, and compassion in one. These suggest if one has limited resources to measure personality traits that certain ones might be better to target. Of course, this also depends on the specific outcome of interest, as different traits are better predictors of different outcomes. The fact that conscientiousness is the most consistent predictor in this context is also not surprising given the prior literature (Borghans et al., 2008).

4 Conclusion

Governments, like any organization, are made of people with different qualities and personalities. We find that measurable differences in government worker personality predict performance both under status-quo incentives as well as who will respond to increased monitoring. Especially at senior levels, the relevance of personality traits is not ex ante clear. First, the small group who succeed in ascending through the hierarchy may all have similar traits. Second, they may see little value in information regarding their subordinates, for example, because political considerations dominate. The patterns we report suggest that selection matters both for performance and shapes how reforms play out at all levels of the hierarchy—that selection and incentives are *complements* for health service delivery in rural Pakistan.

A natural limitation of this study is that we did not randomize the stock of government employees at the time of our experimental change of incentives. Nor did we randomize those employees' personalities. Our data indicate that different workers responded differently to incentives, and that whatever characteristics are driving these differential responses are correlated with personality, however we cannot rule out omitted variables. While future work should investigate these potential omitted variables to inform theory, for purposes of prediction, potential omitted variables could be less important. We have demonstrated that personality can be measured usefully to predict performance and who responds to changes in incentives.

Lastly, this study did not directly measure health impact, and could not comment on how performance differences among doctors with different personalities affect the health of communities they serve. However, exploring the downstream impacts is an important consideration, especially from a policy perspective. While, Donato et al. (2017) makes progress on this question, this is an important area open for further exploration, particularly through the channel of selecting public sector workers with better personalities.

References

Alexander, James Madison Hamilton and John Jay, *The Fedralist: A Collection of Essays, Written in Favour of the New Constitution, as Agreed upon by the Federal Convention, September 17, 1787*, 2 vols. New York: J. and A. M'Lean, 1788.

Almlund, Mathilde, Angela Lee Duckworth, James J Heckman, and Tim D Kautz, "Personality psychology and economics," Technical Report, National Bureau of Economic Research 2011.

Aman-Rana, Shan, "In Self Interest? Meritocracy in a Bureaucracy," *Meritocracy in a Bureaucracy (June 11, 2020)*, 2020.

Anderson, Michael L, "Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects," *Journal of the American statistical Association*, 2008, *103* (484).

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

_ , _ , Edward Davenport, and Scott S Lee, "Losing prosociality in the quest for talent? Sorting, selection, and productivity in the delivery of public services," *American Economic Review*, 2020, *110* (5), 1355–94.

Bandiera, Oriana, Andrea Prat, and Tommaso Valletti, "Active and Passive Waste in Government Spending: Evidence from a Policy Experiment," *American Economic Review*, 2009, *99* (4), 1278–1308.

Banerjee, Abhijit V, Esther Duflo, and Rachel Glennerster, "Putting a band-aid on a corpse: incentives for nurses in the Indian public health care system," *Journal of the European Economic Association*, 2008, *6* (2-3), 487–500.

Barrick, Murray R. and Michael K. Mount, "The Big Five Personality Dimensions and Job Performance: A Meta-Analysis," *Personnel Psychology*, 1991, *44* (1), 1–26.

Bazerman, Max and Don A Moore, *Judgment in Managerial Decision Making, 8th Edition*, Wiley & Sons, 2012.

Benabou, Roland and Jean Tirole, "Intrinsic and extrinsic motivation," *The Review of Economic Studies*, 2003, *70* (3), 489–520.

Bertrand, Marianne and Antoinette Schoar, "Managing with Style: The Effect of Managers on Firm Policies," *Quarterly Journal of Economics*, 2003, *CXVIII*, 1169–1208.

Besley, Timothy, *Principled agents?: The political economy of good government*, Oxford University Press on Demand, 2006.

_ **and Maitreesh Ghatak**, "Competition and incentives with motivated agents," *American Economic Review*, 2005, *95* (3), 616–636.

Bidwell, Kelly, Katherine Casey, and Rachel Glennerster, "Debates: Voting and Expenditure Responses to Political Communication," 2016.

Blattman, Christopher, Julian C. Jamison, and Margaret Sheridan, "Reducing crime and violence: Experimental evidence on adult noncognitive investments in Liberia," 2015. Working paper.

Borghans, Lex, Angela Lee Duckworth, James J. Heckman, and Bas ter Weel, "The Economics and Psychology of Personality Traits," *The Journal of Human Resources*, 2008, *XLIII* (4), 973–1059.

Bowles, Samuel, Herbert Gintis, and Melissa Osborne, "The determinants of earnings: A behavioral approach," *Journal of Economic Literature*, 2001, pp. 1137–1176.

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

Casey, Katherine, Rachel Glennerster, and Edward Miguel, "RESHAPING INSTITUTIONS: EVIDENCE ON AID IMPACTS USING A PREANALYSIS PLAN," *The Quarterly Journal of Economics*, 2012, *1755*, 1812.

Chaudhury, Nazmul, Jeffrey Hammer, Michael Kremer, Karthik Muralidharan, and F Halsey Rogers, "Missing in action: teacher and health worker absence in developing countries," *The Journal of Economic Perspectives*, 2006, *20* (1), 91–116.

Cobb-Clark, Deborah A and Stefanie Schurer, "The stability of big-five personality traits," *Economics Letters*, 2012, *115* (1), 11–15.

Dal Bó, Ernesto, Frederico Finan, and Martín A. Rossi, "Strengthening State Capabilities: The Role of Financial Strengthening State Capabilities: The Role of Financial Incentives in the Call to Public Service," *Quarterly Journal of Economics*, 2013.

Deaton, Angus, "Instruments, Randomization, and Learning about Development," *Journal of Economic Literature*, 2010, *48* (2), 424–55.

DellaVigna, Stefano, Devin Pope, and Eva Vivalt, "Predict science to improve science," *Science*, 2019, *366* (6464), 428–429.

der Linden, Dimitri Van, Jan te Nijenhuis, and Arnold B Bakker, "The general factor of personality: A meta-analysis of Big Five intercorrelations and a criterion-related validity study," *Journal of research in personality*, 2010, *44* (3), 315–327.

Deserranno, Erika, "Financial Incentives as Signals: Experimental Evidence from the Recruitment of Village Promoters in Uganda," 2016.

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

Digman, John M, "Higher-order factors of the Big Five.," *Journal of personality and social psychology*, 1997, *73* (6), 1246.

Donato, Katherine, Grant Miller, Manoj Mohanan, Yulya Truskinovsky, and Marcos Vera-Hernández, "Personality traits and performance contracts: Evidence from a field experiment among maternity care providers in India," *American Economic Review*, 2017, *107* (5), 506–510.

Finan, Frederico, Benjamin A Olken, and Rohini Pande, "The personnel economics of the state," Technical Report, National Bureau of Economic Research 2015.

Gatewood, Robert, Hubert Feild, and Murray Barrick, *Human resource selection*, Cengage Learning, 2010.

Grossman, Guy and Tara Slough, "Government Responsiveness in Developing Countries," *Annual Review of Political Science*, 2022, *25*, 131–153.

Groth-Marnat, Gary, *Handbook of psychological assessment*, John Wiley & Sons, 2009.

Heckman, James J., "Integrating Personality Psychology into Economics," 2011, (NBER WP #17378).

_ , Jora Stixrud, and Sergio Urzua, "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior," *Journal of Labor Economics*, 2006, *24* (3), 411–482.

Hjort, Jonas, Diana Moreira, Gautam Rao, and Juan Francisco Santini, "How research affects policy: Experimental evidence from 2,150 brazilian municipalities," *American Economic Review*, 2021, *111* (5), 1442–80.

Hogan, Joyce and Brent Holland, "Using Theory to Evaluate Personality and Job-Performance Relations: A Socioanalytic Perspective," *Journal of Applied Psychology*, 2003, *88* (1), 100–112.

John, Oliver P., Laura P. Naumann, and Christopher J. Soto, "Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues," in "Handbook of personality: Theory and research," The Guilford Press, 2008, chapter 4.

Johnson, W. Bruce, Robert Magee, Nandu Nagarajan, and Harry Newman, "An Analysis of the Stock Price Reaction to Sudden Executive Deaths," *Journal of Accounting and Economics*, 1985, *7*, 151–174.

Kaplan, Robert M. and Dennis P. Saccuzzo, *Psychological Testing: Principles, Applications, and Issues*, Pacific Grove, Calif.: Brooks/Cole Pub. Co., 1997.

Kautz, Tim D, James J. Heckman, Ron Diris, Bas ter Weel, and Lex Borghans, "Fostering and Measuring Skills: Improving Cognitive and Non-Cognitive Skills to Promote Lifetime Success," 2014, (NBER WP #20749).

Kling, Jeffrey R, Jeffrey B Liebman, and Lawrence F Katz, "Experimental analysis of neighborhood effects," *Econometrica*, 2007, *75* (1), 83–119.

Klinger, Bailey, Asim Ijaz Khwaja, and Carlos del Carpio, *Enterprising Psychometrics and Poverty Reduction*, Springer, 2013.

Miguel, Edward, C Camerer, K Casey, J Cohen, KM Esterling, A Gerber, R Glennerster, DP Green, M Humphreys, G Imbens et al., "Promoting transparency in social science research," *Science*, 2014, *343* (6166), 30–31.

Musek, Janek, "A general factor of personality: Evidence for the Big One in the five-factor model," *Journal of Research in Personality*, 2007, *41* (6), 1213–1233.

National Institute of Population Studies, *Pakistan Demographic and Health Survey 2012-13*, National Institute of Population Studies, 2013.

Nyhus, Ellen K. and Empar Pons, "The Effects of Personality on Earnings," *Journal of Economic Psychology*, 2005, *26* (3), 363–384.

Olken, Benjamin A. and Rohini Pande, "Corruption in Developing Countries," *Annual Review of Economics*, 2012, *4*, 479–509.

Perry, James L., "Measuring Public Service Motivation: An Assessment of Construct Reliability and Validity," *Journal of Public Administration Research and Theory*, 1996, *6* (1), 5–22.

_ **and Lois Recascino Wise**, "The Motivational Bases of Public Service," *Public Administration Review*, 1990, *50*, 367–73.

Petrovsky, Nicolai, "Does Public Service Motivation Predict Higher Public Service Performance? A Research Synthesis," 2009.

Rasul, Imran and Daniel Rogger, "Management of bureaucrats and public service delivery: Evidence from the Nigerian civil service," *The Economic Journal*, 2018, *128* (608), 413–446.

Reinikka, Ritva and Jakob Svensson, "Local Capture: Evidence from a Central Government Transfer Program in Uganda," *The Quarterly Journal of Economics*, 2004, *119* (2), 679–705.

Roberts, Brent W., "Back to the Future: Personality and Assessment and Personality Development," *Journal of Research in Personality*, 2009, *43* (2), 137–145.

_ **, Kate E. Walton, and Wolfgang Viechtbauer**, "Patterns of Mean-Level Change in Personality Traits across the Life Course: A Meta-Analysis of Longitudinal Studies," *Psychological Bulletin*, 2006, *132* (1), 1–25.

Salgado, Jesus F., "The Five Factor Model of Personality and Job Performance in the The Five Factor Model of Personality and Job Performance in the European Community," *Journal of Applied Psychology*, 1997, *82* (1), 30–43.

Schmidt, Frank L and John E Hunter, "The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings.," *Psychological bulletin*, 1998, *124* (2), 262.

Wild, Lena, Vikki Chambers, Maia King, and Daniel Harris, "Common Constraints and Incentive Problems in Service Delivery," Technical Report, Overseas Development Institute 2012.

World Bank, *World Development Report 2004: Making Services Work for the Poor*, World Bank, 2004.