# Training Policymakers in Econometrics:

# Evidence from Two Field Experiments

*By* Sultan Mehmood, Shaheen Naseer, and Daniel L. Chen[1]

We explore the impact of econometrics education on policymakers' decision-making and policy outcomes through two randomized field experiments. In one experiment, junior ministers were trained in the credibility revolution, emphasizing the value of causal evidence. Treated ministers' stated valuation of quantitative evidence and commissioning of RCTs in policymaking increases. One year post-training, these policymakers are more likely to recommend funding for policies supported by RCT evidence, suggesting that econometrics education can enhance the utilization of causal knowledge in policymaking. In a separate field experiment with tax collectors, we assess whether econometrics education can influence tax policy and revenue generation. Using administrative tax data, we find that trained tax collectors are more likely to implement policy changes, such as sending tax reminder letters, leading to significant increases in tax collection. Across both experiments, the evidence indicates that econometrics education not only shapes beliefs and policy choices but also leads to tangible improvements in real-world outcomes—from recognizing and funding evidence-based policies to boosting the efficiency of tax collection efforts. (*JEL D72, D78, O17*)

*Keywords*: randomized control trials, policy, credibility revolution, paradigm shifts.

1 New Economic School (E-mail: smehmood@nes.ru). Lahore School of Economics (E-mail: ranashaheen133@gmail.com) and Toulouse School of Economics (Email: daniel.chen@iast.fr). An IRB is obtained from the Lahore School of Economics with IRB Number RERC-022021-01. We would like to thank the training Director and staff at the Civil Service Academy of Pakistan and National School of Public Policy and seminar participants at Institute for Advanced Study in Toulouse, Paris School of Economics, Aix-Marseille University, American Political Science Association, Econometric Society Annual Meeting 2021, NBER Fall Meeting 2021, New Economic School, Pompeu Fabra University, Erasmus University Rotterdam and World Bank for their helpful suggestions and comments. We also thank Josh Angrist for this feedback and the video lecture. Sajwaar Khalid, Bakhtawar Ali, Hafeez-ul-Hassan and Sameen Tariq provided excellent research assistance. Daniel Chen acknowledges IAST funding from the French National Research Agency (ANR) under the Investments for the Future (Investissements d'Avenir) program, grant ANR-17-EUR-0010. This research has also benefited from financial support of the research foundation TSE-Partnership. Pre-registered at AEA RCT Registry with ID AEARCTR-0006655.

## 1. Introduction

Over the last half century empirical economics has gone through a paradigm shift (Angrist and Pischke 2014). The credibility revolution, with its careful attention to causality and empirical rigor, has presented itself as a new paradigm for “taking the con out of econometrics” (Leamer 1983). Yet, as this paradigm shift reshapes the academic sphere, its ripple effects on the world of policy-making remain opaque. Policymakers, despite their reported openness to research, often find themselves at a crossroads—navigating between novel empirical insights and entrenched policy beliefs. This tension, underscored by a persistent adherence to prior convictions, may inadvertently stifle the implementation of policies that are crucial for sustainable economic development (Hjort et al. 2021; Baekgaard et al. 2019; Banuri et al. 2019; Metzger et al., 2020; Vivalt and Coville 2021; Lu and Chen 2021). How can policymakers be made more receptive to evidence? Will training them in concepts associated with the credibility revolution make them more likely to shift their beliefs and change their policy choices? Is the benefit of econometrics training limited to only the most elite bureaucrats, or does it also extend its impact to frontline policymakers? Using administrative data and two randomized control trials, we explore the relationship between empirical findings, policy beliefs, actual policies, and their broader societal implications, analyzing how much the credibility revolution can affect policy choices and have wider economic effects.

Both experiments use a well-known summary of the credibility revolution, Mastering ’Metrics: The Path from Cause to Effect (Angrist and Pischke 2014), as an instrument. In the first randomized trial, we conducted the experiment with elite bureaucrats in Pakistan, specifically deputy ministers. The Government of Pakistan regards these deputy ministers as the "key wheels on which the entire engine of the state runs" (Government of Pakistan 2019). The equivalent public officials in India, which were also passed down from the British Colonial governance system, are similarly referred to as the "steel frame of India" (Bertrand et al., 2020, p. 627). These bureaucrats are recruited, trained, and incentivized similarly and closely mirror those of numerous other developing nations, particularly India and Bangladesh. These countries, akin to Pakistan, inherited these bureaucratic structures from the British Colonial rule over the Indian subcontinent. The combined populations of Pakistan, India, and Bangladesh account for over a quarter of the global populace, amplifying the significance of this study to a substantial

number of individuals. Given the vast implications for such a significant portion of the world's population, it's worth noting that our initial experiment centered around a policy decision involving deworming. This policy shares many essential characteristics with other development policies as policymakers aim to make a decision in light of its potential consequences. A second experiment involved trying to influence state and fiscal capacity by changing tax policy – the sending of tax reminders in order to increase tax revenue. Across 213 elite deputy ministers in the first experiment and 310 frontline tax officers in the second experiment, we examined how the philosophical foundation associated with the credibility revolution influences their beliefs and policy decisions in a real-world experimental setup.

Our first experiment with deputy ministers had three stages and focused on the mechanisms of behavioral change. We elicited demand for the Mastering 'Metrics book, then randomized the econometrics training. We measured attitudes and beliefs before and after training. The demand for the Mastering 'Metrics book yielded a proxy for potential compliers to the metrics training treatment assignment. We controlled for whether individuals chose the metrics book as well as assess whether similar effects were observed for both high and low-demanders of econometrics training, that is, estimated the effect of treatment on deputy ministers who might be inframarginal—those less likely to be affected by treatment.

To gauge this baseline demand for causal thinking, we presented deputy ministers with a choice between two books. One of the books utilized was "Mastering 'Metrics." Its chapters delve into a range of topics, encompassing the central idea of the credibility revolution that randomized trials are the gold standard for empirical research, regression as a method for comparing means, non-experimental methods such as instrumental variables, regression discontinuity design and difference-in-differences. The other book was Mindsight, a self-help book, which focuses on developing a positive outlook towards life and serves as our placebo (Siegel 2010). We designed the experiment to help isolate the effects of causal thinking separate from the demand for causal thinking by offering a choice between two lotteries. More specifically, deputy ministers chose a high or low probability of receiving one of the two books. The outcome of the lottery served as the instrument for estimating the causal effects of causal thinking controlling for the probability of receiving the metrics or placebo self-help book. The lottery completely determined the random assignment, allowing us to estimate the causal effects

of receiving Mastering 'Metrics for those more likely vs. less likely to comply with the treatment.

The meat of our intervention in the first experiment is intensive training (reading, lectures, writing, discussion, presentations) where we aim to maximize the comprehension, retention, and utilization of the educational materials. Namely, we augmented the book receipt with video lectures from the books' authors, namely, Joshua Angrist and Daniel Siegel, along with competitive writing assignments. We offer deputy ministers both sets of lecture videos and track their click behavior. The econometrician typically cannot observe defiers in an IV framework, so we use the click behavior to proxy for defiers (those who click on the video not assigned to them) and never-takers (those who choose not to click but have training assigned to them). Moving beyond lectures, we designed writing assignments inspired by theory and empirical evidence on the efficacy of social-emotional learning (Heckman and Kautz, 2012). As part of the training program, deputy ministers were assigned to write two essays. The first essay was to summarize every chapter of their assigned book, while the second essay involved discussing how the materials would apply to their career. The essays were graded and rated in a competitive manner. Writers of the top essays were given monetary vouchers and received peer recognition by their colleagues (via commemorative shields, a presentation over Zoom, and a discussion of their essays in a workshop within the treatment arm). Deputy ministers in each treatment group also participated in the Zoom session to present and discuss the lessons and applications of their assigned book in a structured discussion. To summarize, the deputy minister participated in a 20-hour econometrics education program spanning 8 weeks. This program encompassed reading, lectures, writing exercises, structured discussions, and presentations.

The last stage of our experiment was a suite of measurements of deputy ministers. Because we embedded our analyses in administrative data, we had essentially no attrition. We observed a balance on pretreatment quantitative ability as measured from mathematics scores in the entry examinations of the deputy ministers obtained from the Federal Service Commission, an independent government arm that administers the entry examinations of these elite policymakers. Likewise, we observed balance in demographics, pretreatment writing, and interview assessments. This similarity strongly suggests that metrics-trained and untrained ministers are indeed similar, demonstrating our randomization achieved a balance between the

treated and control groups. We analyze all available data shared from the training academy on the ministers' regular policy assessments. These were conducted about 6 months following our workshop and scored deputy ministers in national research methods and policy assessments independently. The latter source of data is particularly important because the regular national policy exams are conducted independently from the experimenter, even independent from the training Academy, thus greatly mitigating potential experimenter demand concerns.

Our first main finding is that training causal thinking shifts policy attitudes. We conducted a survey of policy attitudes on the importance of causal inference several months after the treatment assignment and performed a textual analysis of the high-stakes writing assignment. We find substantial effects. While attitudes on the importance of qualitative evidence are unaffected, treated individuals' beliefs about the importance of quantitative evidence in making policy decisions increase from 35% after reading the book and completing the writing assignment and grows to 50% after attending the lecture, presenting, discussing and participating in the workshop. We also find that deputy ministers randomly assigned to econometrics education have a higher perceived value of causal inference, quantitative data, and randomized control trials. In the writing assignment, treated deputy ministers exhibited a heightened interest in conducting randomized evaluations prior to policy implementation. Their written responses reflected an enhanced comprehension of causal inference and an expressed intention to conduct randomized trials. Moreover, significant increases in scores for standard national research methods and public policy assessments were observed. Treated ministers increased their performance on exams by 0.51 standard deviations. Conversely, in policy evaluations unrelated to econometrics education, the metrics training exhibited no discernible impact. These routine assessments were not prompted by the research team, thus minimizing the possibility of experimenter demand influencing performance improvements.

Our second main result emerges from stated policy decisions of deputy ministers involving deworming, which affects 24% of the world's population (WHO 2021). We provided a signal—an email summary of a recently published randomized evaluation on the long-run impacts of deworming to all deputy ministers (Kremer et al. 2021). We find that treated deputy ministers were twice as likely to demand fiscal support from the Finance Ministry to support the deworming policy. Policy beliefs and hypothetical decisions directly relating to deworming

signal are also affected. First, we elicited initial beliefs about the efficacy of deworming on long-run labor market outcomes. Then, the ministers are asked to choose between implementing a deworming policy versus a policy to build computer labs in schools. After the signal about the impact of deworming, we asked the same deputy ministers about their post-signal beliefs and to make the policy choice again. From this experiment, we observe that only those assigned to receive training in causal thinking showed a shift in their beliefs about the efficacy of deworming: treated ministers assigned to the treatment group updated their beliefs in the direction of the signal. The treated ministers also became more likely to choose deworming as a policy after receiving the RCT evidence signal. The magnitudes are substantial—trained deputy ministers doubled the likelihood of choosing deworming, from 40% to 80%. Notably, this shift occurs only for those ministers who previously believed the impacts of deworming were *lower* than the effects found in the RCT study.

How does econometrics education shift ministers' receptiveness to information? To answer this question, we measured ministers' *stated* willingness-to-pay for three sources of information: RCTs, correlational data, and expert bureaucrat advice. We elicited willingness-to-pay for correlational data and senior bureaucrats' advice because these two alternative sources of information are the status quo that deputy ministers use to inform their policy decisions. We observed that treated deputy ministers were much more willing to spend out-of-pocket and from public funds for RCTs and *less* willing to pay for correlational data. In other words, the treatment did not increase the willingness to seek *any* data and evidence but rather shifted policymakers' beliefs towards the paradigm associated with the credibility revolution. We also find that the effects of metrics training are similar for those who express high versus low demand for econometrics training. This suggests that the initial demand to learn about metrics may not be crucial for policymakers' receptiveness to econometrics education.

Our second experiment works with tax officers, who were randomly assigned a similar econometrics education workshop relative to a placebo macroeconomics workshop. This workshop essentially condensed the 20-hour deputy minister workshop into a 10-hour session: Two three-hour lectures were administered, accompanied by two two-hour self-persuasion exercises. These exercises involved group discussions within each treatment group, all geared towards aiding the integration and application of the lecture's key messages and insights into the

participants' roles as policymakers. A similar group discussion was also employed for the macroeconomics placebo workshop as well. The tax officers were then cross-randomized to receive either a causal signal or a correlational signal regarding the effectiveness of a tax reminder policy. These signals were selected based on the emerging consensus in the literature (as summarized in Abdul Latif Jameel Poverty Action Lab, 2022) that sending tax reminders leads to an increase in tax revenue. Our results reveal a significant impact of econometrics education on tax policy and revenue collection. Specifically, relative to the control, econometrics education training led to a 20% increase in the use of tax reminder policy, with a notable 40% increase in tax collection. To put these findings in perspective, the use of tax reminders increased by 0.25 standard deviations and tax revenue increased by 0.19 standard deviations. Notably, we observe that the impact of econometrics education on the decision to send tax reminders or improve tax collection is evident solely among those who received the causal signal. These findings emphasize how econometrics education can effectively shape tax policy, bolster fiscal capacity, and enhance the generation of tax revenue. A number of robustness tests support these observed effects are valid. First, the distribution of tax collection shows a remarkable shift to the right for the econometrics-trained tax officers relative to the placebo-trained group, suggesting that a large number of tax officers were affected by the econometrics educational workshop. Second, the findings are robust to the inclusion of controls such as the demographic characteristics of the tax officers, consistent with randomization of our treatment assignments. Third, we can trace the pre-treatment collection of tax revenue and pre-treatment sending of tax reminders which we also find to be uncorrelated with the treatment assignments.

Both experiments complement each other in strengthening the evidence that training econometrics impacts the beliefs and decisions of policymakers. The first experiment measures the impact of training on the beliefs and hypothetical decisions of policymakers. It was, however, conducted in conjunction with a companion experiment on training altruism (Mehmood, Naseer, Chen 2023) for the same cohort. The second experiment focuses solely on econometrics training and measures its impact on actual policy and how it relates to state capacity. Further, we show that a lighter version of the original training intervention can also be very successful at shaping actual policy.

Our paper contributes to three key literatures. First, our study pivots the literature on how and why paradigm shifts occur in science (Kuhn 1962) to study its consequences. We studied one of the most prominent schools of thought in science: the credibility revolution (Angrist and Pischke 2010; Pearl and Mackenzie, 2018). We, to the best of our knowledge, are the first to study the causal effects of paradigm shifts using a field experiment with high-stakes decision-makers. Economists, in contrast to philosophers, historians, and sociologists (Kuhn 1962, Shapin 1982, Merton 1973; Foucault 1970) have devoted little attention to paradigm shifts (see Azoulay et al. 2019 for a notable exception). We randomly assigned a book associated with the paradigm and showed its teachings to be highly transmissible via an educational workshop. Mastering 'Metrics provides a concatenation of the school of thought associated with the credibility revolution and provides, in five short chapters, a set of principles for policymakers to abide by. This highlights how sparse thinking and parsimony may be important for influencing human thinking (Gabaix 2014).

Second, we contribute to the literature on state capacity and public sector effectiveness (Besley and Persson, 2010; Acemoglu et al., 2015; Khan et al. 2016; Finan et al., 2017, Slemrod, 2019; Besley et al., 2022). Our experiment shows the potential for building state capacity, particularly fiscal capacity by training public officials. Earlier work showed that training high-stakes decision-makers may impact the private sector by increasing profits and sales on average by 5–10 percent (McKenzie 2021), and economics training impacts high-stakes decisions of policymakers and explains up to 30% of the recent shift towards economic conservatism in the American judiciary (Ash et al. 2021). Relatedly, our study on econometrics education adds to the expansive literature on economics and financial literacy (Lusardi and Mitchell 2014) and financial literacy explains up to 40% of inequality in end-of-life wealth (Lusardi, Michaud, and Mitchell 2017). A recent experiment of eight hours of financial literacy training found impacts on the economic preferences of adolescents (Sutter, Weyland, Untertrifaller, Froitzheim 2020). We, however, study the impact of as little as 10 hours of econometric literacy training — in causal thinking — on the attitudes and behavior of adults who make policy decisions. Policymakers have been pointing out various reasons for poor policy-making, such as lack of skills, time, or incentives to use evidence (Lugo-Gil et al. 2019; Krämer et al. 2021). Against this backdrop, our research concentrates on crafting training materials tailored to a level where both elite and frontline policymakers can grasp technical

concepts. Our research also complements other experiments training altruism (Mehmood, Naseer and Chen, 2021) and artifical intelligence (Mehmood, Naseer and Chen., 2022), to junior and senior civil servants that show the personnel economics of the state can be malleable, proposing a fourth vector alongside selection, monitoring, and incentives as highlighted in Finan et al. (2017) that have been shown to be critical for economic development. Together, this work seeks to enhance our understanding of pathways for imparting technical skills and to design and test interventions that encourage the utilization of evidence by policymakers.

Third, we contribute to the new and vibrant literature on behavioral economics of development and growth (Kremer, Rao, Schilbach 2019). We show that a key factor in demand for and responsiveness to rigorous evidence on the effects of policies is an understanding and appreciation of causal evidence. This, in turn, may promote the implementation of good policies that might otherwise have high rates of return for economic growth. By shaping deputy ministers' and tax officers' causal thinking with a scalable basic econometrics training and measuring its consequences, we show the key role that developing causal thinking plays when evaluating evidence. In our experiment, policymakers without training in causal inference were unresponsive to causal evidence. While many training studies focus on lay populations, we examine high-stakes decision-makers like central bankers (Malmendier et al., 2017) and judges (Chen et al., 2016). We trained deputy ministers' and tax officers' causal thinking and estimated impact on attitudes and subsequent demand for evidence and policy choices. We *cautiously* interpret our findings as suggesting that causal thinking not only increases responsiveness to causal evidence but *may* also correct for some mistakes in the belief-updating process in response to evidence.

The rest of the paper is organized as follows. In Section II, we present the background information that contextualizes the experiments. Section III delves into the specifics of the experimental setup tailored for the deputy ministers' study. The tax officers' experiment is elaborated upon in Section IV, while Section V introduces a sequence of sensitivity tests. A final section concludes.

## II. Background

Our investigation delved into the impact of ideas associated with the credibility revolution on both deputy ministers and tax officers. We executed two separate experiments within the realm of Pakistan's civil bureaucracy. In both scenarios, we utilized a concise presentation of the credibility revolution through "Mastering ’Metrics: The Path from Cause to Effect" (Angrist and Pischke 2014), as the means to impart econometrics education.

In the first randomized trial, our focus was on elite bureaucrats, specifically deputy ministers in Pakistan. Notably, the Government of Pakistan deems these deputy ministers as pivotal elements driving the entirety of the state's machinery (Government of Pakistan, 2019). In India, their counterparts, also a legacy of British Colonial governance, are referred to as the "steel frame" of the State. To conduct this first experiment, we partnered with one of Pakistan's elite civil service training academies (The Civil Service Academy), which grooms top bureaucrats as they enter service. Annually, out of 15,000 exam takers, only about 200 individuals are chosen to become these elite officers and are trained at this Academy. These individuals play pivotal roles, advising top government officials and taking part in policy formulation and execution, ultimately, advising the Prime Minister, President and key public officials. They undergo rigorous training at the academy, participating in multiple workshops and assessments. During our study, we introduced a "mastering metrics" educational workshop for these deputy ministers, which was integrated into their regular program at the Academy. This workshop was delivered both through Zoom and in-person sessions. We had access to ministers’ policy evaluations and national policy exams conducted in other workshops that were running concurrently with our program.

In our second experiment, we enlisted around 10% of the total tax officers (inspectors) in Punjab, which is Pakistan's largest province, resulting in a cohort of 310 participants. These tax officers at the forefront play a vital role in tax collection and compliance, ensuring timely payments, penalizing defaulters, and aiding in the rollout of tax policies. We arranged in-person training sessions for them at Punjab's Tax and Customs Academy. During these sessions, these tax officers were (1) randomly designated for econometrics training or a placebo macroeconomic workshop (2) further cross-randomized to either causal or correlational studies on the impact of tax policy. The tax officers, similar to the self-persuasion writing exercises of the deputy ministers, also engaged in 2 two-hour-long structured discussions after lectures. During these

discussions, they were prompted to succinctly encapsulate the key takeaways from the lecture and contemplate how they could effectively implement the workshop's insights into their policy practices. Additional specifics, including the discussion of the precise training content, can be found in the subsequent sections.

## III. Deputy Ministers Experiment on Beliefs, Willingness-to-Pay and Policy Assessments

### A. *Treatment Design Details: Deputy Ministers*

*October: Baseline survey and book choice.*— We conducted a baseline survey and asked the participants to choose one of two books (1) Mastering 'Metrics: The Path from Cause to Effect by Joshua Angrist and Jörn-Steffen Pischke or (2) Mindsight: The New Science of Personal Transformation by Daniel J. Siegel for all 213 public officers who qualified for service in a single year of examination. The first book is an accessible introduction to the fundamental problem in causal identification and summarizes key concepts associated with the credibility revolution. These include RCT as an experimental ideal, regressions as a comparison of means, instrumental variables, difference-in-differences, and regression discontinuity designs with particular focus on public policy applications. The book is written for undergraduates and is particularly appropriate for our policymakers since all of them at least hold a bachelor's degree. The second book is a popular self-help book emphasizing "personal transformation" and serves as our placebo. For the table of contents of both books, see Figure B1 of Appendix B.

*November: Assignment of treatment.*— In November, the director of the Civil Service Academy tasked all deputy ministers with an assignment linked to the designated book. Subsequently, the deputy ministers within the cohort promptly acknowledged their commitment to completing the assignment by the stipulated deadline. The close cooperation of the director and the Academy's staff translated to an impressive uptake of approximately 90% for our intervention. We randomly assigned the book through a lottery where the person who chose either of the books had a certain probability of actually being assigned that book.[2] That is, the

[2] Specifically, a person choosing the metrics book had a 60% probability of being randomly assigned the metrics training, while the person choosing the placebo self-help book had a 85% probability of being randomly assigned the placebo training.

participants were randomly assigned either Metrics or self-help books, conditional on their choice of lottery. The books were delivered in person. They were then requested to complete two open-ended assignments related to the contents of the respective books:

> "*Main Task 1: After reading the assigned book, we request you provide a chapter-by-chapter summary of the whole book of around 1500 words (+/-100 words).*
>
> *Main Task 2: After reading the assigned book, we request you provide an analysis of how you would apply the lessons learned from the book in your job. This again should be around 1500 words (+/-100 words)*."

All assignments were submitted by 10 December (the set deadline). The full detailed transcript of the message by the director detailing their assignment tasks can be found in Table A1 of Appendix A.

*March: Attitude Survey, lecture, presentation, discussion and workshop.*— On 10 March of the following year, in collaboration with the Academy, we organized two Zoom sessions, one for a randomly assigned metrics group and the other for the placebo self-help group. First, there was an 'endline' survey, i.e., before the lectures and discussion, where we elicited participants' attitudes towards quantitative and qualitative evidence, randomized evaluations and causal inference. This gives us outcomes to assess the impact of metrics books and writing assignment tasks 4 months following the assignment of the books (partial treatment). These writing assignments were high-stakes not only because the overall grade became part of the permanent record at the Academy and may influence the ministers' future career trajectories, but also because we distributed commemorative shields, often accompanied by peer recognition, and monetary gift vouchers to the top 6 performers. After conducting the endline survey on attitudes, we announced the first three positions for both groups and distributed the commemorative shields and gift vouchers to a luxury departmental store. The 1st position received a monetary voucher of USD 150, the 2nd position received a USD 100 voucher, and the 3rd position received a USD 80 voucher. The placebo group also received the vouchers and hence we had 6

winners. These winners also gave a 30-minute presentation summarizing key lessons of the respective books and how the training will inform their policymaking within the treatment arm. This was followed by 30-minute video lectures delivered by the authors of the books to the respective randomly assigned groups. The group assigned Mastering 'Metrics attended the video lecture by Joshua Angrist and the group assigned Mindsight attended the video lecture by Daniel Siegel. A structured discussion of 30 minutes for both arms followed. In particular, we asked participants the following two questions: (a) What do you think is the main point of the lecture? (b) How can you apply the concepts learned in this lecture to your job? In the end, this part of our engagement with the ministers concluded by asking the same questions on attitudes towards quantitative and qualitative evidence, randomized evaluations, and causal inference. This allowed us to assess the short-run impact of our complete metrics training, i.e., essays summarizing the book, essays applying the lessons to policy, attending the video lecture, receiving commemorative shields and gift vouchers, presentation and discussion of key lessons learned. Table B2 in the Appendix B presents screenshots of commemorative shields and gift vouchers distributed to the deputy ministers.

*May: Initial Beliefs, Post-Signal Beliefs, Willingness-to-Pay and Project Choice.*— On 16 May, about 6 months following the book assignments, we elicited policymakers' beliefs on the impact of a policy and asked them to make policy choices. Specifically, we elicited their beliefs on the impact of deworming of children in schools on earnings 20 years later. We also elicited a response on a policy choice between choosing to implement a deworming initiative or an ICT initiative to establish computer labs in a small city of Kuchlak in Balochistan. The decision between the deworming and ICT policies was used because it was an actual policy choice faced by similarly ranked public officials at the time. In particular, we asked them to choose either to implement a computer lab policy that involved installing a computer lab in each school of Kuchlak or implement a deworming policy where they launched deworming campaigns in all schools of the same town.[3] To minimize the experimental demand effect, we nudged policymakers to choose an alternative computer lab project. That is, before the project choice the policymakers see the following prompt:

3 We also inform them that the direct costs of implementation of both projects is roughly the same.

*We suggest that you implement the computer lab project given IT is the future.*

We elicit the policymakers' stated willingness-to-pay (WTP) from both private and public funds for three pieces of information: (1) a RCT assessing the impact of deworming on earnings; (2) correlational data showing a relationship between deworming and earnings in schools with and without a deworming program; (3) expert advice from a senior bureaucrat on the impact of deworming on wages. We then reveal a "signal" that provides experimental evidence of deworming on hourly wages from a 20-year-long study by Kremer et al. (2021). Specifically, the deputy ministers are presented the following prompt which was also sent to all deputy ministers through their official emails addresses the following day:

*Recent randomized evaluation finds deworming impacts on economic outcomes up to 20 years later. Individuals who received deworming experience up to 3 additional years of schooling, 14% increases in consumption expenditure, 13% increases in hourly earnings, 9% in non-agricultural work hours (Source: PNAS, 2021).*

Finally, we collected post-signal beliefs on deworming's impact and updated policy choices between the deworming and computer policies following this signal.[4] For more details on the willingness-to-pay transcripts of questions and project choice, see Table A3 in Appendix A. For the whole set-up and summary of the complete experimental design, see Figure B2 in Appendix B.

*COVID-19, Ongoing Interventions, and Consequences for Our Design.*— At the Academy, the officers typically reside at the Academy in Lahore for the entire period. Nevertheless, the cohort we studied was instructed to remain in their home due to the COVID-19 pandemic. The training, therefore, took place online. So, the geographical dispersion of the ministers due to the pandemic at the time of the training, and the non-shareability of the link

[4] Note, it was not possible for ministers to go back in the survey since it was administered online.

reduced treatment spillovers. However, it should be noted that treatment spillovers would likely suggest that our estimates are underestimated.

The intervention was also conducted in conjunction with a companion experiment for the same cohort on training altruism (Mehmood, Naseer, Chen 2021). Because the randomization was conducted as a factorial design, alternative factors that affect attitudes and beliefs should be uncorrelated with the intervention studied here. The placebo book, *Mindsight*, should be viewed as potentially affecting a vector of thinking that is uncorrelated with the econometrics training intervention. Addressing this potential confound, in the second experiment is conducted in the absence of any altruism training.

B. *Data and Empirical Specification: Deputy Ministers*

*Data.*— The data was collected from about 200 deputy ministers entering service in a single year. The entry year is anonymized to protect their identity. The close collaboration with the Academy implied we had about 90% take-up of our intervention. The administrative data on individual policymakers' characteristics were obtained from the administrative records of the Academy. We used this in our balance check over demographics and as control variables in our regressions. The outcomes of field visits to orphanages, volunteering at low-income schools, teamwork, national public policy and research methods assessments were also obtained from the Academy. The pretreatment mathematics, written and interview assessment scores of the ministers were obtained from the Service Commission that administers the entry examinations for these elite policymakers.[5] Data on WTP, attitudes and beliefs were collected by our research team under the auspices of the Federal Government of Pakistan.

*Outcome Variable on Policy Decision*.— We obtain data for fiscal support or budgetary requests of deputy ministers. These annual budgetary requests for fiscal support, made months after graduation from the Academy, are independent from potential experimental demand effects of the experimenter and the Academy. *Empirical Specification.*— The impact of metrics training can be evaluated in a simple regression framework. For each individual-level outcome, the estimation equation is:

[5] The FPSC is a statutory body of the Government of Pakistan, constituted at the time of independence in 1947. It obtains its jurisdiction from the Constitution of Pakistan and its responsibilities include recruiting elite policy advisors and administering their entry examinations and assessments.

$$Y_i = \alpha + \beta\, Metrics\ Assigned_i + X_i'\mu + \epsilon_i \quad (1)$$

where $Y_i$ is the respective outcome for the policymaker $i$, this includes attitudes, assessment scores, WTP and policy choices. $Metrics\ Assigned_i$ is a dummy equal to one if the policymaker is assigned to metrics training. $X_i$ is a vector of individual-level controls, which includes written test scores, interview test scores, gender, birth in political capitals, asset ownership, income before joining service, age, prior education, foreign visits and occupational designation dummies. Importantly, the list of explanatory variables also includes our randomization strata, and metrics chosen or demanded. This is a dummy variable equal to one if the policymaker chooses the metrics book that we directly control for. We cluster standard errors at the individual level since that is our level of randomization. β is our main coefficient of interest and estimates the causal effect of metrics training conditional on the policymakers choosing metrics.

*Balance and Attrition.* — Table 1 reports the results on the balance check on those randomly assigned to metrics treatment. Differences across treatment groups and placebo are small in magnitude, and statistically insignificant, suggesting that the randomization was effective at creating balance. Salient to note are the policymakers' pretreatment written, interview and mathematics assessments (Table 1, Columns 9, 10 and 11). Since the policymakers obtained these scores *before* the metrics training, the similarity of test scores across written and interview assessments suggest that those assigned the metrics training are likely balanced in their academic and interpersonal ability. Important to note is the balance on pretreatment scores on the mathematics assessment that suggests our sample is also balanced in quantitative ability. The close collaboration with the training Academy and the director resulted in our intervention to have take-up of about 90%, there is, however, a possibility of differential attrition with respect to our treatment. However, this is unlikely because in Table B3 of Appendix B, we find that there is essentially no significant effect of metrics training on attrition.

### C. *Main results: Deputy Ministers*

*Treatment Effect on Policy Assessments*–.The training Academy provided administrative data on policy assessments of the deputy ministers in a high-stakes setting. These assessments in part determine the minister's transfers and postings. It includes policymakers' performance in the national public policy and research method assessments. We obtained scores of these policymakers for three regular assessments that together comprise 15% of their overall training score and become part of their official record. All these "executive courses" style workshops were 2-hour twice weekly sessions with a written assessment at the end of 8 weeks. The first workshop was called "Public Goods and Publicly Provided Private Goods" colloquially referred to as the "Public Policy" assessment at the Academy. This workshop emphasized and evaluated the policymakers on case studies and analysis of past policy decisions of similarly ranked policymakers. The course content covered scenarios that apply concepts of public goods, externalities and the use of data in policymaking in real problems that are currently being faced in the field by similarly ranked ministers. The second was a research methods workshop. Its content included an introduction to hypothesis testing, multivariate regressions with several applications and case studies. Salient to note from the research methods course is that randomized evaluations are also touched on (in passing) in this course. Finally, the workshop on "Teams & Group Decisions" was a policy simulation workshop that included assessments on teamwork and group decision-making that assesses ministers' policy skills to work together in a group. During the simulation, deputy ministers were assessed by a panel of experts. A typical scenario question was as follows:

> "*The Prime Minister wants you to devote more resources to his security detail, while the Chief Minister wants you to aid in the flood relief efforts. How would you organize your team? What decisions will you take? Please detail the exact steps?*" (FPSC, 2021).

The responses were scored by a panel of experts (former supreme court judges, prominent academics, and former senior deputy ministers) and the assessment was high-stakes since it determined their future career trajectories. Table 2 reports the final assessment scores—standardized to mean zero and standard deviation one—6 months after the metrics

training. We observe a large impact on treated policymakers' performance in national research methods and public policy assessments. The treated policymakers score about 0.5 SD higher in national public policy and 0.8 SD higher in the research methods assessments. This suggests a substantial impact of our treatment on their regular policy assessments that take place at the Academy, one that is not solicited by the research team. Scores on teamwork assessments, however, are unaffected (Table 2, Columns 5 and 6), suggesting that the metrics training did not crowd out quality of team decisions, a critical soft skill in effective policymaking (Deming and Weidmann, 2021).

*Treatment Effect shifts beliefs*—.In May 2021, 6 months following the metrics training workshop, we elicit beliefs on the effect of deworming from all policymakers. This was followed by a "signal" on causal evidence regarding the effect of deworming on various outcomes including income. Specifically, the following signal was revealed:

> *Recent randomized evaluation finds deworming impacts on economic outcomes up to 20 years later. Individuals who received deworming experience up to 3 additional years of schooling, 14% increases in consumption expenditure, 13% increases in hourly earnings, 9% in non-agricultural work hours (Source: PNAS, 2021).*

Interestingly, we found that the policymakers underestimated the long-run impact of deworming relative to impact from the randomized evaluation results presented in the signal. Figure 2 reports distributions of initial and post-signal beliefs for placebo and metrics assigned groups. The ministers' initial beliefs on the impact of deworming on long-run income is about 5% (for both treated and placebo groups). Nevertheless, as can be observed from Figure 2, the group assigned the mastering metrics training is substantially more likely to respond to the signal by shifting their prior beliefs towards the signal value of 13%, while the placebo trained group is unaffected by the signal. This strongly suggests that the metrics training shifted policymakers' beliefs on the long-run impact of deworming.

To investigate individual-level shifts in beliefs, we compare the distribution in belief updating following the signal with policymakers' initial beliefs. Figure B8 reports this shift in

beliefs after receiving the signal for the two different groups, the treatment group in the diagonal solid line and the placebo group in the more horizontal dotted line. The top panel shows the treatment and control groups for those who chose the metrics book and bottom panel shows the same for those who chose the placebo book. Three features of the graphs are particularly noteworthy. First, the results are similar regardless of the book choice. Second, the shift in beliefs for the treatment group falls along a diagonal that intersects the x-axis roughly at the signal. That is, treated individuals whose prior beliefs are 13% update very little on average. Moving to the left along the x-axis, treated individuals update their beliefs in a positive direction, though do not necessarily jump to the signal. Likewise, moving along the right, treated individuals update their beliefs in a negative direction. Third, in contrast, the placebo individuals almost never update their beliefs. Some placebo individuals even do not react to the signals in a Bayesian manner. This can be seen from two features of the data. First, placebo individuals whose beliefs are below the signal occasionally update their beliefs in a direction opposite of the signal. Second, a moderate number of placebo individuals whose initial beliefs are near the signal seem to overreact to the signal by updating past the signal, i.e., in a positive or negative direction by an amount almost equal to the original signal in absolute value. This finding suggests that in the absence of econometrics training, policymakers may misinterpret quantitative evidence.

*Effect of Metrics Training on Stated WTP.* — In May 2021, we also elicited policymakers' demand for evidence. Specifically, we elicited the stated WTP from both private and public funds for three pieces of information for: (1) results from a RCT on the impact of deworming on long-run income; (2) correlational data on incomes of schools with and without a deworming program; (3) advice from senior public officials on the impact of deworming policy. The latter two choices are status quo sources of information available to the ministers. In Table B6 (Column 1), we find that metrics trained policymakers' stated willingness-to-pay is about PKR 2000 (USD 13) from their own pocket (private funds) for causal evidence. This is equivalent to about 1% of their monthly salary. It is, however, unclear whether these metrics trained policymakers would also state they will pay for causal evidence from public funds, especially when their budget is constrained and other available sources of information, such as correlational data and advice from senior officials are readily available. Therefore, in Column 4 of Table B6,

we elicit WTP for the same information from public funds.[6] The effects from public funds are substantially larger and statistically significant with metrics assigned policymakers willing to pay about 1400000 (USD 8500) for the information from randomized evaluation. This is about twice as the stated WTP for the placebo group. These results can also be observed from Figure B4 where we plot the distributions of WTP from public and private funds: metrics assigned policymakers are significantly more likely to state they will pay large amounts from personal and public funds for randomized evaluations than those assigned the placebo training. Interestingly, however, we observe that metrics-trained policymakers *decrease* their stated WTP for correlational data. This finding is consistent with metrics trained ministers becoming closer to the paradigm of credibility revolution that discounts simple correlations over well-identified studies. In particular, we find that those assigned the metrics training pay about PKR 1000 (USD 6.40) less from their private funds for correlational data comparing incomes of pupils in schools with and without deworming relative to those assigned the placebo training, and about PKR 55000 (USD 350) less for this information from public funds (Table B6, Columns 2 and 5, respectively). This is equivalent to a 50% decrease in WTP for correlational data from private funds and a 33% decrease from public funds. Figure B5 reports the distributions of WTP for correlational data for personal and public funds in Panel A and B, respectively. We observe a substantially smaller fraction of metrics assigned policymakers willing to pay large amounts for correlational data relative to the placebo group. Finally, we find that metrics training has essentially no effect on policymakers' stated WTP for advice of senior bureaucrats, another competing source of information available to the ministers. These results are reported in Columns 3 and 6 of Table 6 for stated private and public WTP, respectively. We observe that there is no statistically significant difference between WTP for advice from senior bureaucrats among the metrics trained and placebo group policymakers. In Figure B6, we plot the distributions of WTP for advice from senior bureaucrats. Both these distributions are very similar across metrics assigned and placebo policymakers—for both personal and public finances—suggesting that our treatment did not impact WTP for policy advice from senior bureaucrats.

[6] We measure both public and private WTP as Hjort et al. (2021, p. 4) note an important caveat in their study that "WTP measure is rather artificial, and comes out of the policy-maker's private budget, rather than the likely more-relevant municipal budget, which may have other higher-value uses." (Hjort et al., 2021, p. 24).

*Treatment effects on importance of quantitative analysis in policymaking.* — In Table B4, we present the effect of metrics training on attitudes about quantitative evidence 4 months after the training. The dependent variable in this table is a rating from a scale of 1 to 5, with 1 being not important at all and 5 being very important, in response to the question "How important do you think quantitative analysis is in public policy making?" In the first column of the table, we measure the attitudes of policymakers just before the lecture, presentation and discussion of the material related to the assigned book—i.e., we measure the causal effect of partial training via book assignment of summarizing and application of concepts to policy. We find that policymakers treated with this training rate the importance of quantitative evidence in policymaking by about an additional point. This is a substantial effect and equivalent to a nearly 35% increase over the average rating of the placebo group. In Column 2 of Table B4, we present results of our full training where policymakers attend respective video lectures from the authors of the book, receive a gift voucher, and commemorative shields, present the key lessons of the respective book, and partake in a structured discussion of the material in a workshop (full metrics training). The evidence is consistent with metrics training being reinforced: effect sizes increase by about 50% and are statistically different from those obtained prior to the reinforcement training. In particular, full metrics training (book assignments, lectures, presentations and discussions) increases rating on the importance of quantitative analysis in policymaking by about 1.5 points on a 5-point scale. This is a 50% increase in ratings over the placebo mean. Following this, we conducted a falsification test on a similar question that helps mitigate potential concerns that individuals rate *any* evidence higher regardless of its quantitative nature. The dependent variable in this case is the rating on the question "How important do you think *qualitative* analysis is in public policy making?" Columns 3 and 4 of Table B4 report these results. The metrics training (partial or full) has no significant effect on policymakers' beliefs about qualitative evidence. The policymakers' beliefs on the importance of qualitative evidence remain unaffected. This indicates that our training is unlikely to come at the expense of perceived importance of qualitative evidence in policymaking (which may be important in some situations when policymakers must operate under strict time, ethical or budget constraints rendering randomized evaluations unfeasible). For raw comparison of means across treatment and placebo groups, see Figure B7 in Appendix B. In Table B4, we also report results of metrics training on

attitudes about the importance of RCTs in policymaking with partial and full training. The dependent variable is constructed based on the following policy scenario:

> "*You are in charge of assigning people to a public policy program and before rolling it out, you want to learn if the policy is effective, what you would do?*"

One of the options is to "Run a randomized control trial", while other options are unrelated or inconsistent with the main message of the book, such as "Compare two groups of people who had previously benefited most from the policy with those that did not?" and "Survey if there is demand for the policy". These options appear in random order. The dependent variable takes the value of one if the policymaker answered "Run a randomized trial" and zero for all other options. The results of estimating equation (1) with this dependent variable is reported in Table B4 (Columns 5 and 6). We observe that the group assigned the metrics book tasks (partial training) is about 15 percentage points more likely to choose randomized evaluation before rolling out a public policy relative to the placebo group; with full training this effect increases to about 20 percentage points or a 55% increase over the placebo mean. Taken together, these results suggest that months after the training, treated policymakers' perceived importance of quantitative evidence and randomized evaluations increased, while we observe no effect on importance of qualitative evidence.

*Why did the policymakers demand randomized evaluations?—* What explains policymakers attaching greater importance to quantitative analysis and randomized evaluations? Here we present some evidence that the results may be explained by the fact that policymakers learn about causal inference and selection issues. In the last two columns of Table B4, we elicit beliefs on *why* randomized evaluations are important for policymaking. Specifically, we continue with the earlier question and ask: "Continuing with the previous example, why does the previous answer make sense?" One of the options to the above question is "Because comparisons in a RCT are apples to apples comparisons", while other options are unrelated to use of randomization to circumvent selection issues. For instance, "People's feelings are an important determinant whether the public policy will work", "Survey methods are known to produce causal effects", "Comparing two groups of non-randomly selected people allows us to infer causality"

are the other options. The dependent variable takes the value of one if the policymaker chooses "Because comparisons in a RCT are apples to apples comparisons" and zero for all other options. Columns 7 and 8 of Table B4 report these results with partial and full metrics training. Ministers assigned metrics training are 15 percentage points more likely to answer that randomized evaluations are "Apple to apple comparisons" suggesting that they understand that random assignment of subjects in the control and treatment groups solves the selection problem by "comparing apples to apples". This is one likely explanation why our treated group may have higher perceived importance of quantitative evidence and randomized evaluations. In Figure 1, we report all results of Table B4 but standardized to mean zero and standard deviation one. This includes beliefs of policymakers on quantitative and qualitative evidence, as well as importance of RCTs in policymaking. The coefficient estimates and confidence intervals associated with the metrics assigned variable are reported in the figure with equation (1) estimated with all individual-level baseline controls. The group assigned metrics training see about a 0.85–1.32 standard deviation increase in rating assigned to quantitative evidence, a 0.33–0.44 standard deviation increase in request for randomized trial to evaluate effectiveness of public policies, and about 0.30 standard deviation increase in answering that randomized evaluations allow for apple to apple comparisons relative to the placebo group. We find no effect of metrics training on beliefs about qualitative evidence, however. These results are consistent with textual analysis of the ministers' high-stakes assignments. Though, suggestive, the analysis of their writings suggests that the metrics assigned group likely learned many causal inference concepts. Specifically, in Figure B3 we observe that the treated group witnessed a large increase in use of the following phrases: "Causal inference is important", "Correlation is not causation", "Quantitative Evidence" and "Observational studies are not apple to apple comparisons". The metrics training appears to affect policymakers' attitudes towards the paradigm associated with the credibility revolution.

## IV. Tax Officers Experiment

In this section, we describe the tax officers experiment that investigates (i) whether econometrics training impacts Pakistani tax officers' tax reminder "notification" policy, (ii) whether the training affects tax revenue collection, and (iii) whether the kind of information tax officers receives (causal or correlational) impacts their behavior. All outcomes and hypotheses of

the second experiment are pre-specified and the pre-analysis plan can be found at the pre-registry (RCT ID: AEARCTR-0010583).

### A. *Treatment Design Details: Tax Officers*

The workshop for tax officers consisted of two three-hour lectures and two two-hour structured discussions that effectively condensed the material covered by the deputy ministers into a 10-hour session. The workshop on econometrics for tax officers was similar except for three major differences. First, the participants were not given the option to choose one of two workshops and then randomly assigned the workshop, instead, they were randomly assigned into two workshops directly, each comprising 2 three-hour lectures of metrics or macroeconomics. Following the lectures, a self-persuasion exercise involved a group discussion within each treatment arm and case study to apply lessons of the workshop to their career. Second, in addition to the random assignment of the workshop, each group was further randomly assigned to receive either a correlational or causal handout–information on the efficacy of tax reminder policy on tax collection. Third, the experiment with tax officers had pre-specified exact specifications and two pre-defined outcomes in the pre-analysis plan: sending tax reminders and tax revenue collected.

*Assignment of treatment.*–The econometrics workshop delivered to tax officers is also rooted in the principles outlined in the 'Mastering Metrics' book, with a particular emphasis on causal identification. It provides a summary of key concepts related to the credibility revolution. One of the central areas of focus was the concept of random assignment, its significance in causal inference, and how experimental studies diverge from "naive" correlational comparisons when interpreting causal effects. This workshop distilled the essence of the deputy minister workshop into two three-hour lectures and two two-hour structured discussions, effectively compacting the material that the deputy ministers had covered in a 10-hour workshop. Specifically, tax officers were individually randomly assigned to in-person classes at the Tax and Customs' Academy in Punjab. The macroeconomics workshop was of similar length and medium, where the lecturer discussed some macroeconomic facts and trends in Pakistan and held similar structured discussion on this content. A snapshot of the slides for the two lectures, along with the complete

course material, is provided in Figure B9 of Appendix B. Screenshots of these in-person lectures are presented as Figure B10 of Appendix B.

B. *Data, Outcomes and Empirical Specification: Deputy Ministers*

*Data.*—We were granted access to tax officers as they entered their training program. The date of training and cities these officers are stationed at have been anonymized to ensure their anonymity. Our strong partnership with the institute enabled us to effectively match them with administrative data pertaining to individual traits and pre-treatment policy outcomes, such as previous year's policy and tax collection figures. We conducted a balance test on these characteristics and pre-treatment outcome variables, as demonstrated in Table 3, and find the treated and control officers as statistically similar.

*Outcome Variable on Tax Reminders and Revenue*.— We pre-registered two key outcomes for the experiment with tax miniders (AEA RCT Registry ID: AEARCTR-0010583). The first oucome varaible is a dummy variable indicating whether the officer dispatched a "tax notification" to prompt citizens and businesses to settle their tax obligations. This information is sourced from the training institute. The second variable is the individual-level tax collected during the fiscal year following the treatment. This data is acquired from the tax office and cross-referenced with information from the Federal Bureau of Statistics (FBR).

*Causal vs Correlational Signal*.— The “causal signal” is a binary variable that becomes active when tax officers are assigned a handout summarizing the contents of the article titled "Behavioral Interventions in Tax Compliance: Evidence from Guatemala," authored by Kettle, S., Hernandez, M., Ruda, S., and Sanderson, M.A. (2016). In contrast, the binary variable takes the value of zero if the tax officer is assigned the correlational study on tax reminder policy authored by Budiman, I. and Inayati, I. (2021), titled "Effect of Notice of Tax Warning, Notice of Tax Collection, and Tax Education Programs on Tax Compliance in West Sumatera and Jambi," published in Publik (Jurnal Ilmu Administrasi), volume 10, issue 1, pages 45-63. For precise details of the handouts assigned to the tax officers, refer to Figure B11 and B12 in the Appendix. The causal signal is randomly administered across treatment and control groups. This is evidenced from the balance across demographic traits of participants who were assigned to treatment and received the causal signal in Table B14.

*Empirical Specification.*— The evaluation of the impact of metrics training can be carried out within a straightforward regression framework. For each individual-level tax officers' outcome, the estimation equation takes the form:

$$Y_i = \alpha + \theta\, Econometrics\ Training_i + X_i'\mu + \epsilon_i \quad (2)$$

where $Y_i$ represents either tax revenue in millions of Pakistani Rupees (PKR) or a dummy variable that switches on if officer *i* sends a tax reminder notification. $Econometrics\ Training_i$ is a dummy equal to one if the tax officer is assigned to metrics training and zero if assiend the placebo maroeconomic workshop. $X_i$ is a vector of individual-level controls provided in balance check in Table 3. We cluster standard errors at the tax region level. Figure 4 depicts the 78 tax regions in our study. $\theta$ is our primary coefficient of interest, in this second experiment, estimating the causal impact of metrics training on tax reminder policy and tax collection. We also cross-randomized the tax officers to receive causal versus correlational signals within the metrics training and placebo workshops. These signals are depicted in Figure B11 and Figure B12 of the Appendix B report, illustrating experimental versus observational studies showing tax reminder policy increase tax collection.[7] These handouts were randomly assigned within each treatment arm for the tax officers. We have pre-specified the following two hypotheses that we aim to test through this cross-randomization:

"*H5: Econometrics training will increase more the adoption of sending tax reminder for those officers that received the signal study to send tax reminders with RCT evidence (Experiment 2).*

[7] Causal Handout Distributed among Tax Officers uses the following study "Behavioral Interventions in Tax Compliance: Evidence from Guatemala" by Kettle, S., Hernandez, M., Ruda, S., and Sanderson, M.A. (2016). This paper shows that sending tax reminders increases tax collections in a field experiment. Correlational Handout Distributed among Tax Officers uses the following study: "Effect of Notice of Tax Warning, Notice of Tax Collection, and Tax Education Programs on Tax Compliance in West Sumatera and Jambi" by Budiman, I., and Inayati, I. (2021). This paper shows that sending tax reminders increases tax collections in an observational panel data study.

*H6: Econometrics training will not impact or decrease the adoption of sending tax reminder for those officers that received the signal study to send tax reminders with correlational evidence (Experiment 2).*"

Figure 4 offers an illustrative map depicting the fraction of metrics-assigned tax officers in each district alongside the corresponding tax collection. This visualization reveals the geographic correlation between tax officer econometrics workshop assignments and tax collection levels in their respective districts. The assignment pattern of tax officers into the econometrics workshop and the placebo workshop also appears to be randomly distributed throughout Punjab, the setting of the tax officers' experiment. Furthermore, relative to placebo assigned officers, regions where metrics-trained tax officers were stationed seem to exhibit a more pronounced, darker shade, indicative of higher tax collections.

## C. *Main results: Tax Officers*

*Treatment effects on Tax Reminders and Tax Collection.* —Table 4 presents the outcomes of metrics training. As indicated in Columns 1 and 2 of Table 4, tax officers assigned to the metrics group are approximately 12 percentage-points more likely to send tax reminder notifications to citizens. This increase corresponds to a 20% surge in policy adoption compared to the placebo macroeconomic assigned group. Similarly, Columns 3 and 4 of Table 4 correspond with a substantial body of existing evidence summarized in the Abdul Latif Jameel Poverty Action Lab (2022) policy paper that concludes that tax reminders increase tax revenue generation. Our findings, significant at the 10% level, suggest a substantial increase in tax revenue collection attributed to metrics training—approximately PKR 25 million (USD 85,000) or a 40% elevation in revenue over the mean dependent variable. The enhanced performance seen in deputy ministers who underwent metrics training, resulting in a notable improvement on public policy exam scores, is similarly reflected in tax officers' outcomes. Tax officers who received the metrics training displayed a 0.25 standard deviation increase in sending tax reminders and a 0.19 standard deviation improvement in the amount of tax revenue collected.[8]

[8] The MDE for these effects are 0.47, 0.31, and 0.30 respectively. To put these MDEs in perspective, other interventions in training of civil servants have MDEs of 0.11 to 0.26 for the Socratic method, a 5 minute-video, on course outcomes and a behavioral measure of motivated reasoning (Chen, Ramos-Maqueda, Silveira 2023) or 0.4

The MDE for the policy exam effects, sending of tax reminders, and amount of tax collected are 0.47, 0.31, and 0.30 respectively.[9] To put these MDEs in perspective, other interventions in training of civil servants have MDEs of 0.11 to 0.26 for the Socratic method, a 5 minute-video, on course outcomes and a behavioral measure of motivated reasoning (Chen, Ramos-Maqueda, Silveira 2023) or 0.4 for both teacher vaccination and absenteeism in a role model vaccination message campaign (Mehmood, Naseer, Chen 2023). Examining the results visually, the metrics-trained tax officers' revenue distribution experiences a shift toward higher values compared to placebo tax officers (refer to Figure 5 for an estimate of probability density functions).[10] Next, we further unpack the mechanism that explain the increase in tax collection.

*Treatment effects by Causal vs Correlational Signal.* — Table 5 examines whether the nature of the signal—causal vs. correlational—regarding the impact of sending tax reminders to enhance tax collection affects tax reminder policy and tax collection. The objective here is to understand the mechanism at play—how metrics training influences causal thinking—and how it subsequently shapes policy and fiscal capacity. Table 5 reveals that the influence of metrics training is exclusively attributed to those tax officers who receive the causal signal (as illustrated in Figure B11), while metrics trained officers exposed to correlational signal (Figure B12) show minimal impact from the training. These findings suggest that the metrics training is only effective if the information about the policy's causal impact is known. Taken together, the evidence suggests that when faced with policy choices having real reputational costs, implementation challenges and public budgetary constraints, treated policymakers choose policy for which there is causal evidence.

## V. Heterogeneity by Initial Beliefs, Defiers and Compilers, and Baseline Quantitative Scores

*Project choice and heterogeneity by initial beliefs.* — In this subsection, we present evidence that the metrics training impacted the stated project choice of policymakers. Earlier, we observed from Figure 3 that metrics training shifted policymakers' beliefs on impact of deworming on income after being presented with the signal value of 13%, while no such shifting

for both teacher vaccination and absenteeism in a role model vaccination message campaign (Mehmood, Naseer, Chen 2023).

[9] The MDE is calculated as 2.8 times the standard error from the regression coefficient (Ioannidis, et al., 2017).

[10] Values larger than PKR 80 million are dropped for readability. Keeping these values yields a similar conclusion. Results are available on request.

of initial beliefs is observed for the group assigned the placebo training. This indicates that metrics training made the policymakers more responsive to RCT evidence relative to the placebo group. However, we observe interesting heterogeneity in these results. Metrics-assigned policymakers are *only* more likely to choose the deworming relative to the computer lab policy if their initial beliefs on the impact of deworming is below the signal value impact of 13%. In particular, metrics-trained policymakers whose priors are below signal value of 13% impact are about twice as likely —40% to 80%— to choose to implement the deworming policy when causal evidence is provided to them (Panel A of Figure 3). The effect is observed 6 months following the metrics training, suggesting a persistent effect of our intervention. We do not find much evidence of the metrics training impacting policymakers that had above the 13% prior belief on the impact of deworming on income. This indicates policymakers shifted their beliefs and chose policy for which there was causal evidence only insofar as they initially believed that the deworming did not have much impact. As suggested in Cantoni et al. (2019), we further disentangle heterogeneity by prior beliefs via semiparametric estimation at different percentiles of prior beliefs. Panel B of Figure 3 reports these results. The policy choice of only those metrics assigned policymakers are affected who had initial beliefs below the signal value of 13%, while we find not much evidence of the metrics training shifting policy choices of ministers who had priors above 13%. If anything, these policymakers whose initial beliefs are more than the signal value are less likely to choose deworming policy. This result is consistent with the proper interpretation of evidence. However, decision-makers with the strongest beliefs about the efficacy of deworming did not decrease the likelihood of choosing deworming when new evidence becomes available on a policy's efficacy being lower than expected. We cautiously interpret these patterns as broadly consistent with metrics training causing individuals to be more receptive to causal evidence.

*Heterogeneity by Defiers and Never-Takers.* — We next studied heterogeneity by defiers and never-takers. The tracking of the click behavior of policymakers across metrics and placebo lecture videos allowed us to investigate whether individuals assigned the metrics training attempt to defy the treatment assignment and attempt to access the placebo training. Likewise, we were able to investigate whether there were individuals, who were assigned the treatment but never clicked on the lecture videos (never-takers). Therefore, the unique setting allows us to proxy for defiers and never-takers to the metrics training treatment and examine the potential

heterogeneous impacts of our treatment. By matching the individual code and official email of policymakers, which they had used to log in to access the assigned lecture, we could directly observe and track click behavior of individual ministers using oTree (Chen, Schonger, and Wickens, 2016).[11] We also used the expertise of a computer scientist that made it nearly impossible for the policymakers to share, download or access training to which they were not assigned. In Table B7 (Columns 1 and 3) we report the heterogeneous impact of metrics training assignment on defiers—those who were assigned the metrics training lecture but attempted to access the placebo training lecture. We found that defiers are significantly less likely to be impacted by our metrics training. For example, we observed that metrics trained defiers are willing to pay—from public funds—about PKR 2,532,000 (USD 19000) less for randomized evaluations (Table B7, Column 3). We found that never-takers are less likely to be impacted as well but to a lesser extent (Table B7, Column 4). These results suggest that metrics training did not uniformly impact all policymakers and that behavioral data can be a potential method for detecting defiers and never-takers in estimating the average treatment effects in randomized trials.

*Heterogeneity by High and Low Demanders for Econometrics Training.*—Finally, we used our behavioral elicitation of potential compiler status to assess a typical concern of RCTs. Namely, in RCTs, the econometrician estimates the Local Average Treatment Effect (LATE) for the compliers that respond to treatment and the econometrician typically is unable to observe defiers. It is a plausible concern that people who demand to learn causal thinking may be more responsive to the treatment assignment. Thus estimates of the treatment impacts would be uninformative on those who are potential non-compliers. In our unique experimental set-up, we developed a proxy for compliers through those who demanded the metrics book; we show that the effects are the same for both the high and low demanders. As can be seen from Table B5 and B12 of Appendix B, we observe no significant differences between the treatment effects for low and high demanders of metrics training.

*Heterogeneity by Baseline Quantitative Scores.* — Finally, we investigated heterogeneity by pretreatment quantitative scores of the ministers This allowed us to assess possible heterogeneity of our treatment effect by quantitative ability. For instance, if those with higher

[11] We delivered the videos with logins and tracked click behavior using oTree's native features.

pre-treatment quantitative scores respond more to causal evidence. Table B8 presents these results. We find no evidence of the heterogeneous effect of metrics training — on policymakers who had high versus low quantitative ability. The effects of our treatment are very similar in terms of point estimates for those with above or below median quantitative scores. These results suggest that our instrument for the paradigm shift has a similar effect regardless of pretreatment quantitative ability.

## VI. Robustness and Discussion

This section details a series of sensitivity analyses and discusses that our results are unlikely to be explained by lack of balance, idiosyncratic sample, experimental demand effects or multiple hypotheses testing. We also provide additional comments on external validity and mechanisms.

*Balance.* — Earlier, we observed that the sample is balanced across a host of individual characteristics: income, age, years of education, gender, birth in political capitals, asset ownership and foreign visits. It is important to emphasize that the effects we observe are also unlikely to result from lack of balance in the quantitative ability of the deputy ministers who may be more responsive to metrics training. The rich set of outcome variables data gives us access to several pretreatment outcomes including baseline quantitative assessments. In fact, the sample is balanced not just in pretreatment mathematics scores but pretreatment scores on psychological, written and interview assessments—strongly suggesting that the candidates are balanced in underlying cognitive and even noncognitive abilities.

*Sample Size and Statistical Power.* — The focus on deputy ministers allows us to study an elite group of high-stakes decision-makers who can potentially impact long-run economic and political development. However, the selective nature of these policymakers necessitate that they are by design few in number. Therefore, our sample is restricted to about 200 deputy ministers, which raises concerns about lack of statistical power. Nevertheless, even with 200 individuals, our evidence complements several classical experimental studies that train individuals with less statistical power. Still, Imbens and Rubin (2015) recommend, in relatively small sample randomized evaluations, conducting randomization inference where the econometrician scrambles the data, reassigns treatments and compares the distribution of placebo estimates with

the estimate from the experiment. We report the resultant p-values in Tables B9, B10, and B11 of Appendix B with 1000 iterations of this process.[12] Even though the p-values slightly increase, the treatment effects are still statistically significant at conventional significance levels. These results strongly suggest that idiosyncratic small sample bias is unlikely to explain our results.

*Experimental Demand.* — It is also unlikely that experimental demand drives our results, i.e., deputy ministers in the metrics training treatment are providing responses in a manner they feel are simply expected by the experimenter. This is due to several reasons. First and foremost, the tax collection is independent from both the experimenter and the training academy. Especially relevant is the fact these correlational signals about tax collection did not affect tax officers in the econometrics training group. Second, two policies for which no causal evidence signal were also included in the budget documents and they are unaffected by the metrics training. Neither placebo policy was affected by metrics training. Third, in addition to increased WTP for commissioning RCTs, we find no evidence of metrics training on WTP for advice of senior bureaucrats, a potential source of experimenter demand. Fourth, we use a method inspired by the methodology of De Quidt et al. (2018) to mitigate concerns of experimenter demand. We requested the subjects to choose an opposite ICT project than the deworming policy for which causal evidence was provided. In other words, we demanded that a policy other than what we would expect based on RCT evidence presented should be chosen.[13] Last, national policy exams were administered separately from the experimental team. All of these patterns are inconsistent with experimental demand explaining our results.[14]

*External Validity.* — In this subsection, we discuss external validity. First, the elite bureaucrats of Pakistan, their selection procedures and training are similar to many other developing countries, especially India and Bangladesh who, like Pakistan, inherited these of bureaucratic institutions during the British Colonial rule of the Indian subcontinent. Pakistan, India and Bangladesh alone consist of more than a quarter of world's population making this study particularly relevant for a large number of people. Second, we follow List (2020)'s *Selection-Attrition-Naturalness-Scaling* (SANS) conditions in our discussion of the

[12] *ritest* in Stata is implemented to compute p-values corresponding to the permutation inference test by Heb (2017).
[13] Before the policy choice, all deputy ministers were told that ICT policy is important for the 21st century economy.
[14] The results are also robust to an alternative specification of a saturated model that also includes an interaction between metrics assigned and metrics chosen (see Table B11 in Appendix B).

generalizability of our results. First, in terms of selection, our sample consisted of almost all deputy ministers that entered service in Pakistan via competitive examinations in a given year (that we have anonymized). Considering the nature of the setting, time frame and choice task, we obtained natural measures such as policy assessments and simulations. The policymakers performed natural tasks in the field such as policy choices, teamwork assessments, national public policy assessments and field visits. Finally, in terms of scaling our intervention in other settings, the intervention was cheap to deliver since it was largely online. Since the training was delivered online it may also be scaled to other high-stakes decision-makers such as judges and CEOs in other settings. However, we view these results as a WAVE1 insight, in the nomenclature of List (2020), and replications need to be completed to understand if the effect sizes can be applied to general populations as well as high-stakes decision-makers in other contexts.

*Multiple Hypothesis Testing*. — Given the fact that we are testing multiple hypotheses, we also examine if our results are driven by false rejections. Under the assumption that treatment has no effect on any of our outcomes (all our null hypotheses are true), then the probability of one or more false rejections when using a critical value of 0.05 is about 40%. As a result, in order to reduce the likelihood of false rejections, we adjust for the fact that we are testing for multiple hypotheses. Following the literature, we use sharpened False Discovery Rate (FDR) q-values suggested by Anderson et al., 2008 (see for instance Heckman et al., 2018 for an application). These sharpened q-values are presented in square brackets in Table B13 where we also show standard p-values from our regressions in parentheses for comparison. Our results remain robust at conventional significance levels.

## VII. Conclusion

In this paper, we explore the effects of econometrics training on deputy ministers' demand and responsiveness to certain types of evidence. We find that training the principles of a new paradigm, associated with the credibility revolution, causes substantial shifts in attitudes and behavior. Our rich data on individual deputy ministers allows us to provide a deeper understanding of the behavioral consequences of the paradigm shift.

One year after the training, in their official duties, treated policymakers are twice as likely to hypothetically choose policies for which there is causal evidence. Training econometrics

through an influential book that summarizes concepts associated with the credibility revolution yielded significant impacts on demand for and responsiveness to causal evidence. After six months, treated individuals' perceived importance of quantitative analysis increased by 50%. Their performance in national research methods and public policy exams improve by 0.5–0.8 sigma. Text analyses of their writings suggest an understanding of concepts associated with the credibility revolution. We also found that treated individuals' stated WTP for commissioning RCTs using public funding increased by 150% and decreased by 50% for correlational studies. This suggests that econometrics training changed how policymakers perceived the inputs to their decisions.

We also provide evidence regarding the mechanisms that may affect how policymakers respond to causal information. Econometrics training focusing on causal thinking likely corrected belief updating in how policymakers respond to causal evidence. Deputy ministers assigned to the placebo group sometimes updated their beliefs in the opposite direction of the signal or updated their beliefs past the signal. These findings suggest that training causal thinking not only increases responsiveness to causal evidence but may even correct for mistakes in the belief updating process.

Our second randomized field experiment provides strong evidence that econometrics training can improve tax policy decisions and revenue collection. Specifically, we find that a short 10-hour workshop led to a 20% increase in the use of tax reminder notices and a 40% increase in tax revenue collection. Moreover, the training only affected tax reminder policy among those officers who also received a causal research signal about the efficacy of reminders. A series of robustness checks—including examining the full distribution of tax revenue, controlling for demographics, and placebo tests—lend further credibility to these sizable treatment effects.

These results have important practical and theoretical implications. On the practical side, integrating econometrics training into standard curricula for tax officers could significantly strengthen fiscal capacity in developing countries. More broadly, the returns to even short, focused training demonstrate the value of investing in civil servant skills. Theoretically, our findings highlight the synergies between causal research and training in facilitating

evidence-based policymaking. Overall, this study points to new approaches for building state capability and improving governance through targeted capacity-building programs. Even small interventions to equip public officials with empirical skills can yield substantial dividends and supercharge their demand for and receptiveness to causal evidence (Hjort et al. 2021).

Understanding the properties of paradigm shifts and what makes them readily transmissible is an open question for future research. A school of thought may help focus decision-makers to pay attention to salient features associated with a decision (Falk and Tirole 2016). Mastering 'Metrics provides a concatenation of the school of thought associated with credibility revolution. We *cautiously* interpret the training program as welfare improving in light of the body of evidence that finds ambiguous effect of ICT policies on learning outcomes (Cristia, Ibarraran, Cueto, Santiago, and Severin 2012; Kremer, Brannen, and Glennerster 2013; Mbiti 2016), and positive welfare effects of deworming policies (Miguel and Kremer, 2004; Croke et al., 2016; Ahuja et al. 2017). Future research can follow the long-term effects of metrics training on these deputy ministers and tax officers as well as study the welfare consequences of training in the school of thought associated with the credibility revolution.

## REFERENCES


**Ahuja, Amrita, Sarah Baird, Joan Hamory Hicks, Michael Kremer, and Edward Miguel.** 2017. "Economics of Mass Deworming Programs." *Child and Adolescent Health and Development. 3rd edition*.

**Anderson, Michael L.** 2008. "Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects." *Journal of the American Statistical Association* 103 (484): 1481-1495.

**Angrist, Joshua D., and Jörn-Steffen Pischke.** 2010. "The credibility revolution in empirical economics: How better research design is taking the con out of econometrics." *Journal of economic perspectives* 24 (2): 3-30.

**Angrist, Joshua D., and Jörn-Steffen Pischke.** 2014. *Mastering 'metrics: The path from cause to effect*. Princeton university press.

**Ash, Elliott, Daniel L. Chen, and Suresh Naidu.** 2019. "Ideas have consequences: The impact of law and economics on American justice." *Center for Law & Economics Working Paper Series* 4.

**Azoulay, Pierre, Christian Fons-Rosen, and Joshua S. Graff Zivin.** 2019. "Does science advance one funeral at a time?." *American Economic Review* 109 (8): 2889-2920.

**Baekgaard, Martin, Julian Christensen, Casper Mondrup Dahlmann, Asbjørn Mathiasen, and Niels Bjørn Grund Petersen.** 2019. "The role of evidence in politics: Motivated reasoning and persuasion among politicians." *British Journal of Political Science* 49 (3): 1117-1140.

**Banuri, Sheheryar, Stefan Dercon, and Varun Gauri.** 2019. "Biased policy professionals." *The World Bank Economic Review* 33 (2): 310-327.

**Bauman, Yoram, and Elaina Rose.** 2011. "Selection or indoctrination: Why do economics students donate less than the rest?." *Journal of Economic Behavior & Organization* 79 (3): 318-327.

**Besley, T., Burgess, R., Khan, A. and Xu, G., 2022**. Bureaucracy and development. Annual Review of Economics, 14, pp.397-424.

**Berry, James, Greg Fischer, and Raymond Guiteras.** 2020. "Eliciting and utilizing willingness to pay: Evidence from field trials in Northern Ghana." *Journal of Political Economy* 128 (4): 1436-1473.

**Bonnefon, Jean-François, Azim Shariff, and Iyad Rahwan.** 2016. "The social dilemma of autonomous vehicles." *Science* 352 (6293): 1573-1576.

**Cantoni, Davide, David Y. Yang, Noam Yuchtman, and Y. Jane Zhang.** 2019. "Protests as strategic games: experimental evidence from Hong Kong's antiauthoritarian movement." *The Quarterly Journal of Economics* 134 (2): 1021-1077.

**Callen, Michael, Adnan Khan, Asim I. Khwaja, Asad Liaqat and Emily Myers M**. 2017. "These 3 barriers make it hard for policymakers to use the evidence that development researchers produce." *The Washington Post*.

**Chen, Daniel L., Tobias J. Moskowitz, and Kelly Shue.** 2016. "Decision making under the gambler's fallacy: Evidence from asylum judges, loan officers, and baseball umpires." *The Quarterly Journal of Economics* 131 (3): 1181-1242.

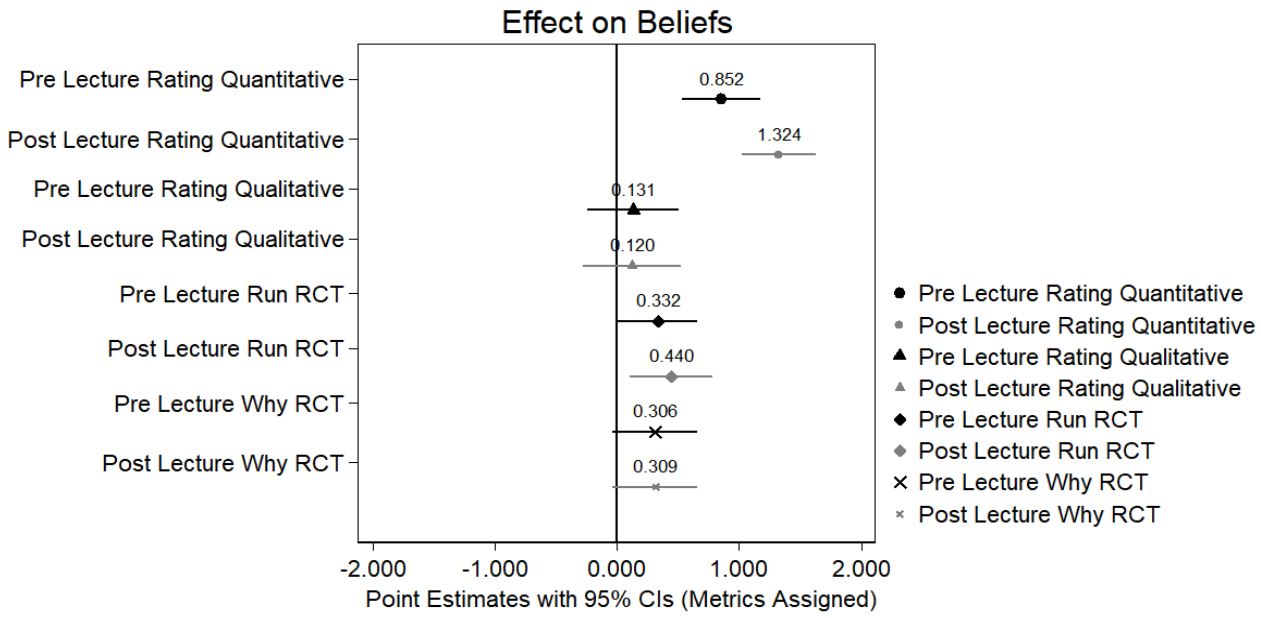
**Chen, Daniel L., Martin Schonger, and Chris Wickens.** 2016. "oTree—An open-source platform for laboratory, online, and field experiments." *Journal of Behavioral and Experimental Finance* 9: 88-97.
**Cole, Shawn, Thomas Sampson, and Bilal Zia.** 2011. "Prices or knowledge? What drives demand for financial services in emerging markets?." *The journal of finance* 66 (6): 1933-1967.
**Conte, Elisabetta, Ilaria Grazzani, and Alessandro Pepe.** 2018. "Social cognition, language, and prosocial behaviors: a multitrait mixed-methods study in early childhood." *Early Education and Development* 29 (6): 814-830.
**Cristia, Julian, Pablo Ibarrarán, Santiago Cueto, Ana Santiago, and Eugenio Severín.** 2017. "Technology and child development: Evidence from the one laptop per child program." *American Economic Journal: Applied Economics* 9 (3): 295-320.
**Croke, Kevin, Joan Hamory Hicks, Eric Hsu, Michael Kremer, and Edward Miguel.** 2016. *Does mass deworming affect child nutrition? Meta-analysis, cost-effectiveness, and statistical power* (No. w22382). National Bureau of Economic Research,
**David J. Deming. and Weidmann, Ben.** 2021. *Team Players: How Social Skills Improve Team Performance*. Forthcoming Econometrica.
**De Quidt, Jonathan, Johannes Haushofer, and Christopher Roth.** 2018. "Measuring and bounding experimenter demand." *American Economic Review* 108 (11): 3266-3302.
**Durkin, Kevin, and Gina Conti‐Ramsden.** 2007. "Language, social behavior, and the quality of friendships in adolescents with and without a history of specific language impairment." *Child development* 78 (5): 1441-1457.
**Frank, Robert H., Thomas Gilovich, and Dennis T. Regan.** 1993. "Does studying economics inhibit cooperation?." *Journal of economic perspectives* 7 (2): 159-171.
**Frey, Bruno S., and Stephan Meier.** 2003. "Are political economists selfish and indoctrinated? Evidence from a natural experiment." *Economic Inquiry* 41 (3): 448-462.
**Foucault, M.** 1970. "The order of things Routledge."
**Gabaix, Xavier.** 2014. "A sparsity-based model of bounded rationality." *The Quarterly Journal of Economics* 129 (4): 1661-1710.
**Grantham-McGregor, Sally M., Christine A. Powell, Susan P. Walker, and John H. Himes.** 1991. "Nutritional supplementation, psychosocial stimulation, and mental development of stunted children: the Jamaican Study." *The Lancet* 338 (8758): 1-5.
**Hjort, Jonas, Diana Moreira, Gautam Rao, and Juan Francisco Santini.** 2021. "How research affects policy: Experimental evidence from 2,150 brazilian municipalities." *American Economic Review* 111 (5): 1442-80.
**Heß, Simon.** 2017. "Randomization inference with Stata: A guide and software." *The Stata Journal* 17 (3): 630-651.
**Heckman, J. J., & Kautz, T.** (2012). Hard evidence on soft skills. Labour economics, 19(4), 451-464. is about non-cognitive skills in general. but Dweck, C. S. (2006). Mindset: The new psychology of success. Random House Incorporated. is about writing for self-reflection and visualization.
**Heckman, James J., and Ganesh Karapakula.** 2019. *The Perry Preschoolers at late midlife: A study in design-specific inference* (No. w25888). National Bureau of Economic Research.

**Ifcher, John, and Homa Zarghamee.** 2018. "The rapid evolution of homo economicus: Brief exposure to neoclassical assumptions increases self-interested behavior." *Journal of Behavioral and Experimental Economics* 75: 55-65.
**Ioannidis, J.P., Stanley, T.D. and Doucouliagos, H., 2017.** The power of bias in economics research. The Economic Journal, Volume 127, Issue 605, October 2017, Pages F236–F265,
**Kremer, Michael, Conner Brannen, and Rachel Glennerster.** 2013. "The challenge of education and learning in the developing world." *Science* 340 (6130): 297-300.
**Hamory, Joan, Edward Miguel, Michael Walker, Michael Kremer, and Sarah Baird.** 2021. "Twenty-year economic impacts of deworming." *Proceedings of the National Academy of Sciences* 118 (14).
**Khan, A.Q., Khwaja, A.I. and Olken, B.A., 2016**. Tax farming redux: Experimental evidence on performance pay for tax collectors. The Quarterly Journal of Economics, 131(1), pp.219-271.
**Krämer, M., Jechel, L., Kretschmer, T. and Schneider, E.,** 2021. Rigorous Impact Evaluation: Evidence Generation and take-up in German Development Cooperation.
**Kremer, Michael, Gautam Rao, and Frank Schilbach.** 2019. "Behavioral development economics." In *Handbook of Behavioral Economics: Applications and Foundations 1* (Vol. 2, pp. 345-458). North-Holland.
**Kuhn, T. S.** 1962. "The Structure of Scientific Revolution. 1st Edn Chicago." *IL: University of Chicago Press*.
**LaLonde, Robert J.** 1986. "Evaluating the econometric evaluations of training programs with experimental data." *The American economic review*: 604-620.
**Leamer, Edward E.** 1983. "Let's take the con out of econometrics." *The American Economic Review* 73 (1): 31-43.
**Lu, Wei, Elliott Ash, and Daniel L. Chen.** 2021." Motivated Reasoning in the Field: Polarization of Precedent, Prose, and Policy in U.S. Circuit Courts, 1930-2013." *Mimeo*.
**Lugo-Gil, J., Jean-Baptiste, D. and Jaramillo, L.F., 2019.** Use of evidence to drive decision-making in government. Mathematica Policy Research, 1, pp.1-80.
**Lusardi, Annamaria, and Olivia S. Mitchell.** 2014. "The economic importance of financial literacy: Theory and evidence." *Journal of economic literature* 52 (1): 5-44.
**Lusardi, Annamaria, Pierre-Carl Michaud, and Olivia S. Mitchell.** 2017. "Optimal financial knowledge and wealth inequality." *Journal of Political Economy* 125 (2): 431-477.
**Malmendier, Ulrike, Stefan Nagel, and Zhen Yan.** 2017. *The making of hawks and doves: Inflation experiences on the FOMC* (No. w23228). National Bureau of Economic Research.
**Marwell, Gerald, and Ruth E. Ames.** 1981. "Economists free ride, does anyone else?: Experiments on the provision of public goods, IV." *Journal of public economics* 15 (3): 295-310.
**Mbiti, Isaac M.** 2016. "The need for accountability in education in developing countries." *Journal of Economic Perspectives* 30 (3): 109-32.
**McKenzie, D.,** 2021. Small business training to improve management practices in developing countries: re-assessing the evidence for 'training doesn't work'. Oxford Review of Economic Policy, 37(2), pp.276-301.
**Mehmood, Sultan, Shaheen Naseer, and Daniel L. Chen.** 2021. Enhancing Public Officials' Altruistic Behavior through Training Programs. Mimeo.

**Miguel, Edward, and Michael Kremer.** 2004. "Worms: identifying impacts on education and health in the presence of treatment externalities." *Econometrica* 72 (1): 159-217.
**Metzger, L., Svoronos, T. and Khan, A., 2020.** Policy decisions and evidence use among civil servants: A group decision experiment in Pakistan. CID Working Paper Series.
**Muennig, Peter, Dylan Robertson, Gretchen Johnson.** 2011. "The effect of an early education program on adult health: the Carolina Abecedarian Project randomized controlled trial." *American journal of public health* 101 (3): 512-516.
**Merton, Robert K.** 1973. *The sociology of science: Theoretical and empirical investigations*. University of Chicago press.
**Pearl, J. and Mackenzie, D**., 2018. The book of why: the new science of cause and effect. Basic books.
**Rubinstein, Ariel.** 2006. "Dilemmas of an economic theorist." *Revista de Economía Institucional* 8 (14): 191-213.
**Shapin, Steven.** 1982. "History of science and its sociological reconstructions." *History of science* 20 (3): 157-211.
**Slemrod, J., 2019**. Tax compliance and enforcement. Journal of Economic Literature, 57(4), pp.904-954.
**Sidorov, Grigori, Alexander Gelbukh, Helena Gómez-Adorno, and David Pinto.** 2014. "Soft similarity and soft cosine measure: Similarity of features in vector space model." *Computación y Sistemas* 18 (3): 491-504.
**Siegel, Daniel J.** 2010. *Mindsight: The new science of personal transformation*. Bantam.
**Sutter, Matthias, Michael Weyland, Anna Untertrifaller, and Manuel Froitzheim.** 2020. "Financial literacy, risk and time preferences–Results from a randomized educational intervention." *MPI Collective Goods Discussion Paper* (2020/17).
**Vivalt, Eva, and Aidan Coville.** 2021. "How Do Policymakers Update?" *Mimeo*.
**World Health Organization. (2023)**. Soil Transmitted Helminth Infections. Retrieved from [https://www.who.int/news-room/fact-sheets/detail/soil-transmitted-helminth-infections]

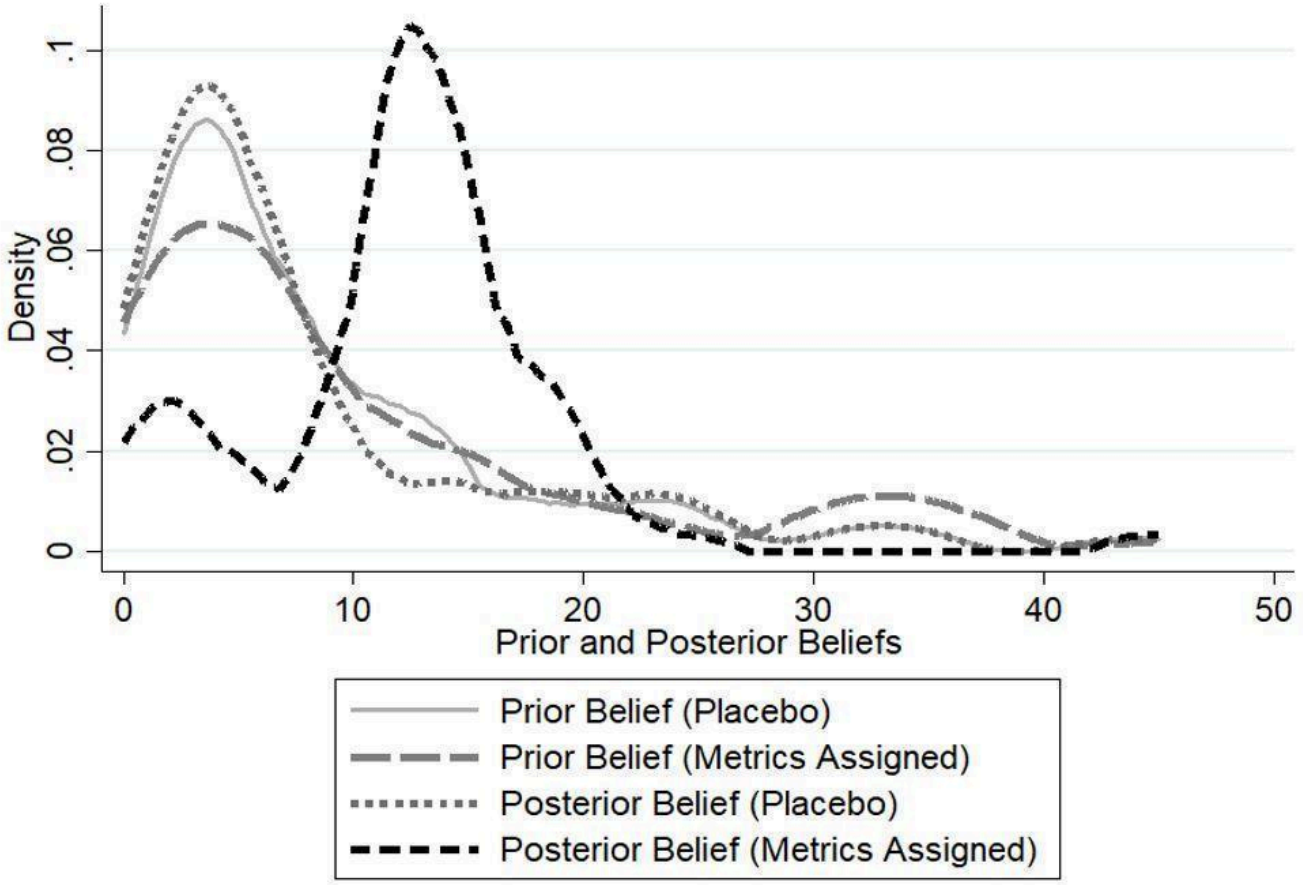
# Figures and Tables

**Figure 1: Impact of Metrics Training on Beliefs – Standardized**




*Note:* The figure above estimates our main specification including choice of book and individual level controls with all dependent variables standardized to mean zero and standard deviation one. Point estimate and 95% confidence interval on the randomly assigned metrics training is presented for each dependent variable pre (partial training) and post lecture and discussions (full training). The post-lecture results correspond to participants attending the complete metrics training: reading the Mastering metrics book, completing the required assignments, attending a lecture and participating in a discussion on the content.

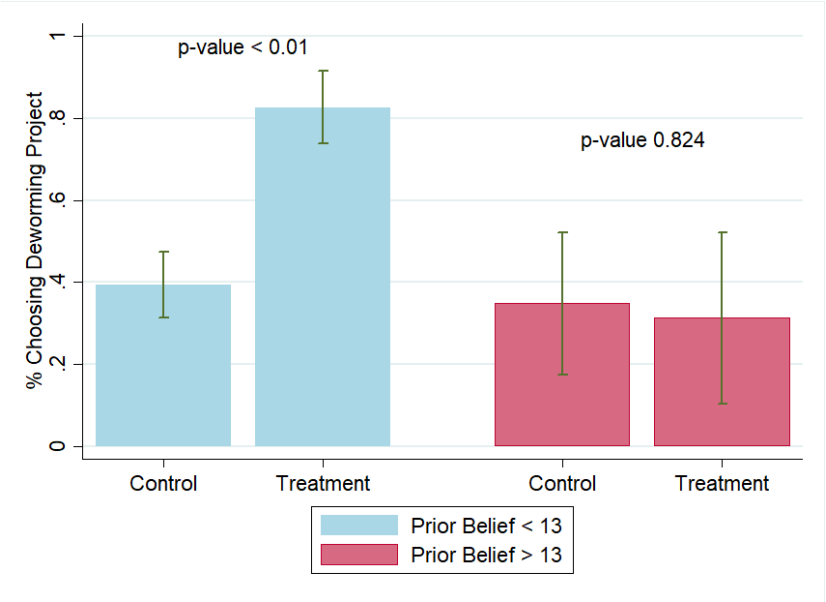
**Figure 2**: **Distribution of Initial Beliefs and Post-Signal Beliefs**




*Note*: The figure plots the distribution of prior and posterior beliefs on the effect of deworming on income after information on the estimate from a randomized evaluation on the effect of deworming is revealed to all participants. The estimate for impact of deworming is taken from a 25 year long randomized evaluation by Kremer et al. 2020. The metrics assigned group corresponds to participants attending the complete metrics training: reading the Mastering Metrics book, completing the required assignments, attending a lecture and participating in a discussion on the content.

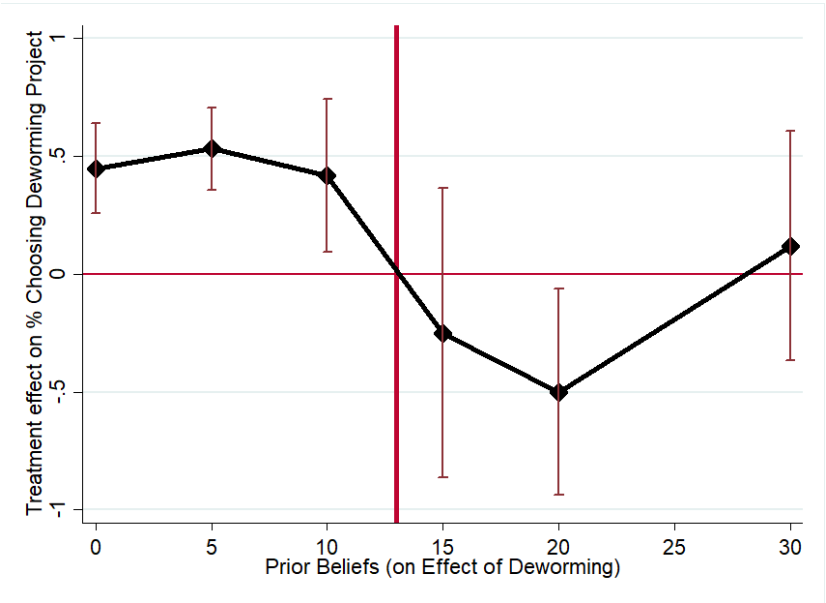
**Figure 3: Effect of Metrics Training on Deworming Project Choice by Initial Beliefs**

**Panel A:** Parametric Estimation



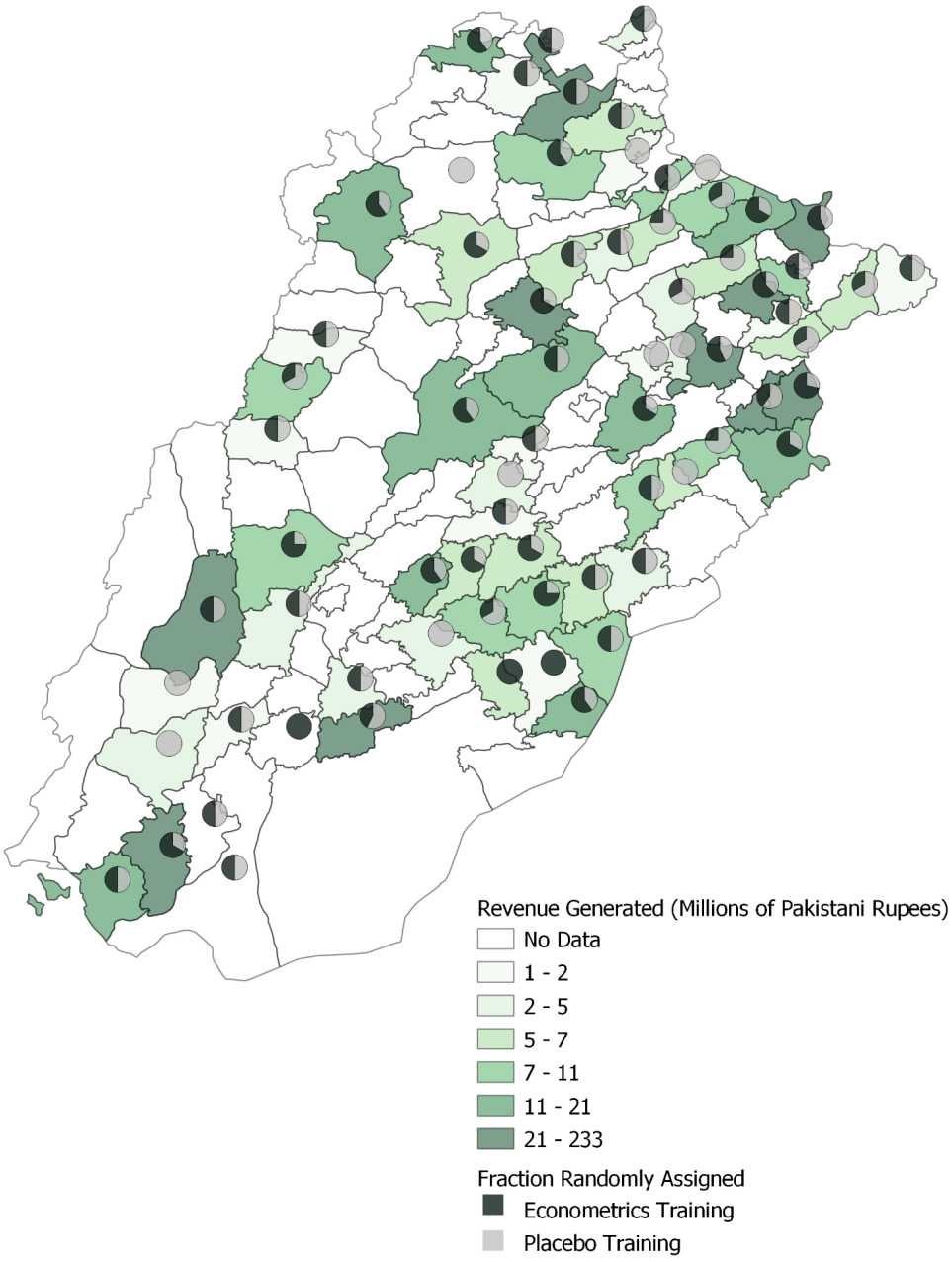

*Note*: The figure plots bar charts documenting the average impact of metrics training by those that had priors less and greater than the signal estimate of 13% increased effect of deworming on hourly wages along with 95% confidence intervals. The treatment corresponds to participants attending the complete metrics training: reading the Mastering Metrics book, completing the required assignments, attending a lecture and participating in a discussion on the content.

**Panel B:** Semi-Parametric Estimation



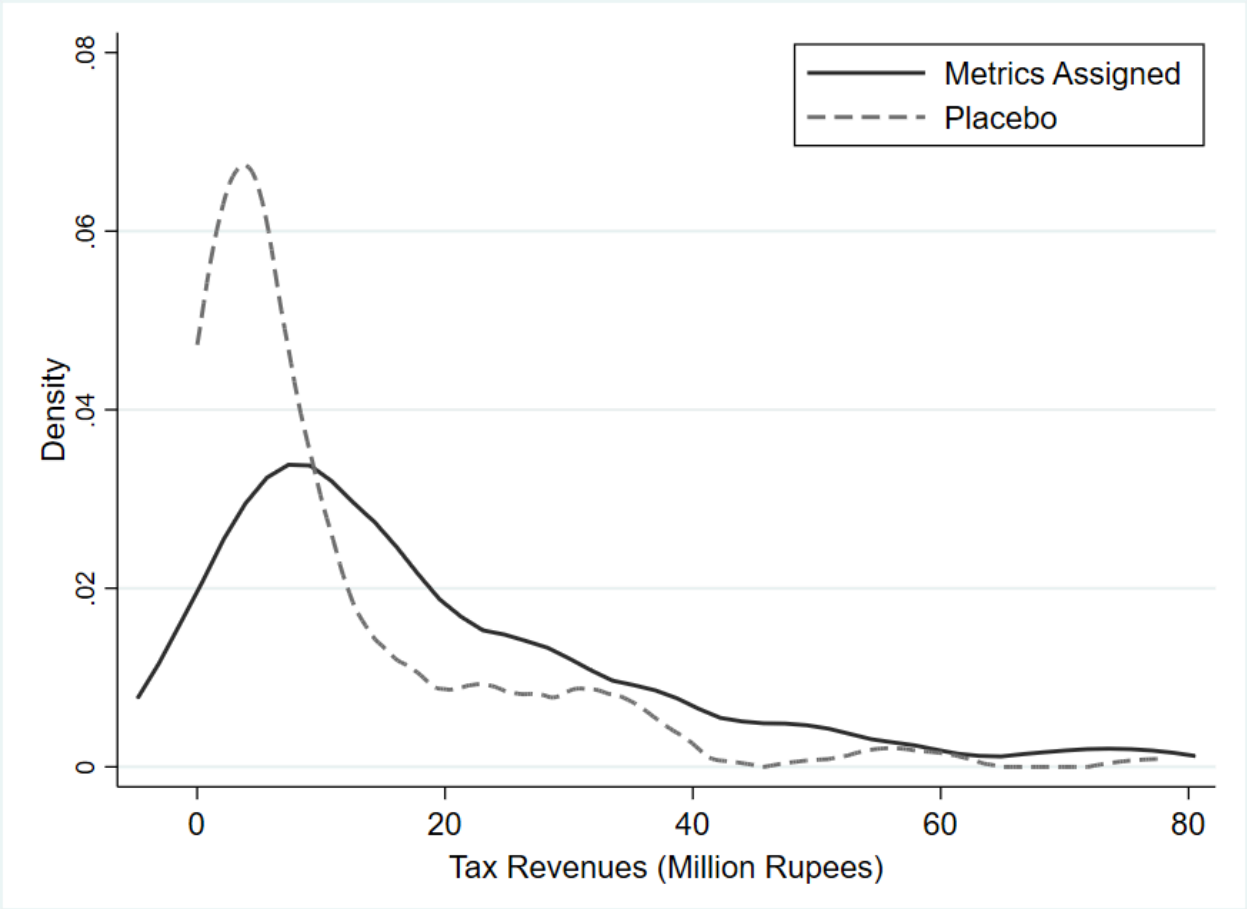

*Note*: The figure presents estimates of the impact of metrics training on choosing the deworming project by prior beliefs of the participants. The treatment corresponds to participants attending the complete metrics training: reading the Mastering Metrics book, completing the required assignments, attending a lecture and participating in a discussion on the content.

## Figure 4: Treated versus Control Regions by Tax Revenue Units in Punjab




Note: The figure illustrates tax units in Punjab, with darker colors representing higher tax collections in the region and lighter regions representing lower tax revenues. The units are in millions of rupees. The pie charts on the regions indicate the proportion of metrics-trained tax officers versus placebo-trained tax officers. Within each region, tax officers were either treated with a metrics workshop or a placebo macroeconomics workshop. Full dark circles represent all tax officers as treated, while full light pie charts represent no one in the region being treated with metrics training.

**Figure 5: Tax Revenue distributions of Metrics Trained versus Placebo Trained Tax Officers**




Note: The figure presents estimates of probability density functions for tax revenue collected by metrics-trained and placebo-trained tax officers. The units are in millions of Pakistani rupees.

**Table 1: Deputy Ministers - Balance of Treatment on Individual Characteristics and pre-treatment outcomes**

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Birth in political capitals | Income | Age | Education | Visited Foreign Country | PAS | PSP | Other groups | Pre-Treatment Written Assessment | Pre-Treatment Interview Assessment | Pre-Treatment Mathematics Assessment |
| Metrics Assigned | 0.0528 | -7,327 | 0.212 | 0.104 | -0.00229 | -0.0130 | -0.0549 | 0.0235 | 0.960 | 2.208 | 0.0627 |
| | (0.0902) | (4,601) | (0.395) | (0.0873) | (0.0712) | (0.0438) | (0.0348) | (0.0570) | (5.049) | (3.091) | (0.218) |
| Controls | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Observations | 190 | 190 | 190 | 190 | 190 | 190 | 190 | 190 | 190 | 190 | 190 |
| Mean of dep. variable | 0.324 | 34258.26 | 26.775 | 0.516 | 0.225 | 0.169 | 0.099 | 0.610 | 655.585 | 131.085 | 7.221 |

Robust standard errors appear in parentheses (clustered at the individual level). Metrics assigned is a dummy variable that switches on when a causal inference book is randomly assigned to participants. The causal inference book is randomly assigned conditional on the book being chosen. The controls include Metrics Chosen (a dummy variable that switches on when causal inference book is chosen by the participants), and all other available individual characteristics obtained from administrative data (i.e. all remaining column dependent variable except the dependent variable used in the respective column). *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

**Table 2: Impact of Metrics Training on Policy Making Assessments - Administrative Data - Standardized**

| | *Assessment Public Policy* | | *Assessment Research Methods* | | *Assessment Teamwork* | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Metrics Assigned | 0.574*** | 0.505*** | 0.814*** | 0.799*** | -0.004 | 0.017 |
| | (0.165) | (0.168) | (0.173) | (0.180) | (0.189) | (0.194) |
| Individual Controls | No | Yes | No | Yes | No | Yes |
| Observations | 190 | 190 | 190 | 190 | 190 | 190 |
| R-squared | 0.097 | 0.171 | 0.186 | 0.219 | 0.001 | 0.123 |

Robust standard errors appear in parentheses (clustered at the individual level). The dependent variables are standardized scores with mean 0 and standard deviation 1 from regular public policy training workshops at the training Academy. Columns (1) and (2) are scores on the workshop called *Public Sector Economics, Public Goods and Publicly Provided Private Goods* that consisted of case studies and analysis of past (actual) decisions of similar policymakers. The course content cover scenarios that apply concepts of public goods, externalities, the use of data in policymaking. Columns (3) and (4) present scores on *Research Methods* are reported. The workshop content included an introduction to hypothesis testing, multivariate regression, randomized controlled trials with particular focus on application to policy. Finally, in Columns (5) and (6) scores *Teams & Group Decisions* workshop. This assessment scores decisions pertaining to teamwork and group work these policymakers typically make in the field. Metrics assigned is a dummy variable that switches on when causal inference book is randomly assigned to participants. Consistent with all our earlier regressions, we always control for metrics chosen. The causal inference book, Mastering Metrics, is randomly assigned conditional on the book being chosen. The estimations obtained from OLS regressions includes the following controls: metrics chosen, written test scores, interview test scores, gender, birth in political capitals, asset ownership, income before joining public service, age, education, foreign visits and occupational group dummies. *** p<0.01, ** p<0.05, * p<0.1.

**Table 3: Tax Officers - Balance over individual characteristics and pre-treatment outcomes**

| | (1) Gender | (2) Age | (3) Land | (4) Income | (5) Birth capital | (6) Religion | (7) Experience | (8) Pre-treatment Revenue | (9) Pre-treatment Reminder Letter | (10) Metrics Assigned |
|---|---|---|---|---|---|---|---|---|---|---|
| Metrics Assigned | -0.0263 | -0.369 | 0.0497 | 1,300 | -0.00211 | -0.0364 | 0.00459 | -10.76 | -0.0321 | |
| | [0.0558] | [0.268] | [0.0527] | [964.7] | [0.0531] | [0.0258] | [0.257] | [14.79] | [0.0580] | |
| Gender | | | | | | | | | | -0.0283 |
| | | | | | | | | | | [0.0601] |
| Birth capital | | | | | | | | | | -0.00255 |
| | | | | | | | | | | [0.0639] |
| Land | | | | | | | | | | 0.0609 |
| | | | | | | | | | | [0.0646] |
| Income | | | | | | | | | | 0.0000046 |
| | | | | | | | | | | [0.00000339] |
| Age | | | | | | | | | | -0.0170 |
| | | | | | | | | | | [0.0121] |
| Religion | | | | | | | | | | -0.187 |
| | | | | | | | | | | [0.125] |
| Experience | | | | | | | | | | 0.000233 |
| | | | | | | | | | | [0.0131] |
| Pre-treatment Revenue | | | | | | | | | | -0.000165 |
| | | | | | | | | | | [0.000213] |
| Pre-treatment Notification | | | | | | | | | | -0.0319 |
| Observations | 310 | 310 | 310 | 310 | 310 | 310 | 310 | 310 | 310 | 310 |
| R-squared | 0.049 | 0.291 | 0.054 | 0.050 | 0.031 | 0.029 | 0.273 | 0.036 | 0.013 | 0.026 |
| Mean | 0.610 | 32.86 | 0.300 | 55203 | 0.297 | 0.948 | 4.906 | 53.42 | 0.539 | 0.500 |
| F Statistics | 0.222 | 1.898 | 0.891 | 1.817 | 0.00158 | 2.001 | 0.000318 | 0.529 | 0.306 | 0.965 |
| p-values | 0.638 | 0.169 | 0.346 | 0.179 | 0.968 | 0.158 | 0.986 | 0.468 | 0.580 | 0.469 |

Robust standard errors appear in brackets (clustered at the tax region level). Metrics assigned is a dummy variable that switches on when the causal inference workshop is randomly assigned to participants relative to the placebo macroeconomics workshop. The causal inference book is randomly assigned conditional on the book being chosen. The controls include other available individual characteristics obtained from administrative data (i.e. all remaining column dependent variable except the dependent variable used in the respective column). *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

**Table 4: Impact of Metrics Training on Tax Reminder Letters and Tax Revenue**

| | (1) Tax Reminders | (2) Tax Reminders | (3) Tax Revenues | (4) Tax Revenues |
|---|---|---|---|---|
| Metrics Assigned | 0.123** | 0.123** | 28.34* | 24.53* |
| | [0.0545] | [0.0543] | [14.81] | [13.65] |
| Controls | No | Yes | No | Yes |
| Observations | 310 | 310 | 310 | 310 |
| R-squared | 0.015 | 0.053 | 0.012 | 0.041 |
| Mean | 0.584 | 0.584 | 55.32 | 55.32 |

Robust standard errors appear in brackets (clustered at the tax region level). Metrics assigned is a dummy variable that switches on when a causal inference workshop is randomly assigned to participants relative to the placebo macroeconomics workshop. The controls include available individual characteristics in the balance table. *** p<0.01, ** p<0.05, * p<0.1.

**Table 5: Impact of Metrics Training and Signals on Tax Reminder Letters and Tax Revenue**

| | (1) Tax Reminders | (2) Tax Reminders | (3) Tax Revenues | (4) Tax Revenues |
|---|---|---|---|---|
| Metrics Assigned | -0.00616 | -0.00544 | 15.63 | 11.44 |
| | [0.0770] | [0.0756] | [13.79] | [13.23] |
| Causal Signal | -0.00616 | 0.0159 | -1.849 | -7.233 |
| | [0.0812] | [0.0799] | [16.15] | [14.59] |
| Metrics Assigned x Causal | 0.256** | 0.259** | 25.29** | 26.26** |
| Signal | [0.108] | [0.107] | [12.36] | [10.61] |
| Controls | No | Yes | No | Yes |
| Observations | 310 | 310 | 310 | 310 |
| R-squared | 0.042 | 0.080 | 0.024 | 0.042 |
| Mean of Dep. Variable | 0.587 | 0.587 | 55.32 | 55.32 |

Robust standard errors appear in brackets (clustered at the tax region level). Metrics assigned is a dummy variable that switches on when a causal inference workshop is randomly assigned to participants relative to the placebo macroeconomics workshop.The "causal signal" is a binary variable activated when tax officers receive a summary handout of the article "Behavioral Interventions in Tax Compliance: Evidence from Guatemala" by Kettle et al. (2016). Alternatively, the variable takes zero if officers are assigned the correlational study "Effect of Notice of Tax Warning, Notice of Tax Collection, and Tax Education Programs on Tax Compliance in West Sumatera and Jambi" by Budiman and Inayati (2021). The controls include available individual characteristics in the balance table. *** p<0.01, ** p<0.05, * p<0.1.

# Online Appendix to:

## Training Policymakers in Econometrics: Evidence from Two Experiments

*By* Sultan Mehmood, Shaheen Naseer and Daniel Chen

## Contents

**A. Additional Experimental Detail**

**B. Additional Figures and Tables**

## A. Additional Experimental Detail

### Section 1: Deputy Ministers

### Table A1: Transcript of Email sent by the Training Academy

Dear Officers,

We request you to carefully read the book assigned to you. It may or may not be different from the Your Tasks for this assignment are listed below. Main Task/Assignment 1: After reading the assigned book, we request you provide a chapter-by-chapter summary of the whole book of around 1500 words (+/-100 words). Main Task/Assignment 2: After reading the assigned book, we request you provide an analysis of how you would apply the lessons learned from the book in your job. This again should be around 1500 words (+/-100 words).

Please send your complete assignment by 10th December … to our office. You should write your answer in a word document, convert it in PDF, put your CSA ID and Book assigned to you on top of the document.

Please note that due to some logistical considerations, we have not sent books (printed version), to your addresses as we initially planned. For the assignment we have shared the (both) e-Books and hardcopies with you so we are certain everyone receives the material on time. However, each of you will get these books, when you are on campus and we have a debriefing session (where we will share the results) of the soft skills workshop with you. There are two major tasks and three minor tasks within this assignment that you must complete after reading carefully your assigned book. Please only read the book assigned to you.

Best of Luck,

**Table A2: Structured Discussion Post-Lecture**

Each training lecture was followed by a structured group discussion. In this discussion, the following structure was followed. After the lecture, 3 candidates from treatment and control were randomly drawn to answer these two questions:

*Candidate 1*:
Q1. What do you think were the main messages of the lectures? Q2. How do you think you may apply lessons from today's lecture in your career? Give at least 3 examples.

*Candidate 2*:
Q1. What struck you most about today's lectures and why? Please be specific on what you think are the key takeaways of today's lectures. Q2. Can you give three examples on how the lessons of today's workshop could be applied in your official duties?

*Candidate 3*:
Q1. What are your thoughts on todays' talk? Q2. How may they apply in your official duties?

**Table A3:** The Metrics Trainees Beliefs, Willingness to Pay and Policy Choice

Prompt 1: The following is a 5-minute task where you read the following text and answer 8 short questions. Just enter the number what you think is correct in each of the 8 questions.

Question 1) What is the impact of deworming on hourly earnings of children, 20 years later? Please provide the number that indicates your belief about the percent change in hourly earnings. i.e., if you report 30, you believe it would be 30% increase.

[Enter Number]

You are appointed as a policy maker of Kuchlak district in Baluchistan (which of course is a distinct possibility you may become as you graduate your training program).

Kutchlak being a very poor region, it neither has good IT facilities nor does it have a school deworming project. You have limited budget to allocate among two projects, and you have to choose 1. The direct costs of implementation of both projects is roughly equal.

First project is a school deworming program where students all across Kuchlak schools are dewormed.

The second project is the Computer Lab program where in each school of Kuchlak, a computer lab is established.

We suggest that you implement the computer lab project given IT is the future.

Question 2) Before you decide you can choose to pay from your pocket (personal spending) in Pakistan Rupees for each the three pieces of information.

a) Recommendation from an expert bureaucrat on which project to implement.

[Enter Number PKR]

b) A randomized control trial assessing the impact of deworming vs. building computer lab. The effectiveness of the policy is measured on schooling outcomes, labor force outcomes such as additional years of schooling, hourly earnings and non-agricultural work hours.

[Enter Number PKR]

c) Last year's administrative data from Kuchlak showing that schools that implemented a deworming program have 10% higher overall test scores and 5% higher incomes, while schools that installed a computer lab had 20% higher test scores and 15% higher incomes.

[Enter Number PKR]

Question 3) Before you decide you can choose to pay from government budget in Pakistan Rupees for each the three pieces of information.

a) Recommendation from an expert bureaucrat on which project to implement.

[Enter Number PKR]

b) A randomized control trial assessing the impact of deworming vs. building computer lab. The effectiveness of the policy is measured on schooling outcomes, labor force outcomes such as additional years of schooling, hourly earnings and non-agricultural work hours.

[Enter Number PKR]

c) Last year's administrative data from Kuchlak showing that schools that implemented a deworming program have 10% higher overall test scores and 5% higher incomes, while schools that installed a computer lab had 20% higher test scores and 15% higher incomes.

[Enter Number PKR]

Question 4) What project would you choose between 1 or 2? One being deworming and two being the computer lab.

Prompt 2: Recent randomized evaluation finds deworming impacts on economic outcomes up to 20 years later. Individuals who received deworming experience up to 3 additional years of schooling, 14% increases in consumption expenditure, 13% increases in hourly earnings, 9% in non-agricultural work hours (Source: PNAS, 2021).

We suggest that you implement the computer lab project given IT is the future

Question 5) What is the impact of deworming on hourly earnings of children, 20 years later? Please provide the number that indicates your belief about the percent change in hourly earnings. i.e., if you report 30, you believe it would be 30% increase.

[Enter Number]

Question 6) Now, how much would you be willing to pay from your pocket (personal spending) in Pakistan Rupees for each the three pieces of information.

a) Recommendation from an expert bureaucrat on which project to implement.

[Enter Number PKR]

b) A randomized control trial assessing the impact of deworming vs. building a computer lab. The effectiveness of the policy is measured on schooling outcomes, labor force outcomes such as additional years of schooling, hourly earnings and non-agricultural work hours.

[Enter Number PKR]

c) Last year's administrative data from Kuchlak showed that schools that implemented a deworming program have 10% higher overall test scores and 5% higher incomes, while schools that installed a computer lab had 20% higher test scores and 15% higher incomes.

[Enter Number PKR]

Question 7) Before you decide you can choose to pay from the government budget Pakistan Rupees for each the three pieces of information.

a) Recommendation from an expert bureaucrat on which project to implement.

[Enter Number PKR]

b) A randomized control trial assessing the impact of deworming vs. building computer lab. The effectiveness of the policy is measured on schooling outcomes, labor force outcomes such as additional years of schooling, hourly earnings and non-agricultural work hours.

[Enter Number PKR]

c) Last year's administrative data from Kuchlak showing that schools that implemented a deworming program have 10% higher overall test scores and 5% higher incomes, while schools that installed a computer lab had 20% higher test scores and 15% higher incomes.

[Enter Number PKR]

Question 8) Now, what project would you choose 1 or 2? One being deworming and two being computer lab.

[Enter Number]

*Metrics Training and Prosocial Behavior.* — A large body of evidence documents that economics training may make individuals less prosocial. Individuals trained in neoclassical economic concepts are more likely to free ride, less likely to donate or cooperate (see, e.g., Marwell and Ames, 1981; Frey and Meier, 2003; Bauman and Rose, 2011). When it comes to evidence, with its attention to utilitarian cost-benefit calculations, it has been suggested that practitioners of data science can become less prosocial (Bonnefon et al. 2016). We, however, present evidence that the metrics training program does not come at the expense of reduced prosociality. We measure prosocial behavior by field measures such as visits to orphanages and volunteering in impoverished schools, as well as language use in their writings. The field measures are obtained from the Academy on the policymakers' "syndicate field trip" workshops. The policymakers undertook two field trips, one 4 months and the other 6 months following the training. In the first, they were provided a choice to either visit a prominent orphanage (*Dar-ul-Aman*) or attend lectures on a specific government program from a senior bureaucrat. In the second syndicate field trip, the policymakers are asked to choose between volunteering to teach in any impoverished government school that falls under the government's Progressive Education Network (PEN) or once again choosing to attend a lecture on government programs from a senior public official. From the top of Figure B3, we can observe that metrics training is unlikely to come at the expense of reduced field visits to orphanages or volunteering in low-income schools. Treated ministers are neither less likely to visit orphanages nor volunteer in impoverished schools. We also found that the metrics training program did not come at the expense of reduction in the use of prosocial language in social media. The dependent variable is the Soft-Cosine Measure (SCM) representing the similarity of writings of policymakers with specific phrases related to prosocial behavior with higher values representing greater similarity.[15] Prosocial phrases were chosen based on recent work that shows that these phrases were correlated with field measures of prosocial behavior such as blood donations of these deputy ministers (Mehmood, Naseer and Chen 2021). Specifically, in Figure B3**,** we find words associated with prosociality are unaffected by our treatment; if anything, the metrics training is likely to reduce the use of "them" and "I", words associated with lack of social cohesion.

[15] It is a continuous variable with 0 denoting no similarity and 1 indicating perfect match with the phrase. SCM is a machine learning textual analysis algorithm that compares similarity between words and accurately detects similarity when they have no words in common between phrases (using pre-trained word embeddings). It is shown to outperform many of the state-of-the-art methods in the semantic text similarity tasks and is widely used commercially e.g. by Google Translate (for more details, see for instance, Sidorov et al., 2014)

# Section 2: Tax Officers - Experimental Details

**Illustration 1**: Summary statistics and variable description for Tax Officers Experiment

**Panel A: Dependent Variables**

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| Tax Reminders | 310 | .584 | .494 | 0 | 1 |
| Tax Revenue | 310 | 55.315 | 128.198 | .01 | 774.374 |

Note: The table contains pre-registered dependent variables used in the regressions in the tax officers' experiment.

**Panel B: Baseline Characteristics**

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| Gender (Male) | 310 | .61 | .489 | 0 | 1 |
| Birth capital | 310 | .297 | .458 | 0 | 1 |
| Land | 310 | .3 | .459 | 0 | 1 |
| Income | 310 | 55203.226 | 8531.722 | 40000 | 70000 |
| Age | 310 | 32.865 | 2.733 | 28 | 42 |
| Religion | 310 | .948 | .222 | 0 | 1 |
| Experience | 310 | 4.906 | 2.572 | 1 | 12 |
| Pre-Tax Revenue | 310 | 53.422 | 128.735 | .005 | 879.622 |
| Pre-Tax Reminders | 310 | .539 | .499 | 0 | 1 |

Note: The table contains all baseline or control variables used in the regressions with controls in the tax officers experiment.

**Illustration 2. Baseline Characteristics**

| Variable | Description |
|---|---|
| Gender (Male) | Dummy Variable equals 1 if the person is male |
| Birth capital | Dummy Variable equals 1 if the birth city is the provincial capital |
| Land | Dummy Variable equals 1 if the person owns any land |
| Income | Monthly income in PKRs |
| Age | Age in Years |
| Religion | Dummy Variable equals 1 if the person is Muslim |
| Experience | Work experience in Years |

Note: The table describes baseline or control variables used in the regressions in the tax officers experiment.

# B. Additional Figures and Tables

## Figure B1: Contents of Treatment and Placebo Books

**Panel A***:* Mastering Metrics by Joshua Angrist and Jörn-Steffen Pischke



CONTENTS


*List of Figures* vii
*List of Tables* ix
*Introduction* xi
1 Randomized Trials 1
1.1 In Sickness and in Health (Insurance) 1
1.2 The Oregon Trail 24
Masters of 'Metrics: From Daniel to R. A. Fisher 30
Appendix: Mastering Inference 33
2 Regression 47
2.1 A Tale of Two Colleges 47
2.2 Make Me a Match, Run Me a Regression 55
2.3 Ceteris Paribus? 68
Masters of 'Metrics: Galton and Yule 79
Appendix: Regression Theory 82
3 Instrumental Variables 98
3.1 The Charter Conundrum 99
3.2 Abuse Busters 115
3.3 The Population Bomb 123
Masters of 'Metrics: The Remarkable Wrights 139
Appendix: IV Theory 142
4 Regression Discontinuity Designs 147
4.1 Birthdays and Funerals 148
4.2 The Elite Illusion 164
Masters of 'Metrics: Donald Campbell 175
5 Differences-in-Differences 178
5.1 A Mississippi Experiment 178
5.2 Drink, Drank, … 191
Masters of 'Metrics: John Snow 204
Appendix: Standard Errors for Regression DD 205
6 The Wages of Schooling 209
6.1 Schooling, Experience, and Earnings 209
6.2 Twins Double the Fun 217
6.3 Econometricians Are Known by Their … Instruments 223
6.4 Rustling Sheepskin in the Lone Star State 235
Appendix: Bias from Measurement Error 240


**Panel B***:* "Mindsight" by Daniel J. Siegel.

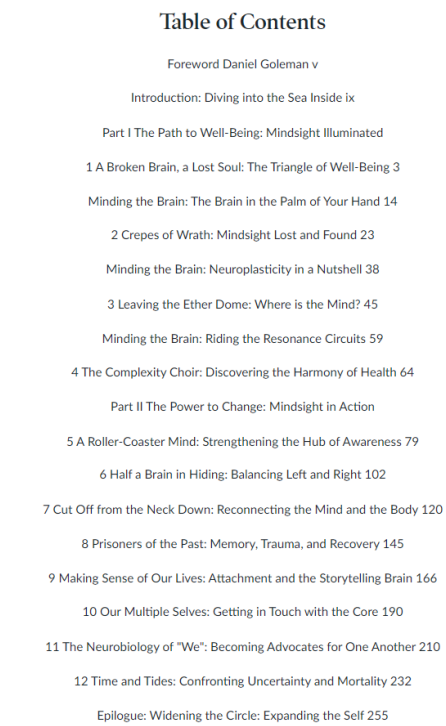
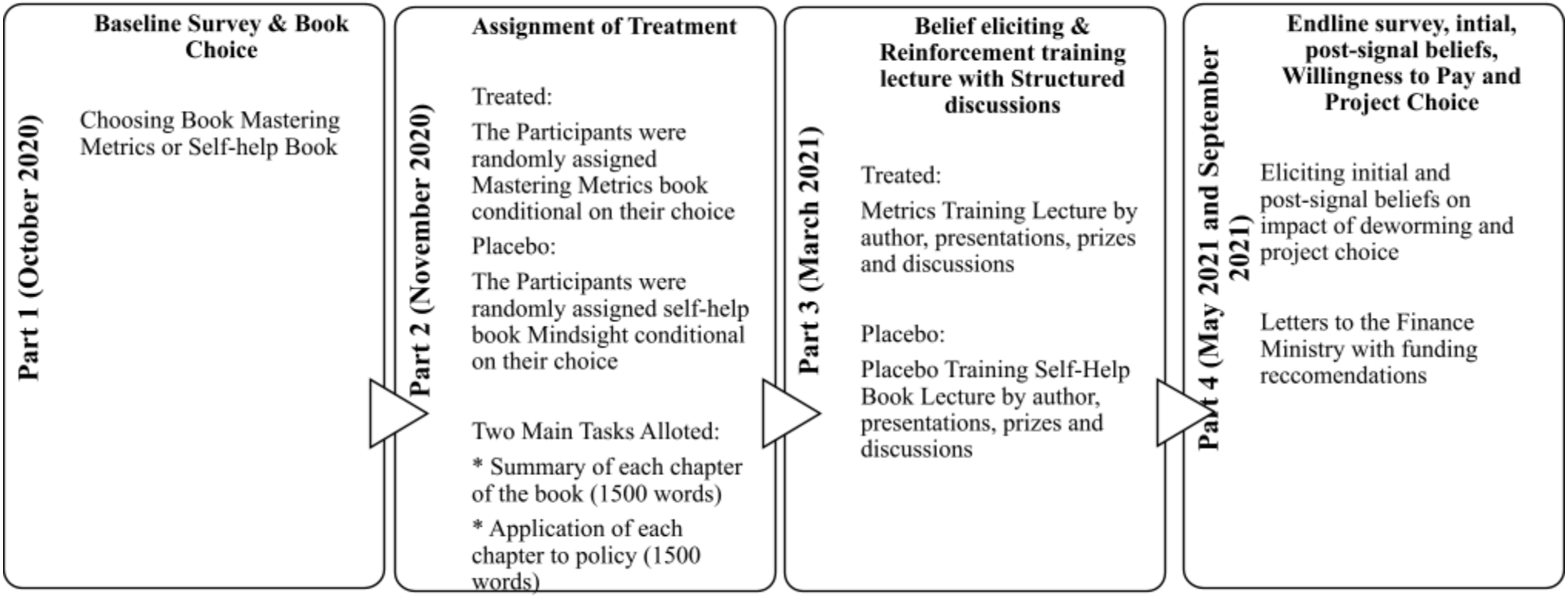


Table of Contents


Foreword Daniel Goleman v
Introduction: Diving into the Sea Inside ix
Part I The Path to Well-Being: Mindsight Illuminated
1 A Broken Brain, a Lost Soul: The Triangle of Well-Being 3
Minding the Brain: The Brain in the Palm of Your Hand 14
2 Crepes of Wrath: Mindsight Lost and Found 23
Minding the Brain: Neuroplasticity in a Nutshell 38
3 Leaving the Ether Dome: Where is the Mind? 45
Minding the Brain: Riding the Resonance Circuits 59
4 The Complexity Choir: Discovering the Harmony of Health 64
Part II The Power to Change: Mindsight in Action
5 A Roller-Coaster Mind: Strengthening the Hub of Awareness 79
6 Half a Brain in Hiding: Balancing Left and Right 102
7 Cut Off from the Neck Down: Reconnecting the Mind and the Body 120
8 Prisoners of the Past: Memory, Trauma, and Recovery 145
9 Making Sense of Our Lives: Attachment and the Storytelling Brain 166
10 Our Multiple Selves: Getting in Touch with the Core 190
11 The Neurobiology of "We": Becoming Advocates for One Another 210
12 Time and Tides: Confronting Uncertainty and Mortality 232
Epilogue: Widening the Circle: Expanding the Self 255

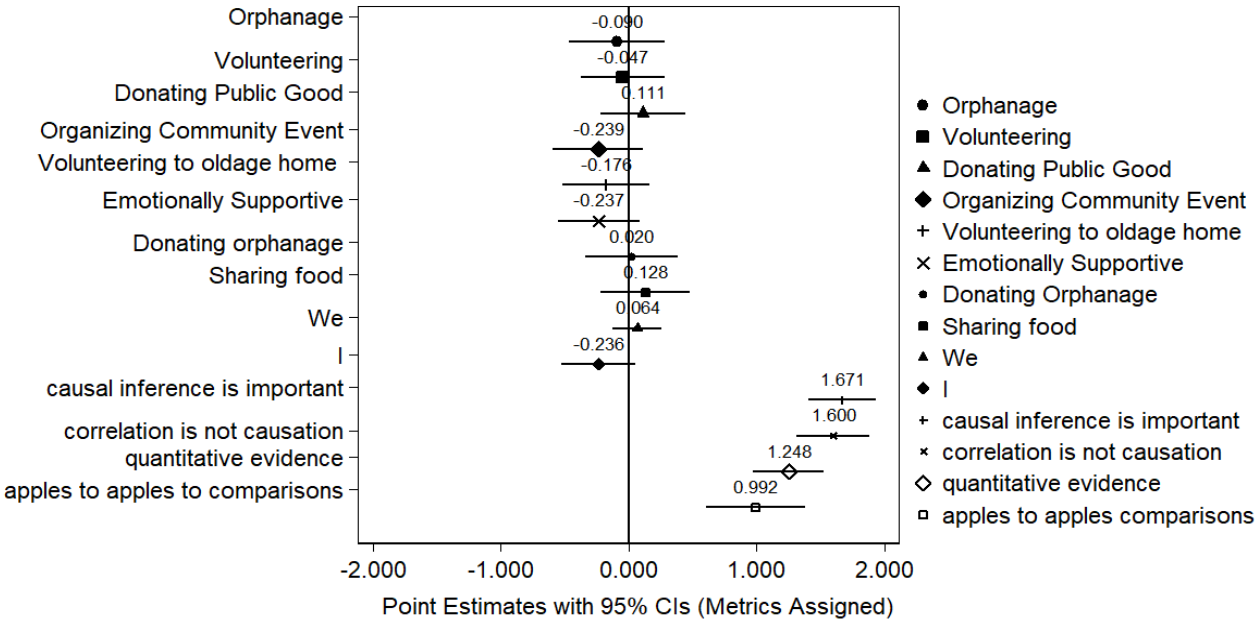

*Note:* The Figure provides table of contents for our treatment and placebo book. Panel A displays the table of contents for the book "Mastering Metrics" by Joshua Angrist and Jörn-Steffen Pischke. Panel B display the placebo book, "Mindsight: the New Science of Personal Transformation" by Daniel J. Siegel.

**Figure B2: Set-up of the Experiment and Intervention Detail**




*Note:* The Figure summarizes the main deputy minister's experiment and its corresponding timeline.

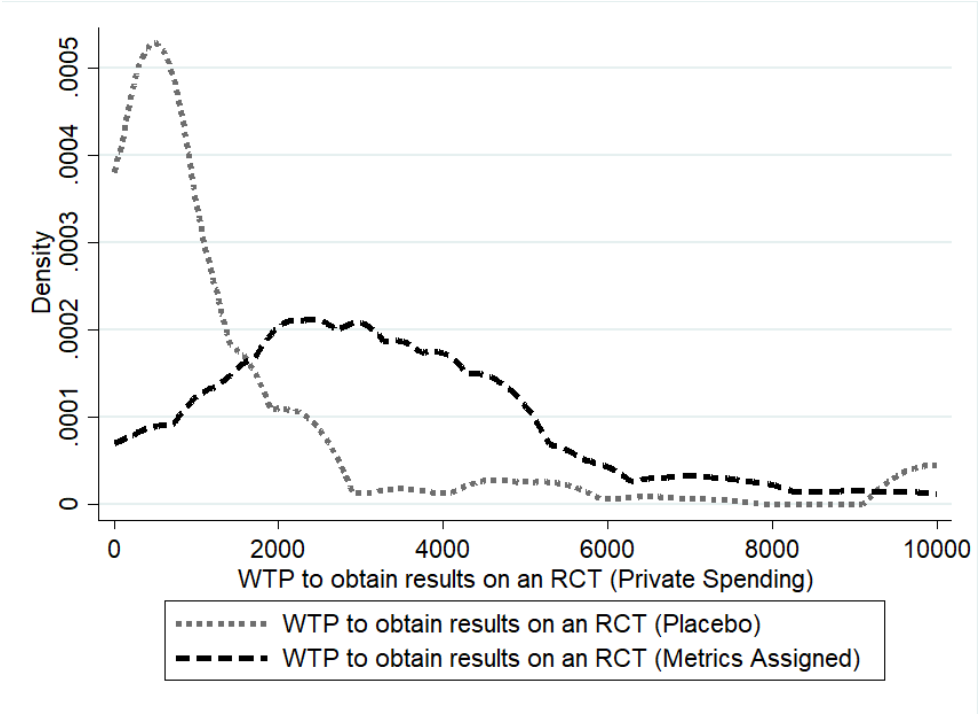
**Figure B3: Impact of Metrics Training on Prosociality and Causal Language - Standardized**



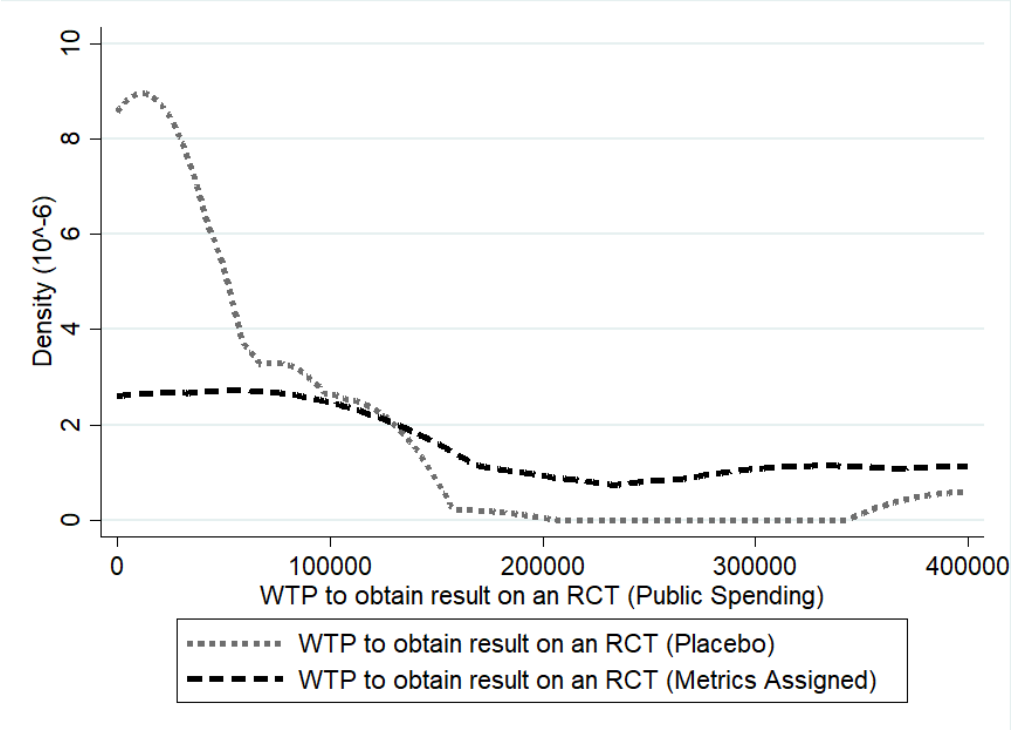

*Note:* The figure above estimates our main specification including choice of book and individual level controls on different dependent variables standardized to mean zero and standard deviation one. Point estimate and 95% confidence interval on the randomly assigned metrics training is presented for each dependent variable. The metrics assigned correspond to participants attending the complete metrics training: reading the Mastering metrics book, completing corresponding assignments, attending a lecture and participating in a discussion on the content.

**Figure B4**: **Distribution of Post-Signal WTP for RCT**

**Panel A**: Private Spending



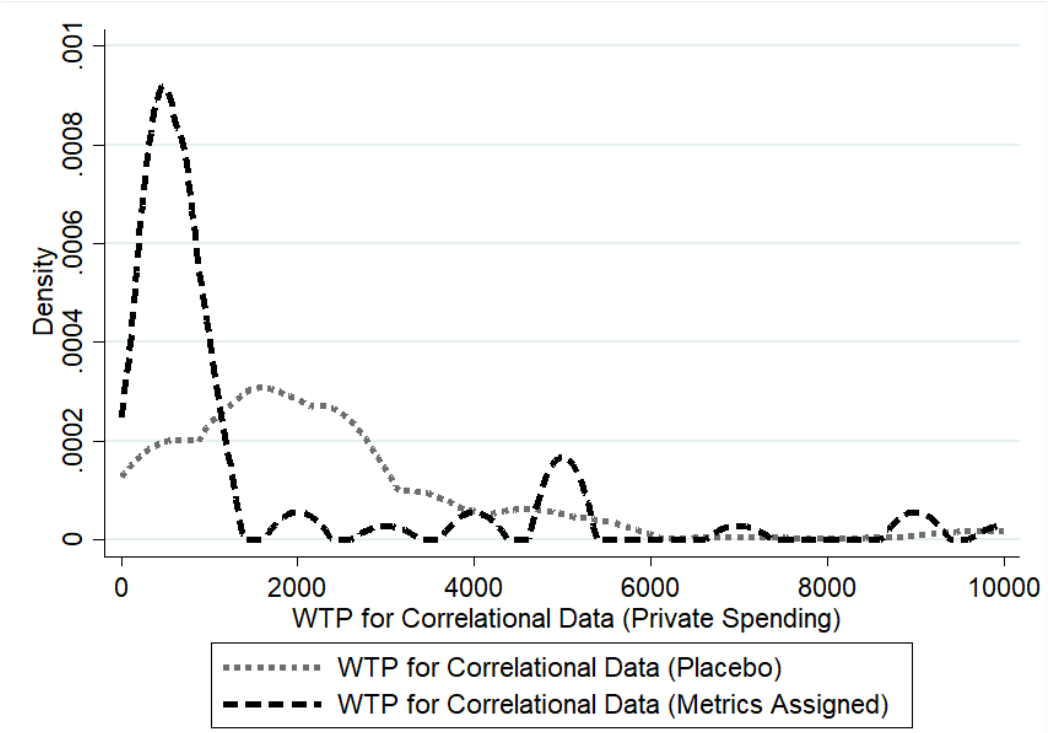

**Panel B**: Public Spending



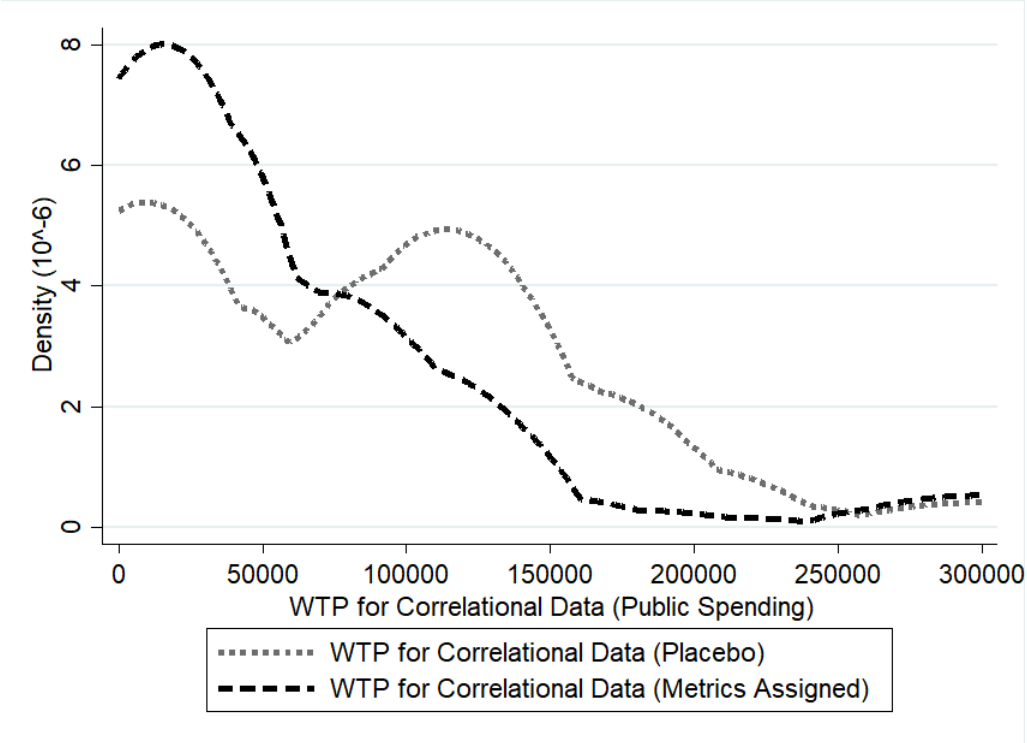

*Note*: The figure plots distribution of WTP in Pakistani Rupees for policymaker to obtain estimate from an RCT for a policy decision (choosing deworming or computer lab project). Panel A is for private spending whereas Panel B is the willingness to pay from the public exchequer.

**Figure B5**: **Distribution of Post-Signal WTP for Correlational Data**

**Panel A**: Private Spending



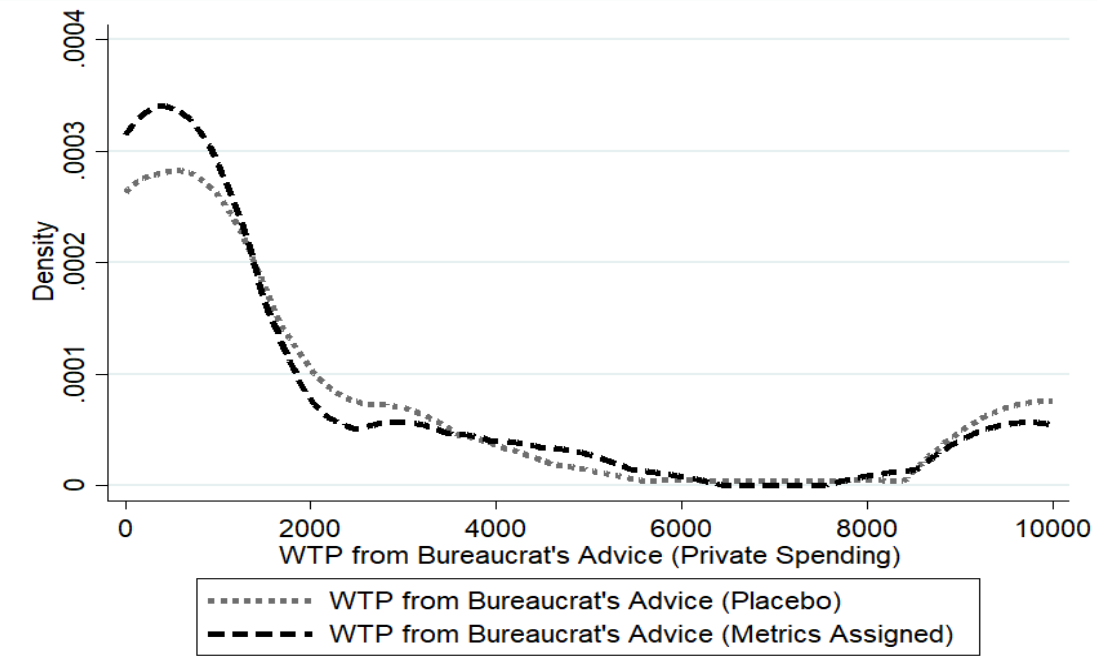

**Panel B**: Public Spending



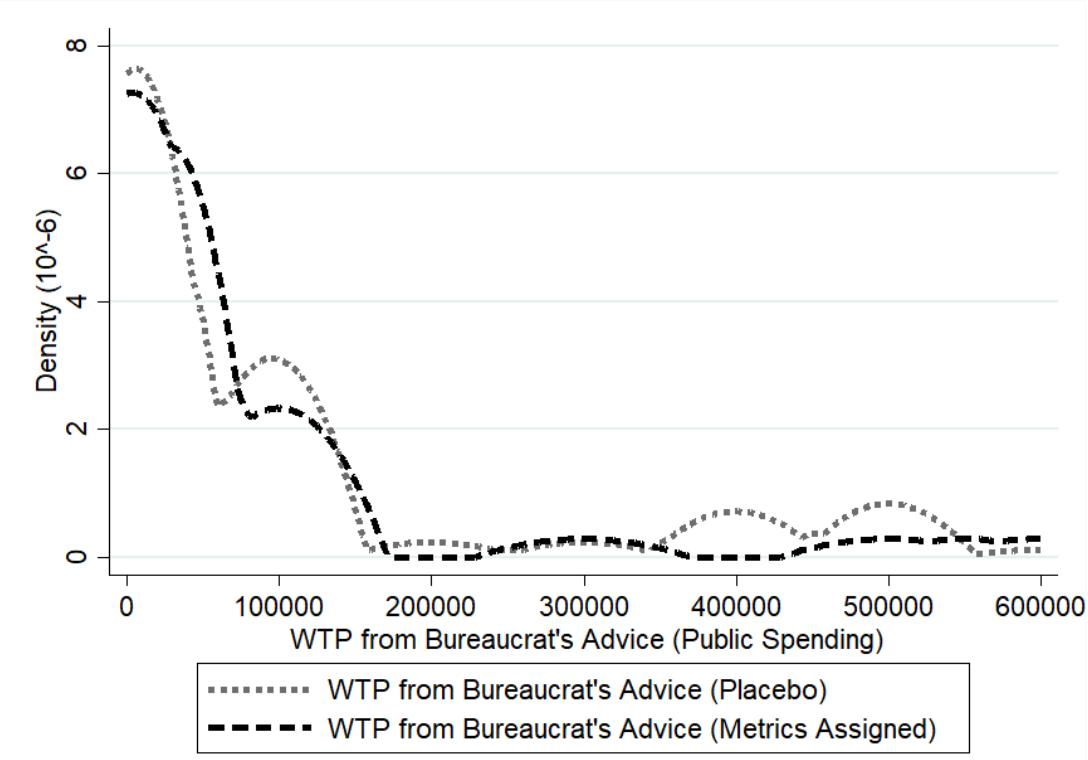

*Note*: The figure plots distribution of WTP in Pakistani Rupees for policymaker to obtain relevant correlational data for a policy decision (choosing deworming or computer lab project). Panel A is for private spending whereas Panel B is the willingness to pay from the public exchequer.

**Figure B6**: **Distribution of Post-Signal WTP for Bureaucrat's Advice**

**Panel A**: Private Spending



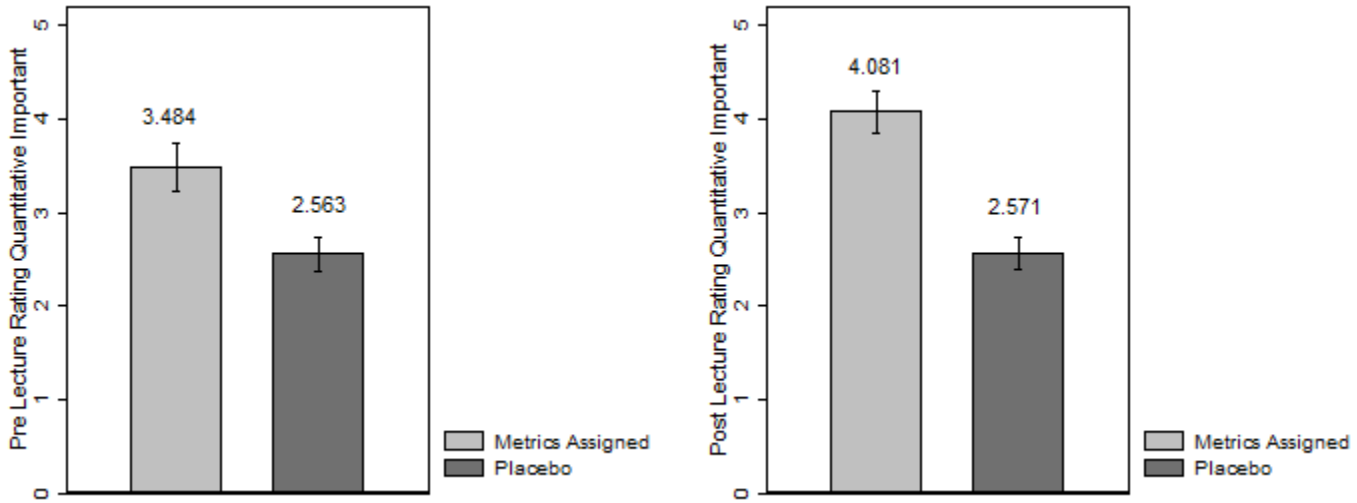

**Panel B**: Public Spending



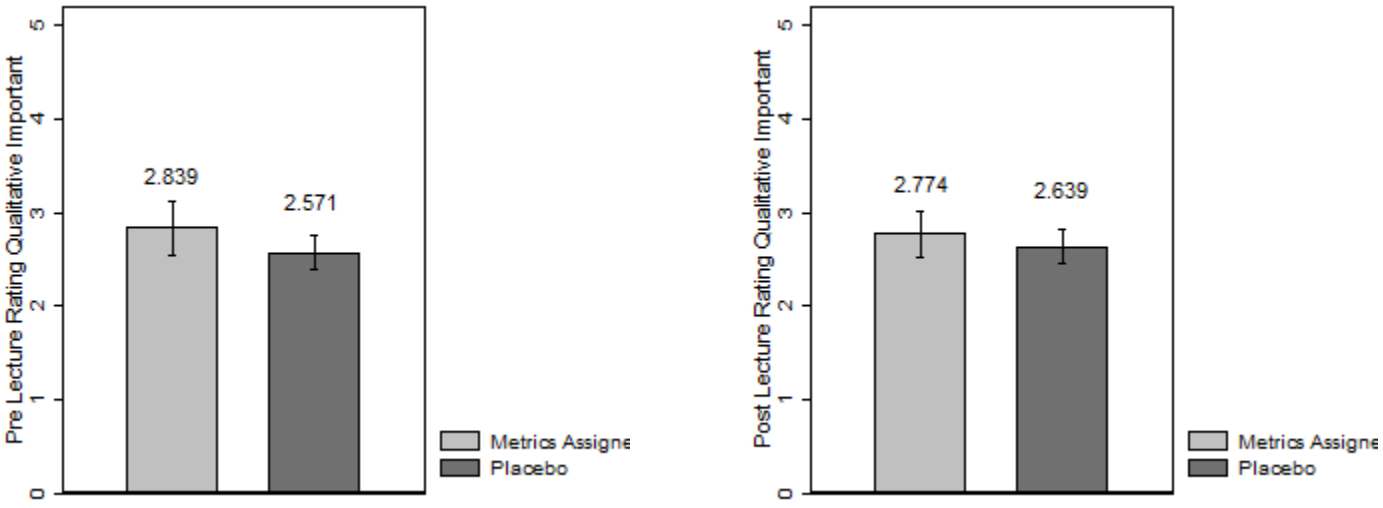

*Note*: The figure plots distribution of WTP in Pakistani Rupees for policymaker to obtain expert bureaucrat's advice on a policy decision (choosing deworming or computer lab project). Panel A is for private spending whereas Panel B is the willingness to pay from the public exchequer.

## Figure B7: Impact of Metrics Training

**Panel A:** Beliefs on Importance of Quantitative Evidence



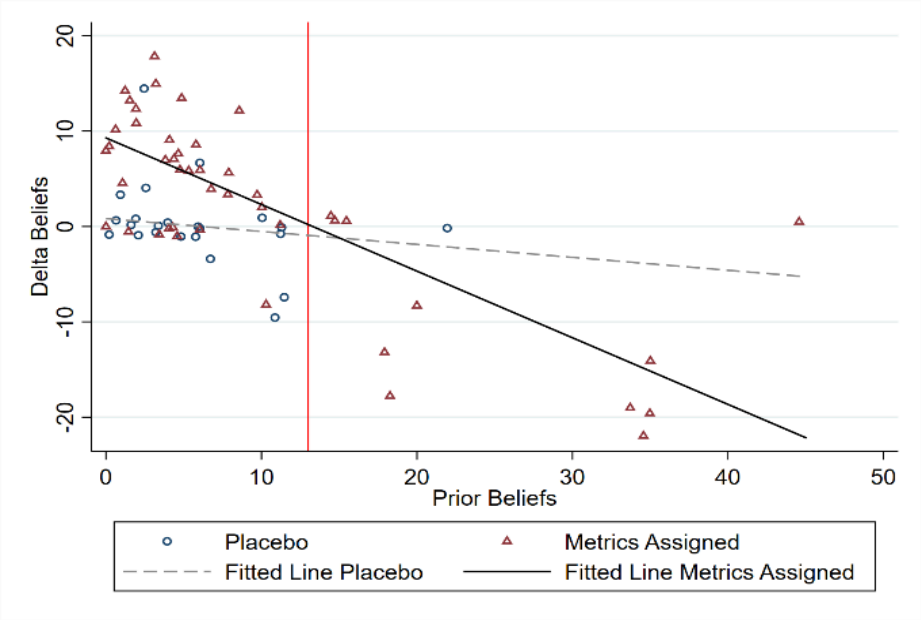

**Panel B:** Beliefs on Importance of Qualitative Evidence



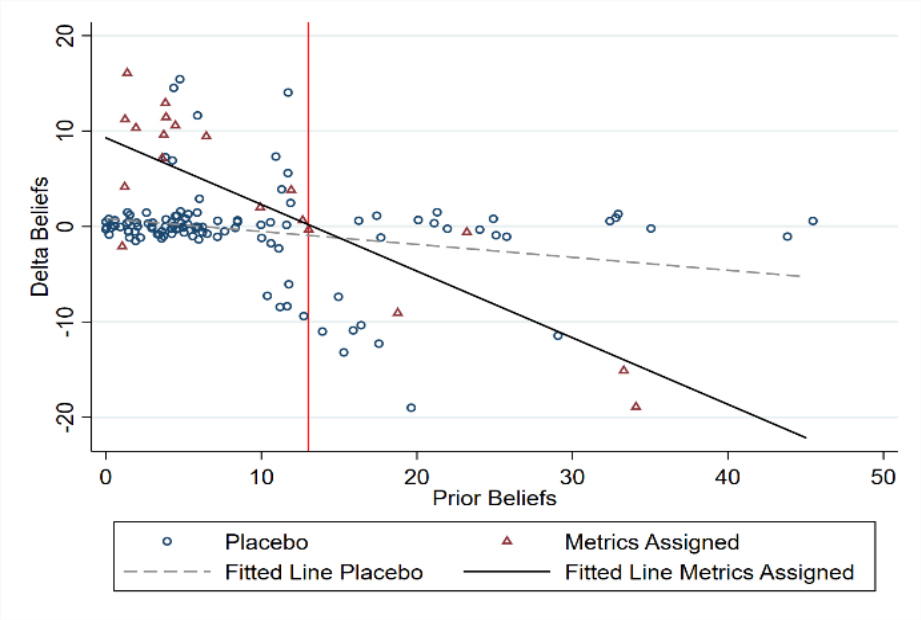

*Note*: The figures above compare beliefs on importance of quantitative (Panel A) and qualitative (Panel B) evidence for those before and after lecture and discussion for treated and control group. The left panels are those treated with metrics book and corresponding assignments, while figures on the right are post lecture and discussion ratings. The results on post-lecture correspond to participants attending the complete metrics training: reading the Mastering metrics book, completing corresponding assignments, attending a lecture and participating in a discussion on the content. Panel A presents results on rating of quantitative evidence and Panel B presents results on rating of importance of qualitative evidence for our metrics trained relative to the placebo group.

**Figure B8: Shifts in Beliefs versus Initial Beliefs by Metrics Selection**

**Panel A**: Metrics Selected



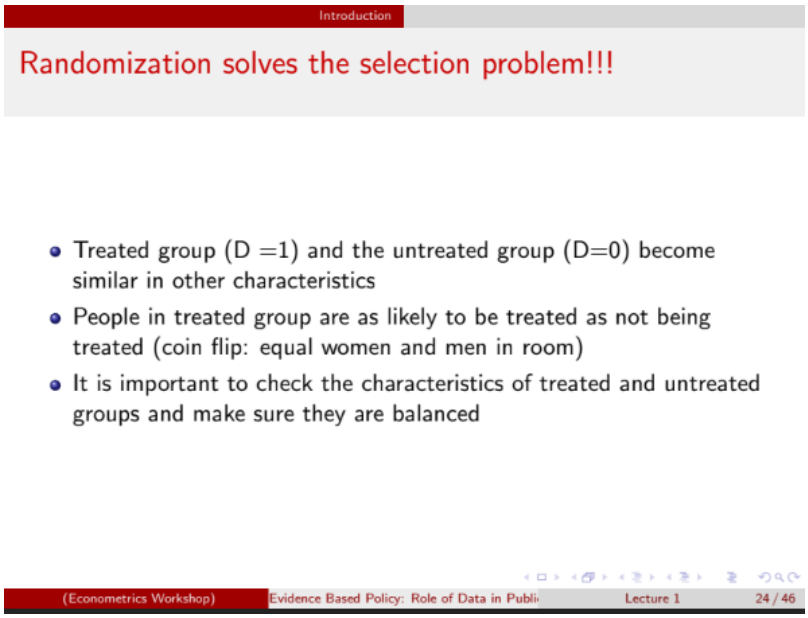

**Panel B**: Placebo Selected



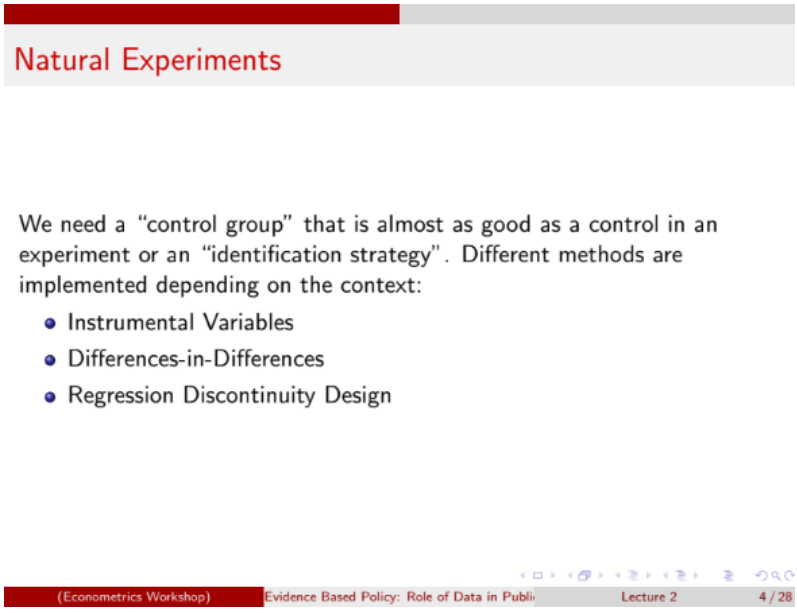

*Note*: The figure above plots shifts in beliefs from initial to post-signal belief relative to initial beliefs for individuals assigned metrics training and those assigned placebo training. Line of best fits are also reported. Panel A provides the plots for those who selected the metrics training, while Panel B provides the plots for those who selected the placebo training. Fitted lines using *lfit* in Stata are shown for both metrics trained treated and placebo group. Vertical line represents the signal value of 13% impact of deworming on income.

**Figure B9: Slides of Tax Offiers Workshop**

Panel A: A slide from Lecture 1




Panel B: A slide from Lecture 2




Note: Two representative slides from Lecture 1 and 2 in Panel A and B, respectively, are presented in the figure above. The link to access the complete slide deck for both Lecture 1 and Lecture 2 can be found HERE. Placebo lecture on macroeconomics can be found HERE.

**Figure B10: Tax Officers Econometrics Workshop**

Panel A: Illustration 1



Panel B: Illustration 2




Note: Two images from the tax officers' workshop are displayed in this figure.

**Figure B11: Causal Signal Distributed to Tax Officer**

July 5, 2022

# BEHAVIORAL INTERVENTIONS IN TAX COMPLIANCE

## Evidence from Guatemala

## Policy Insights

- Sending Tax Notices or Tax Reminders increase tax collection.
- The individuals assigned Tax Reminder letters (notifications) increased tax compliance relative to those that were assigned no letters in control
- The effects are large and statistically significant.

## Study Summary

In 2014, taxpayers (individuals and firms) who had failed to pay their income tax for the 2013 tax year Guatemala were randomly allocated to receive either no letter or tax reminder letters.

The randomized evaluation finds that although all letters increased the rate of declaration, the best performing treatments were a deliberate choice letters framing non-declaration as an intentional and deliberate choice.

These interventions increased the rate of payment as well as the average amount paid conditional on paying.

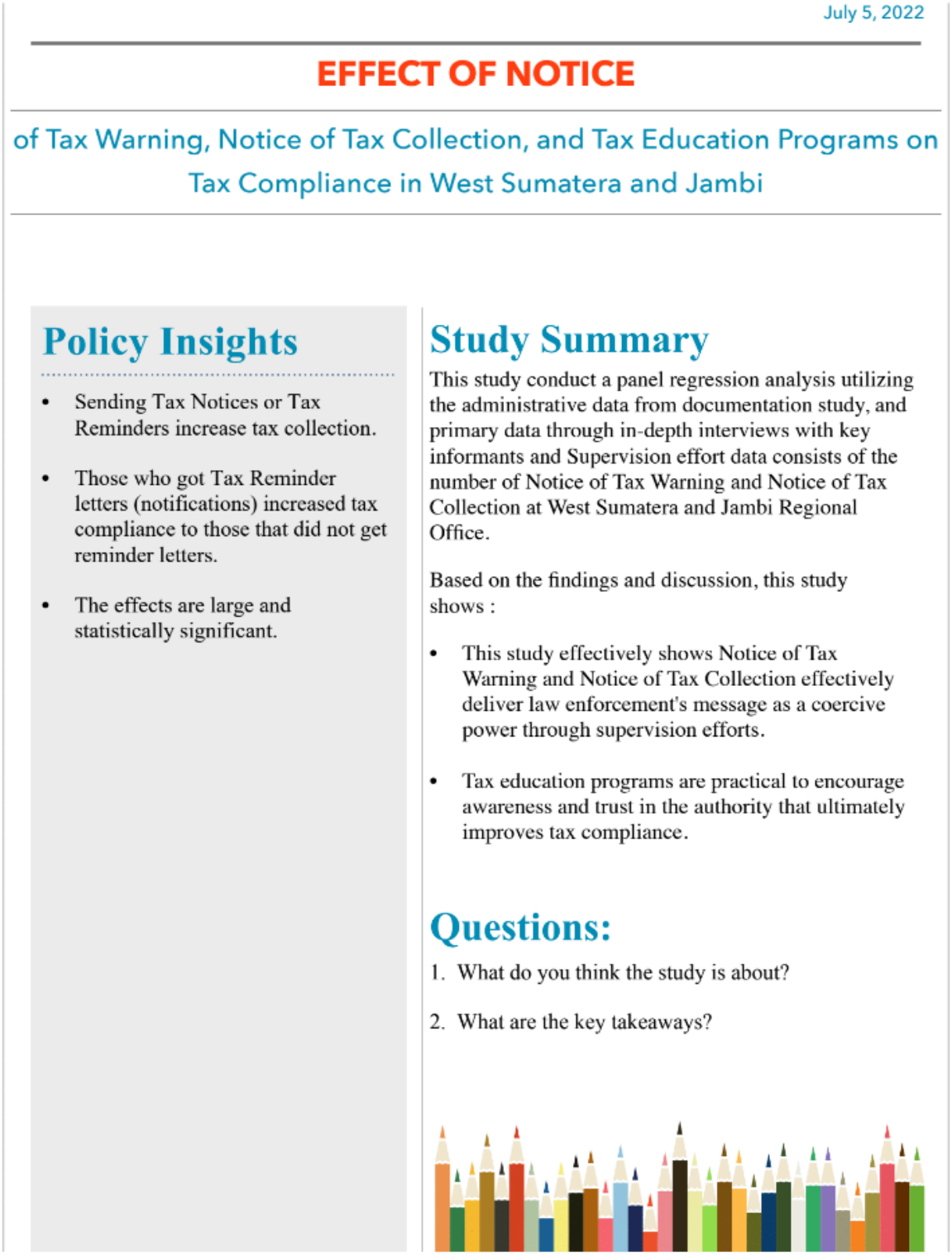
## Questions:

1. What do you think the study is about?
2. What are the key takeaways?



Note: The figure is an illustration of a causal handout seen by the tax officers. The study used to prepare this document causal handout is the following: Kettle, S., Hernandez, M., Ruda, S. and Sanderson, M.A., 2016. Behavioral interventions in tax compliance: Evidence from Guatemala. World Bank Policy Research Working Paper, (7690).

**Figure B12: Correlational Signal Assigned to Tax Officer**



July 5, 2022

# EFFECT OF NOTICE

## of Tax Warning, Notice of Tax Collection, and Tax Education Programs on Tax Compliance in West Sumatera and Jambi

### Policy Insights

- Sending Tax Notices or Tax Reminders increase tax collection.
- Those who got Tax Reminder letters (notifications) increased tax compliance to those that did not get reminder letters.
- The effects are large and statistically significant.

### Study Summary

This study conduct a panel regression analysis utilizing the administrative data from documentation study, and primary data through in-depth interviews with key informants and Supervision effort data consists of the number of Notice of Tax Warning and Notice of Tax Collection at West Sumatera and Jambi Regional Office.

Based on the findings and discussion, this study shows :

- This study effectively shows Notice of Tax Warning and Notice of Tax Collection effectively deliver law enforcement's message as a coercive power through supervision efforts.
- Tax education programs are practical to encourage awareness and trust in the authority that ultimately improves tax compliance.

### Questions:

1. What do you think the study is about?
2. What are the key takeaways?

Note: The figure is an illustration of the correlational handout seen by the tax officers. The study used to prepare this document is the following: Budiman, I. and Inayati, I., 2021. Effect of Notice of Tax Warning, Notice of Tax Collection, and Tax Education Programs on Tax Compliance in West Sumatera and Jambi. Publik (Jurnal Ilmu Administrasi), 10(1), pp.45-63.

**Table B1: Letter of Support from Director of Training Academy**

GOVERNMENT OF PAKISTAN
CIVIL SERVICES ACADEMY
CTP-WING, WALTON CAMPUS, LAHORE
**********

**SUBJECT: LETTER OF SUPPORT**

The Civil Services Academy (CSA) is the primary institution in Pakistan responsible for post induction training of the elite civil servants up to rank of deputy ministers. Common Training Program (CTP) is the flagship training program of CSA. **Ms. Shaheen Naseer (Assistant Professor, Lahore School of Economics)** has been helping the Civil Services Academy in carrying out research related to various aspects of CTP and its various training components since 2019. She is on board with us for training workshops. This includes organization and design of **mastering metrics** workshop on training deputy ministers in causal inference.

Our collaboration with the researchers/ academia is primarily meant to improve the training courses being offered at civil services academy and it helps us in evidence-based design of the training courses.



**Table B2: Commemorative shields and vouchers**

**Panel A:** Commemorative Shield




*Note*: The figure shows one of the commemorative shields presented to the deputy ministers.

**Panel B:** Gift Vouchers




*Note*: The figure shows cash gift vouchers at a luxury departmental store. The monetary amount is designated in Pakistan Rupees. The vouchers for the first three positions within each treatment arm, and are worth about USD 150, USD 100, and USD 80, respectively.

**Table B3: Impact of Treatment on Attrition**

| | *Attrition in Sample 1* | | *Attrition in Sample 2* | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Metrics Assigned | -0.058 | -0.053 | 0.001 | 0.008 |
| | (0.045) | (0.047) | (0.031) | (0.035) |
| Individual Controls | No | Yes | No | Yes |
| Observations | 213 | 213 | 213 | 213 |
| R-squared | 0.01 | 0.09 | 0.014 | 0.052 |
| Mean of dep. var. (placebo) | 0.091 | 0.091 | 0.06 | 0.06 |

Robust standard errors (clustered at the individual level) appear in parentheses. The dependent variable in Columns (1) and (2) are dummies that switch on when there is attrition in the sample for quantitative and qualitative evidence responses and the dependent variable in Column (3) and (4) are dummies that switch on when there is attrition in the sample for deworming and willingness to pay responses. Metrics assigned is a dummy variable that switch on when causal inference book is randomly assigned to participants. Consistent with all our earlier regressions, we always control for metrics chosen. The causal inference book, Mastering 'Metrics, is randomly assigned conditional on the book being chosen. The metrics training is randomly assigned conditional on it being chosen. The estimations obtained from OLS regressions includes the following controls: metrics chosen, written test scores, interview test scores, gender, birth in political capitals, asset ownership, income before joining public service, age, education, foreign visits and occupational group dummies. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

**Table B4: The Effect of Metrics Training on Beliefs – Original Units**

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
|---|---|---|---|---|---|---|---|---|
| | Pre-Lecture Rating Quantitative | Post-Lecture Rating Quantitative | Pre-Lecture Rating Qualitative | Post Lecture Rating Qualitative | Pre-Lecture Run RCT | Post-Lecture Run RCT | Pre-Lecture Why Run RCT | Post-Lecture Why Run RCT |
| Metrics Assigned | 0.912*** | 1.538*** | 0.136 | 0.122 | 0.167** | 0 .220** | 0.151* | 0.153* |
| | (0.176) | (0.178) | (0.196) | (0.206) | (0.082) | (0.085) | (0.087) | (0.087) |
| Controls | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Observations | 190 | 190 | 190 | 190 | 190 | 190 | 190 | 190 |
| Mean of dep. variable | 2.745 | 2.979 | 2.490 | 2.596 | 0.362 | 0.404 | 0.396 | 0.396 |

Robust standard errors clustered at individual level appear in parantheses. In Columns 1-2 dependent variable is a rating on a scale of 1 to 5 on the statement "How important do you think quantitative analysis is in public policy making?" In Columns 3 and 4 the dependent variable is a rating on "How important do you think qualitative analysis is in public policy making? "While in Columns 5 and 6 dependent variable is constructed from "You are in charge of assigning people to a public policy program and before rolling it out, you want to learn if the policy is effective, what you would do?" One of the options is to "Run a randomized control trial" while the other options are "Survey feelings of respondents regarding the policy", and "Survey if there is demand for policy" and option 4 is to "compare two groups of people who had previously benefited most from the policy with those that did not?" The dependent variable takes the value of 1 if the run a randomized control trial option is chosen and zero otherwise. In Columns 7 and 8 the dependent variable is constructed based on the question: "Continuing with previous example, why the previous answer makes sense?": (1) Because people in a RCT are apples to apples comparisons (2) People feelings are important determinant whether the public policy will work (3) Survey methods are known to produce causal effects (4) Comparing two groups of non-randomly selected people allows us to infer causality. The dependent variable takes the value of one if option 1 is chosen and 0 otherwise. Beliefs or ratings are measured before and after the lecture. Metrics assigned is a dummy that switches on when a causal inference book is randomly assigned to participants. The metrics training is randomly assigned conditional on it being chosen. The results on post-lecture correspond to participants attending the complete metrics training: reading the Mastering metrics book, completing corresponding assignments, attending a lecture and participating in a discussion on the content. *** p<0.01, ** p<0.05, * p<0.1

**Table B5: Impact of Treatment by Metrics Chosen**

| | *Rating Quantitative* | *Run RCT* | *Why Run RCT* | ***Private Spending*** *Amount Randomized Trial* | ***Public Spending*** *Amount Randomized Trial* | *Assessment Research Methods* |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6 ) |
| Metrics Assigned X Metrics Chosen | -0.127 | -0.138 | -0.124 | 1218.829 | 1591257 | 0.481 |
| | (0.351) | (0.171) | (0.175) | (1286.544) | (1505679) | (0.353) |
| Metrics Assigned | 1.605*** | 0.290** | 0.218* | 1414.802* | 545008.5 | 0.537** |
| | (0.240) | (0.122) | (0.125) | (737.532) | (1300118) | (0.258) |
| Individual Controls | Yes | Yes | Yes | Yes | Yes | Yes |
| Observations | 190 | 190 | 190 | 180 | 180 | 190 |
| R-squared | 0.426 | 0.101 | 0.090 | 0.221 | 0.098 | 0.228 |
| Mean of dep. var. (placebo) | 2.657 | 0.400 | 0.379 | 1539.453 | 928546.1 | -0.317 |

Robust standard errors clustered at individual level appear in parantheses. In the first column, the dependent variable is a rating on a scale of 1 to 5 with 1 being not important at all and 5 being very important on the statement "How important do you think quantitative analysis is in public policy making?" In the second column, a dependent variable is constructed based on the statement "You are in charge of assigning people to a public policy program and before rolling it out, you want to learn if the policy is effective, what you would do?" These and other dependent variables are identical to those reported and explained in the accompanied notes of Table 2-4. Metrics assigned is a dummy variable that switches on when a causal inference book is randomly assigned to participants. The metrics book is randomly assigned conditional on it being chosen. The estimations obtained from OLS regressions include the following controls: metrics chosen, written test scores, interview test scores, gender, birth in political capitals, asset ownership, income before joining public service, age, education, foreign visits and occupational group dummies. *** p<0.01, ** p<0.05, * p<0.1.

**Table B6: Effect of Metrics Training on Willingness to Pay**

| | (1) Amount Randomized Trial | (2) Amount Correlational Data | (3) Amount Expert Bureaucrat | (4) Amount Randomized Trial | (5) Amount Correlational Data | (6) Amount Expert Bureaucrat |
|---|---|---|---|---|---|---|
| Metrics Assigned | 2,063*** | -1,020*** | -1,986 | 1391308** | -35,274*** | -2,048 |
| | (587.5) | (340.4) | (1,390) | (665,160) | (13,033) | (34,090) |
| Controls | Yes | Yes | Yes | Yes | Yes | Yes |
| Observations | 180 | 180 | 180 | 180 | 180 | 180 |
| R-squared | 0.217 | 0.202 | 0.104 | 0.094 | 0.148 | 0.080 |
| Mean of dep. variable | 1539.453 | 2214.07 | 4490.008 | 928546.1 | 94136.72 | 115937.5 |

Robust standard errors clustered at individual level appear in parantheses. Dependent variables are the stated willingness to pay from private funds in columns 1, 2 and 3 and thestated willingness to pay from public funds in columns 4, 5 and 6 = for a randomized control trial, correlational comparison of means data, and suggestion from expert bureaucrat respectively. The metrics training is randomly assigned conditional on it being chosen. The estimations obtained from OLS regressions includes the following controls: metrics chosen, written test scores, interview test scores, gender, birth in political capitals, asset ownership, income before joining public service, age, education, foreign visits and occupational group dummies. All dependent variables are denominated in Pakistani Rupees. *** p<0.01, ** p<0.05, * p<0.1.

**Table B7: Effect of Treatment on WTP by Defiers and Never Takers**

| | *Private Spending* | | *Public Spending* | |
|---|---|---|---|---|
| | Amount Randomized Trial | Amount Randomized Trial | Amount Randomized Trial | Amount Randomized Trial |
| | (1) | (2) | (3) | (4) |
| Defiers X Metrics Assigned | -888.410 | | -2531651* | |
| | (1301.347) | | (1367816) | |
| Never Takers X Metrics Assigned | | -1987.019 | | -1304302 |
| | | (1837.863) | | (1558128) |
| Metrics Assigned | 2179.581*** | 2165.335*** | 1775415*** | 15028283** |
| | (648.553) | (641.327) | (635063.5) | (702559.8) |
| Defiers X Metrics Chosen | -428.640 | | 3326010 | |
| | (1207.932) | | (2437943) | |
| Never Takers X Metrics Chosen | | -643.626 | | 2190478* |
| | | (1385.195) | | (1303170) |
| Defiers | -463.802 | | -3101800 | |
| | (857.643) | | (2201009) | |
| Never Takers | | 1756.372 | | -1250269 |
| | | (1887.535) | | (1201626) |
| p-values: | (1) = (2): | {0.590} | (3) = (4): | {0.532} |
| Individual Controls | Yes | Yes | Yes | Yes |
| Observations | 180 | 180 | 180 | 180 |
| R-squared | 0.224 | 0.223 | 0.114 | 0.098 |
| Mean of dep. var. (placebo) | 1539.453 | 1539.453 | 928546.1 | 928546.1 |

Robust standard errors clustered at individual level appear in parentheses. In first two columns, the private willingness to pay for a randomized control trial. In the last two columns, the willingness to pay from public funds or government budget is elicited for the same piece of information. The metrics training is randomly assigned conditional on it being chosen. Defiers are all participants that clicked on the opposite lecture link, while never takers were assigned the book but chose to not to click any link. p-values testing equality of coefficients of Metrics Assigned X Defiers = Metrics Assigned X Never Takers are presented in curly brackets. The estimations obtained from OLS regressions include the following controls: metrics chosen, written test scores, interview test scores, gender, birth in political capitals, asset ownership, income before joining public service, age, education, foreign visits and occupational group dummies. All dependent variables are denominated in Pakistani Rupees *** p<0.01, ** p<0.05, * p<0.1.

**Table B8: Heterogeneity by Pretreatment Quantitative Assessment Scores**

| | *Rating Quantitative* | | *Run RCT* | | *Amount Randomized Trial (Private Spending)* | | *Amount Randomized Trial (Public Spending)* | |
|---|---|---|---|---|---|---|---|---|
| Quantitative Scores | *Below Median* | *Above Median* | *Below Median* | *Above Median* | *Below Median* | *Above Median* | *Below Median* | *Above Median* |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Metrics Assigned | 1.865*** | 0.942*** | 0.289** | 0.259 | 447.309 | 843.157 | -511279.2 | -284572.3 |
| | (0.244) | (0.282) | (0.145) | (0.119) | (1080.277) | (642.236) | (1312014) | (453235.8) |
| Metrics Assigned Equal (p-value): | (1) = (2): | [0.0135] | (3) = (4): | [0.8722] | (5) = (6): | [0.7528] | (7) = (8): | [0.8703] |
| Individual Controls | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Observations | 79 | 80 | 79 | 80 | 79 | 80 | 79 | 80 |
| R-squared | 0.469 | 0.594 | 0.246 | 0.168 | 0.310 | 0.257 | 0.167 | 0.285 |

Robust standard errors clustered at individual level appear in parentheses. The baseline regressions are run on individuals with quantitative scores below and above median respectively, for our main outcomes of interest. Similar results are found for other variables and available on request. p-values corresponding to equality of coefficients are reported in square brackets. Metrics assigned switch on when metrics book is randomly assigned. All regressions always contain the metrics chosen dummy as in baseline regressions i.e. a dummy variable that switches on when the metrics book is chosen. These results on post-lecture correspond to participants attending the complete metrics training: reading the Mastering metrics book, completing corresponding assignments, attending a lecture and participating in a discussion on the content. The estimations obtained from OLS regressions include the following controls: metrics chosen, written test scores, interview test scores, gender, birth in political capitals, asset ownership, income before joining public service, age, education, foreign visits and occupational group dummies. *** p<0.01, ** p<0.05, * p<0.1.

**Table B9: The Effect of Metrics Training on Beliefs – Randomization Inference**

| | *Pre Lecture Rating Quantitative* | *Post Lecture Rating Quantitative* | *Pre Lecture Rating Qualitative* | *Post Lecture Rating Qualitative* | *Pre Lecture Run RCT* | *Post Lecture Run RCT* | *Pre Lecture Why Run RCT* | *Post Lecture Why Run RCT* |
|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Metrics Assigned | 0.912 | 1.538 | 0.136 | 0.122 | 0.166 | 0 .220 | 0.151 | 0.153 |
| | (0.001) *** | (0.001) *** | (0.487) | (0.554) | (0.046) ** | (0.011) ** | (0.085) * | (0.080) * |
| | {0.001} *** | {0.001} *** | {0.491} | {0.533} | {0.075} * | {0.012} ** | {0.104} | {0.096} * |
| Individual Controls | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Observations | 190 | 190 | 190 | 190 | 190 | 190 | 190 | 190 |
| Mean of dep. var. (placebo) | 2.657 | 2.658 | 2.656 | 2.714 | 0.400 | 0.400 | 0.386 | 0.379 |

p-values from our baseline regressions appear in parentheses for comparison, while p-values from randomization inference due to Heß (2017) are reported in curly brackets. In Columns 1-2 dependent variable is a rating on a scale of 1 to 5 with 1 being not important at all and 5 being very important on the statement "How important do you think quantitative analysis is in public policy making?" Likewise, in Columns 3 and 4 dependent variable is a rating on the statement "How important do you think qualitative analysis is in public policy making? In Columns 5 and 6 the dependent variable is the rating based on the statement "You are in charge of assigning people to a public policy program and before rolling it out, you want to learn if the policy is effective, what you would do?" One of the options is to "Run a randomized control trial or a pilot" while the other options are "Survey feelings of respondents regarding the policy", and "Survey if there is demand for policy" and option 4 is to "compare two groups of people who had previously benefited most from the policy with those that did not?" The dependent variable takes the value of 1 if the run a randomized control trial option is chosen and zero otherwise. In Columns 7 and 8 the dependent variable is constructed based on the question: "Continuing with previous example, why the previous answer makes sense?": 1) Because people in a RCT are apples to apples comparisons 2) People feelings are important determinant whether the public policy will work 3) Survey methods are known to produce causal effects 4) Comparing two groups of non-randomly selected people allows us to infer causality. The dependent variable takes the value of 1 if option 1 is chosen and zero otherwise. Every time the beliefs or ratings are measured before and after the lecture. Metrics assigned is a dummy variable that switch on when causal inference book is randomly assigned to participants. The metrics book is randomly assigned conditional on it being chosen. All estimations include the following controls: metrics chosen, written test scores, interview test scores, gender, birth in political capitals, asset ownership, income before joining service, age, education, foreign visits and occupational group dummies.*** p<0.01, ** p<0.05, * p<0.1.

**Table B10: Impact of Metrics Training on Policy Making Course Grades – Randomization Inference**

| | *Policy Assessment* | *Research Methods Assessment* | *Teamwork Assessment* |
|---|---|---|---|
| | (1) | (2) | (3) |
| Metrics Assigned | 0.505 | 0.799 | 0.017 |
| | (0.003) *** | (0.001) *** | (0.931) |
| | {0.007} *** | {0.001} *** | {0.930} |
| Individual Controls | Yes | Yes | Yes |
| Observations | 190 | 190 | 190 |
| R-squared | 0.171 | 0.219 | 0.123 |

p-values from our baseline regressions appear in parentheses for comparison, while p-values from randomization inference due to Heß (2017) are reported in curly brackets. The dependent variables are standardized scores with mean 0 and standard deviation 1 from regular public policy training workshops at the training Academy. Column (1) is the score on the workshop called Public Sector Economics, Public Goods and Publicly Provided Private Goods that focuses on case studies of past (actual) decisions of similar policymakers. The course content covers scenarios that apply concepts of public goods, externalities, the use of data in policymaking. Columns (2) presents scores on Public Sector Management Committees, Teams & Group Decisions course that simulate real decisions these policymakers make in the field. Both teamwork and group decisions are marked by a committee of senior bureaucrats. Finally, in Columns (3) score on Research Method is reported. The workshop content included a statistical inference course with emphasis on hypothesis testing, multivariate regression analysis with applications to policy-making and a particular focus on randomized evaluation trials as a "gold standard". Metrics assigned is a dummy variable that switches on when a causal inference book is randomly assigned to participants. Consistent with all our earlier regressions, we always control for metrics chosen. The causal inference book, Mastering Metrics, is randomly assigned conditional on the book being chosen. The estimations obtained from OLS regressions include the following controls: metrics chosen, written test scores, interview test scores, gender, birth in political capitals, asset ownership, income before joining public service, age, education, foreign visits and occupational group dummies. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

## Table B11: Effect of Metrics Training on Willingness to Pay – Randomization Inference

*Panel A: Private Spending*

| | *Before Signal Amount Randomized Trial* | *Before Signal Amount Correlational Data* | *Before Signal Amount Expert Bureaucrat* | *After Signal Amount Randomized Trial* | *After Signal Amount Correlational Data* | *After Signal Amount Expert Bureaucrat* |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Metrics Assigned | 1486.662 | -963.938 | 46.442 | 2063.028 | -1020.318 | -1986.22 |
| | (0.035) ** | (0.121) | (0.781) | (0.001) *** | (0.003) *** | (0.155) |
| | {0.079} * | {0.079} * | {0.119} | {0.791} | {0.005} *** | {0.254} |
| Individual Controls | Yes | Yes | Yes | Yes | Yes | Yes |
| Observations | 180 | 180 | 180 | 180 | 180 | 180 |
| R-squared | 0.108 | 0.108 | 0.074 | 0.217 | 0.202 | 0.104 |
| Mean of dep. var. (placebo) | 2503.594 | 2267.188 | 430.977 | 1539.453 | 2214.07 | 4490.008 |

*Panel B: Public Spending*

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Metrics Assigned | 757363.300 | -53577.520 | 10256.68 | 1391308 | -35273.920 | -2047.595 |
| | (0.001) *** | (0.000) *** | (0.835) | (0.038) ** | (0.007) *** | (0.952) |
| | {0.000} *** | {0.001} *** | {0.873} | {0.151} | {0.072} * | {0.977} |
| Individual Controls | Yes | Yes | Yes | Yes | Yes | Yes |
| Observations | 180 | 180 | 180 | 180 | 180 | 180 |
| R-squared | 0.235 | 0.17 | 0.069 | 0.094 | 0.148 | 0.080 |
| Mean of dep. var. (placebo) | 224604.7 | 109746.9 | 168464.8 | 928546.1 | 94136.72 | 115937.5 |

p-values from our baseline regressions appear in parentheses for comparison, while p-values from randomization inference due to Heß (2017) are reported in curly brackets. In Panel A, the private willingness to pay for a randomized control trial, correlational comparison of means data, suggestion from expert bureaucrat is elicited before and after the signal of effect of deworming on hourly wages is revealed to all participants. In Panel B, the willingness to pay from public funds or government budget is elicited for the same pieces of information. The metrics training is randomly assigned conditional on it being chosen. The estimations obtained from OLS regressions include the following controls: metrics chosen, written test scores, interview test scores, gender, birth in political capitals, asset ownership, income before joining public service, age, education, foreign visits and occupational group dummies. All dependent variables are denominated in Pakistani Rupees. *** p<0.01, ** p<0.05, * p<0.1.

**Table B12: The Effect of Metrics Training – Interaction Specification**

| | *Pre-Lecture Rating Quantitative* | *Post Lecture Rating Quantitative* | *Pre-Lecture Rating Qualitative* | *Post Lecture Rating Qualitative* | *Pre-Lecture Run RCT* | *Post Lecture Run RCT* | *Pre-Lecture Why Run RCT* | *Post Lecture Why Run RCT* |
|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Metrics Assigned | 1.133*** | 1.605*** | 0.085 | 0.090 | 0.274** | 0.290** | 0.220* | 0.218* |
| | (0.229) | (0.240) | (0.276) | (0.316) | (0.110) | (0.122) | (0.125) | (0.125) |
| Metrics Assigned X Metrics Chosen | -0.423 | -0.127 | 0.098 | 0.062 | -0.214 | -0.138 | -0.131 | -0.124 |
| | (0.344) | (0.351) | (0.380) | (0.408) | (0.165) | (0.171) | (0.175) | (0.175) |
| Individual Controls | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Observations | 190 | 190 | 190 | 190 | 190 | 190 | 190 | 190 |
| Mean of dep. var. (placebo) | 2.657 | 2.657 | 2.657 | 2.714 | 0.400 | 0.400 | 0.386 | 0.379 |

Robust standard errors clustered at individual level appear in parentheses. The dependent variables are identical to those in Table 2. The key difference of this table with respect to Table 2 is the additional interaction term between Metrics Assigned X Metrics Chosen Metrics assigned is a dummy variable that switch on when causal inference book is randomly assigned to participants. The metrics training is randomly assigned conditional on it being chosen. The results on post-lecture correspond to participants attending the complete metrics training: reading the Mastering metrics book, completing corresponding assignments, attending a lecture and participating in a discussion on the content. The estimations obtained from OLS regressions includes the following controls: metrics chosen, written test scores, interview test scores, gender, birth in political capitals, asset ownership, income before joining public service, age, education, foreign visits and occupational group dummies. *** p<0.01, ** p<0.05, * p<0.1.

**Table B13: Robustness to Multiple Hypothesis Testing**

| | *Lecture Rating Quantitative* | *Amount Randomization Trial* | *Amount Correlational Data* | *Assessment Public Policy* | *Assessment Research Methods* |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| Metrics assigned | 1.541 | 1391308 | -35,274 | 0.505 | 0.799 |
| p-value | (0.001) *** | (0.038) ** | (0.007) *** | (0.003) *** | (0.001) *** |
| Sharpened q-value | [0.002] *** | [0.009] *** | [0.004] *** | [0.002] *** | [0.002] *** |
| Observations | 190 | 180 | 180 | 190 | 190 |
| R-squared | 0.425 | 0.094 | 0.148 | 0.071 | 0.218 |
| Mean of dep. var. (Placebo) | 2.745 | 928546.1 | 94136.720 | 0.171 | 0.219 |

Note: p-values from our baseline regressions appear in parentheses for comparison, while Anderson q-values and are reported in square and curly brackets, respectively. The dependent variables in Column (1) is a rating on a scale of 1 to 5 with 1 being not important at all and 5 being very important on the statement "How important do you think quantitative analysis is in public policy making?". In Columns 2 and 3 the dependent variables are letters sent and funds requested from Pakistan's Finance Ministry (in Pakistani Rupees) for allocation of budget for deworming policy. Columns 4 and 5 have willingness-to-pay from public funds for randomized control trials and correlational comparison of means evidence, respectively. Column 6 contains the assessment scores on the courses *Public Sector Economics, Public Goods and Publicly Provided Private Goods* that consisted of case studies and analysis of past (actual) decisions of similar policymakers. The course content covers scenarios that apply concepts of public goods, externalities, the use of data in policymaking. Column 7 present scores on *Research Methods* are reported. This assessment scores decisions pertaining to teamwork and group work these policymakers typically make in the field. Metrics assigned is a dummy variable that switches on when a causal inference book is randomly assigned to participants. Consistent with all our earlier regressions, we always control for metrics chosen. The causal inference book, Mastering Metrics, is randomly assigned conditional on the book being chosen. The estimations obtained from OLS regressions includes the following controls: metrics chosen, written test scores, interview test scores, gender, birth in political capitals, asset ownership, income before joining public service, age, education, foreign visits, and occupational group dummies. *** p<0.01, ** p<0.05, * p<0.1.

## Table B14: Balance with interactions

| | (1) Gender | (2) Age | (3) Land | (4) Income | (5) Birth capital | (6) Religion | (7) Experience | (8) Pre-treatment Revenue | (9) Pre-treatment Notification | (10) Metrics Assigned |
|---|---|---|---|---|---|---|---|---|---|---|
| Metrics Assigned | 0.0313 | -0.415 | 0.0746 | 1,634 | 0.0534 | -0.0342 | -0.106 | -2.060 | -0.0176 | |
| | [0.0763] | [0.387] | [0.0723] | [1,360] | [0.0761] | [0.0291] | [0.366] | [19.36] | [0.0817] | |
| Causal Signal | 0.0145 | 0.0290 | 0.0335 | 516.3 | -0.0328 | -0.0390 | -0.739** | 9.084 | 0.0260 | |
| | [0.0790] | [0.406] | [0.0724] | [1,328] | [0.0734] | [0.0302] | [0.372] | [24.34] | [0.0816] | |
| Metrics Assigned x Causal Signal | -0.114 | 0.0917 | -0.0498 | -667.4 | -0.110 | -0.00347 | 0.225 | -17.33 | -0.0290 | |
| | [0.110] | [0.535] | [0.103] | [1,927] | [0.103] | [0.0520] | [0.504] | [28.69] | [0.115] | |
| Gender | | | | | | | | | | -0.0314 |
| | | | | | | | | | | [0.0302] |
| Birth capital | | | | | | | | | | -0.0341 |
| | | | | | | | | | | [0.0319] |
| Land | | | | | | | | | | -0.0155 |
| | | | | | | | | | | [0.0320] |
| Income | | | | | | | | | | -0.000000598 |
| | | | | | | | | | | [0.00000173] |
| Age | | | | | | | | | | 0.00107 |
| | | | | | | | | | | [0.00626] |
| Religion | | | | | | | | | | -0.00455 |
| | | | | | | | | | | [0.0682] |
| Experience | | | | | | | | | | 0.00296 |
| | | | | | | | | | | [0.00666] |
| Pre-treatment Revenue | | | | | | | | | | -0.0000673 |
| | | | | | | | | | | [0.000109] |
| Pre-treatment Notification | | | | | | | | | | -0.00731 |
| Observations | 310 | 310 | 310 | 310 | 310 | 310 | 310 | 310 | 310 | 310 |
| R-squared | 0.054 | 0.291 | 0.055 | 0.050 | 0.043 | 0.037 | 0.288 | 0.037 | 0.013 | 0.675 |
| Mean | 0.610 | 32.86 | 0.300 | 55203 | 0.297 | 0.948 | 4.906 | 53.42 | 0.539 | 0.255 |
| F Statistic | 1.080 | 0.0294 | 0.233 | 0.120 | 1.135 | 0.00445 | 0.198 | 0.365 | 0.0632 | 0.407 |
| p-values | 0.299 | 0.864 | 0.629 | 0.729 | 0.288 | 0.947 | 0.656 | 0.546 | 0.802 | 0.931 |

Robust standard errors appear in brackets (clustered at the tax region level). Metrics assigned is a dummy variable that switches on when a causal inference workshop is randomly assigned to participants. The "causal signal" is a binary variable activated when tax officers receive a summary handout of the article "Behavioral Interventions in Tax Compliance: Evidence from Guatemala" by Kettle et al. (2016). Alternatively, the variable takes zero if officers are assigned the correlational study "Effect of Notice of Tax Warning, Notice of Tax Collection, and Tax Education Programs on Tax Compliance in West Sumatera and Jambi" by Budiman and Inayati (2021). The controls include other available individual characteristics obtained from administrative data (i.e. all remaining column dependent variable except the dependent variable used in the respective column). *** p<0.01, ** p<0.05, * p<0.1.