Training Policymakers in Econometrics

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

August 2021

The credibility revolution triggered a paradigm shift in economics. This paper examines its causal effects on deputy ministers in a "mastering metrics" training program. We separated the demand for econometrics training from its impact with a simplified Becker Degroot Marshak mechanism. Policymakers could choose a high or low probability for randomly receiving a popular econometrics or a self-help placebo book. After receiving the book, policymakers participated in an intense training workshop that included watching lecture videos made by the authors of the book, summarizing each chapter, discussing, presenting, and applying the book's concepts in their policymaking. Three results emerge. First, we document large persistent effects. After six months, treated individuals' ratings on the importance of quantitative analysis increase by 50%. Treated individuals' performance in national research methods and public policy exams improves by 0.5-0.8 sigma. Text analysis of their writings reflect an increase in perceived importance of causal inference. Second, treated individuals' willingness-to-pay for commissioning Randomized Control Trials using public funding increases by 300% and decreases by 50% for correlational studies. Third, treated ministers are twice as likely to choose a policy for which there is RCT evidence. We use click behavior as a behavioral proxy of IV defiers. Few defiers are observed, and they are less affected by treatment. Overall, we provide experimental evidence that training policymakers in the school of thought associated with the credibility revolution increases demand and responsiveness to causal evidence. (JEL D72, D78, O17, O18)

Keywords: randomized evaluations, 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). We thank the Government of Pakistan for their cooperation and access to these policymakers. Daniel L. 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 benefited from financial support of the research foundation TSE- Partnership and ANITI funding. We thank Josh Angrist, Peter Hull, Ben Olken, Micheal Kremer, Mathias Sutter and Gautam Rao for their helpful comments and suggestions. Mohammad Ahmed Nasif provided excellent research assistance.

"*RCTs can play an important role in the rigorous evaluation of how policies actually work in practice. Theory is often ambiguous on the effects of policy intervention. Thus, trials can help shed light on the overall effect of policy interventions.*"

Deputy Minister in Pakistan (after our workshops)

1. Introduction

Over the last half century empirical economics has gone through a paradigm shift (Angrist and Pischke 2010). The credibility revolution, with its careful attention to causality, has presented itself as a new paradigm for "taking the con out of econometrics" (Leamer 1983). We study causal effects of a paradigm shift in science (Kuhn 1962) on practitioners–policymakers–using the training of the paradigm as its instrument. Policymakers demand and even respond to causal evidence (Hjort et al. 2021), but they are unlikely to distinguish between different types of evidence and change their policy choices in response to new evidence. There seems to be consensus emerging in the literature that policymakers are highly averse to shifting their beliefs and engage in motivated reasoning to justify their initial policy choices (Baekgaard et al. 2019; Banuri et al. 2019; Vivalt and Coville 2021; Lu and Chen 2021). Sticking to priors and being inattentive to evidence may stymie the implementation of good policies that might otherwise spur economic development (Kremer, Rao, Schilbach 2019). 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? Will it induce them to change their policy choices?

To address these questions, we conducted a randomized trial. We identified the causal effects of the credibility revolution among deputy ministers in Pakistan using an instrument: Mastering 'Metrics: The Path from Cause to Effect, a prominent summary of the credibility revolution (Angrist and Pischke 2014). These deputy ministers are considered by the government of Pakistan the "key wheels on which the entire engine of the state runs" (Government of Pakistan, 2019). We studied the impact of training causal thinking in a policy decision involving deworming, a policy that shares many essential characteristics with other development policies as policymakers aim to make a decision in light of its potential consequences. In this context, we experimentally modified individuals' causal thinking and studied how the school of thought associated with the credibility revolution affects their beliefs and policymaking in a framed field experiment.

Our experiment had three stages. We elicited demand for causal thinking, then randomized the econometrics training, and measured consequences on attitudes, behavior, and decisions. The demand for the Mastering 'Metrics book yielded a proxy for potential compliers to the treatment assignment. We controlled for this proxy as well as assessed whether similar effects were observed for both high and low demanders, that is, estimated the effect of treatment on deputy ministers who might be inframarginal—those less likely to be affected by treatment.

At baseline, we measured policymakers' demand for causal thinking by presenting deputy ministers with a choice between two books. One book was Mastering 'Metrics. The chapters of this book cover the following topics: randomized trials as experimental ideal, regression as mean comparisons, instrumental variables, regression discontinuity design, differences-in-differences, and estimating impact of schooling on wages. 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 isolated effects of causal thinking separate from the demand for causal thinking with a simplified Becker Degroot Marshak mechanism. 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 demand. 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 is intensive training where we aim to maximize the comprehension, retention, and utilization of the educational materials. Namely, we augmented the book receipt with lectures from the books' authors, namely, Joshua Angrist and Daniel Siegel, along with high-stakes 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 high-stakes writing assignments inspired by theory and empirical evidence on the efficacy of social-emotional learning. 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 and discussion of their essays in a workshop within the treatment arm). Deputy ministers in each treatment group also participated in a zoom session to present, discuss the lessons and applications of their assigned book in a structured discussion.

The last stage of our experiment was a suite of measurements of attitude, behavior, and decisions in a framed field experiment. Because we embedded our analyses in administrative data, we had essentially no attrition. Our collaboration with the training academy gave us direct access to administrative baseline measures of ability and background. Importantly, we observed balance on pretreatment quantitative ability as measured from mathematics scores in the entry examinations of the deputy ministers obtained from the Federal Public Service Commission (FPSC). Likewise, we observed balance on demographics, pretreatment writing and interview assessments. Finally, we obtained data on the ministers' regular policy assessments from the training academy. These were conducted 4–6 months following our workshop and scored deputy ministers in national research methods and policy assessments.

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 treatment assignment and performed a textual analysis of the high-stakes writing assignment. We find substantial effects. While attitudes on importance of qualitative evidence are unaffected, treated individuals' beliefs about the importance of quantitative evidence in making policy decisions increases 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 causal training have higher perceived value of causal inference, quantitative data, and randomized control trials. Metrics training increases how policymakers rate the importance of quantitative evidence in policymaking by about 1 full standard deviation. In the writing assignment and demand assessment, treated deputy ministers also showed an increased desire to run a randomized evaluation before rolling out a policy. In the text of their writings, the treated policymakers discussed their understanding of concepts such as "selection bias", "correlation is not causation" and "randomized evaluations allow for apples to apples comparisons". When asked what actions to undertake before rolling out a new policy, they were more likely to choose to run a randomized trial, with an effect size of 0.33 sigma after completing the book and writing assignment (partial training) and 0.44 sigma after attending the lecture, presentation, discussion and workshop (full training). We also observe substantial performance improvements in scores on national research methods and public policy assessments. These regular assessments were not specially requested by the research team, so performance improvements are unlikely to be due to experimenter demand.

Our second main result emerges from a framed field experiment designed to measure deputy ministers' willingness-to-pay for evidence and to measure changing of policy decisions in response to evidence. We measured ministers' 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 (50% more) and from public funds (300% more) for RCTs and less willing to pay for correlation data (50% less). Demand for senior bureaucrats' advice is unaffected. This indicates that econometrics training increased demand for causal evidence by deputy ministers involved in high-stakes policymaking.

In this framed field experiment, we also studied the impact of causal thinking in a policy decision involving deworming. First, we elicited initial beliefs about the efficacy of deworming on long-run labor market outcomes. Then, they were asked to choose between implementing a deworming policy versus a policy to build computer labs in schools. This scenario was particularly realistic as these policy choices were an actual decision facing ministers during the timeframe of our study. Next, we provided a signal—a summary of a recently published randomized evaluation on the long-run impacts of deworming (Kremer et al. 2021). After this signal, 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: the treated ministers 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 to choose deworming, from 40% to 80%. Notably, this shift occurs only for those ministers whose previously believed the impacts of deworming were *lower* than the effects found in the RCT study, while those who previously believed the effects were larger than the estimate reported in the signal did not shift their choice of policy. In contrast, some deputy ministers assigned to the placebo group displayed confirmation bias and motivated reasoning—updating their beliefs in the opposite direction of the signal or updating their beliefs past the signal. We interpret this policy choice 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). Our findings suggest that causal thinking may increase responsiveness to causal evidence, correct for mistakes in the belief-updating process, and potentially raise welfare.

Econometrics training likely induced deputy ministers to choose deworming and rate quantitative evidence differently because they learned causal inference concepts, a fact corroborated by analysis of their writings: metrics trained policymakers demonstrate their knowledge of these concepts by using phrases such as "Observational studies are not apple to apple comparisons" and "Correlation is not causation" in their writings. Their willingness-to-pay for causal evidence increases and willingness-to-pay for correlations decreases. In other words, the treatment did not increase willingness to seek *any* data and evidence, but rather shifted policy makers' beliefs towards the paradigm associated with the credibility revolution.

The study was conducted during the COVID-19 pandemic, which implied deputy ministers were not together on site at the regular training facility. This likely minimized spillovers. Roughly ten percent of deputy ministers attrited, but this is balanced across treatment and control. In the click behavior (classifying as a defier someone who clicked on a video lecture different from the book assigned to them), we saw that less than 10% are defiers. Roughly 5% are never-takers (classified as someone who does not click). As one might expect, defiers and never-takers are less affected by the treatment (with defiers significantly less affected). This suggests that behavioral data can be a potential method for detecting defiers in estimating the Average Treatment Effects in randomized evaluations.

Our results are robust to a series of sensitivity analyses. In addition to the randomly assigned groups being balanced across individual characteristics and in pretreatment quantitative ability as measured by entry mathematics assessment scores, the results are also robust to randomization inference tests. The results are also similar for those with high and low pretreatment mathematics assessment scores. Together, they indicate that small or idiosyncratic samples assigned to treatment or control are unlikely to explain our results.

In addition, we observed variation in the data that is inconsistent with experimenter demand since not everyone in the treatment group responded positively to information—only those individuals whose priors are less than the signal value of 13% impact change their policy choices. Experimenter demand is also not reflected in assessment of teamwork and attitudes on other sources of evidence since our treatment had no effect on any of these outcomes. Finally, we also bound experimental demand by using a methodology proposed by De Quidt et al. (2018). We suggested the subjects choose the ICT policy, so any shift in the direction of the deworming policy would be the opposite of experimenter demand.

The administrative data also included a suite of behavioral data in the field, for example, a choice of field visits to orphanages and volunteering in low-income schools. This allowed us to assess potential crowdout of prosociality, an oft-raised concern about the teaching of neoclassical economics (Selten and Ockenfels 1998; Frank et al. 1993; Rubinstein 2006; Ifcher 2018). We detected no evidence of econometrics training crowding out prosocial behavior—orphanage field visits, volunteering in low-income schools and language associated with compassion, kindness and social cohesion is not significantly impacted. Scores on teamwork assessments as a proxy of soft skills were also unaffected (Deming and Weidmann, 2021).

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 empirical economics: the credibility revolution (Angrist and Pischke 2010). 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 a training 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, our study on econometrics literacy adds to the expansive literature on economics and financial literacy (Lusardi and Mitchell 2014). Recent work attributes up to 40% of inequality in end-of-life wealth to financial literacy through the mediating channel of financial decision-making (Lusardi, Michaud, and Mitchell 2017). Economics training also impacts high-stakes decisions of policymakers and explains up to 30% of the recent shift towards economic conservatism in the American judiciary (Ash, Chen, and Naidu 2021). Our study is closest to a RCT of eight hours of financial literacy training that impacts economic preferences of adolescents (Sutter, Weyland, Untertrifaller, Froitzheim 2020) and a RCT that included two hours of financial literacy training that impacted those who had low levels of financial literacy (Cole, Sampson, and Zia 2011). We, however, study the impact of econometric literacy training — in causal thinking — on attitudes and behavior of adults who make policy decisions.

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 (Hjort et al., 2021) 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' 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' causal thinking and estimated impact on attitudes and subsequent demand for evidence and policy choices.

The rest of the paper is organized as follows. Section II provides the background and details on the experimental set-up. Section III describes the data and empirical specification, while Section IV presents the main results. Section V conducts an heterogeneity analysis. Section VI discusses a series of sensitivity tests. A final section concludes.

II. Background and Study Design

A. *Background*

Study Context.— We conducted a randomized evaluation implemented through close collaboration with an elite training academy. The Academy in Pakistan is one of the most prestigious training facilities that prepares top brass policymakers—deputy ministers—for their jobs. These high-ranking policy officials are selected through a highly competitive exam: about 200 are chosen among 15,000 test-takers annually. The jobs of these deputy ministers entail policymaking, implementation and advisory positions to the President, Prime Minister and cabinet ministers. These deputy ministers attend rigorous training programs at the Academy, taking part in several workshops and assessments that are designed to hone and assess their policy skills. All the workshops are mandatory and count toward their future career trajectories. We obtained unique and unprecedented access to almost all deputy ministers entering service in a single year.[2] We designed a rigorous "mastering metrics" training workshop for these deputy ministers and delivered it as they participated in the Academy's regular common training program. We also obtained access to each ministers' policy assessments from the Academy's administrative data. Our econometrics workshop was 5% of their overall graduation requirements, which elevated the importance of the workshop since their transfers and posting are in part determined by their overall performance at the Academy. We obtained extensive support and cooperation from the director of this elite training academy, before, during and after organization of our workshop. This ensured that we had freedom to design and conduct the workshop with these elite policymakers and embed our analysis within administrative data. The director's letter of support is attached as Table B1 in Appendix B.

B. *Treatment design details*

October 2020: 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 on 20th October 2020. 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

[2] Not only do we anonymize the names of the ministers but the year of their entry to public service. This is done on the request of the Academy that cites political concerns.

ideal, regressions as 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.[3]

November 2020: Assignment of treatment. — On 10th November, the director of Academy sent an official email to complete an assignment associated with the designated book to all deputy ministers. All the deputy ministers in the cohort sent a confirmation message that they will complete the assignment within the deadline. The mandatory nature of the workshop and close collaboration of the director and the staff at the Academy implied we had about 90% take-up of 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.[4] That is, the participants were randomly assigned either Metrics or self-help books but conditional on their choice. 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 2020 (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.

[3] For the table of contents of both books, see Figure B1 of Appendix B.

[4] Specifically, a person choosing the metrics book had 60% probability of being randomly assigned the metrics training, while the person choosing placebo book had 85% probability of being randomly assigned the placebo training. Shipment of the books caused these probabilities to differ.

March 2021: Attitude Survey, lecture, presentation, discussion and workshop.— On 10 March 2020, 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, while 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. 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 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 2021: Initial Beliefs, Post-Signal Beliefs, Willingness-to-Pay and Project Choice.— On 16 May 2021, about 6 months following the book assignments, we elicited policymakers initial beliefs on policy impact and policy choices. Specifically, we elicited their initial 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.[5] To minimize experimental demand effect, we nudged policymakers to choose the computer lab project. That is, before the project choice the policymakers see the following prompt:

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

We elicit the policymakers' 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:

> *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. For more details on the willingness-to-pay transcripts of questions and project choice, see Table A2 in Appendix

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

A. For the whole set-up and summary of the complete experimental design, see Figure B2 in Appendix B.

COVID-19 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 cities due to the COVID-19 pandemic. The training, therefore, took place online. The Academy has strict training protocols that do not allow for random assignment by experimenters on this "elite group" of public officials. However, these procedures were valid only for on-site training, therefore, the unique circumstances arising due to the COVID-19 pandemic provided us an unusual opportunity to randomly assign training at the individual level. The combination of the Academy's express instructions that the participants may not share or discuss our workshop material with their peers, the geographical dispersion of the ministers due to the pandemic at the time of the training, and the non-shareability of the link likely reduced treatment contamination. However, it should be noted that treatment contamination would only mean that our estimates are underestimated.

III. Data and Empirical Specification

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 Pakistan's Federal Public Service Commission (FPSC) that administers the entry examinations for these elite policymakers.[6] Data on WTP, attitudes and beliefs were collected by our research team under the auspices of the Federal Government of Pakistan.

[6] 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.

Empirical Specification. — The impact of metrics training can be evaluated in a simple regression framework. For each individual-level outcome, the estimation equation is:

$$Y_i = \alpha + \beta \text{Metrics Assigned}_i + \boldsymbol{X}_i{}' \mu + \epsilon_i \tag{1}$$

where Y_i is the respective outcome for the policymaker i, this includes attitudes, assessment scores, WTP and policy choices. $\text{Metrics Assigned}_i$ is a dummy equal to one if the policymaker is randomly assigned to metrics training. $\boldsymbol{X_i}$ 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, Metrics Chosen. This is a dummy variable equal to one if the policymaker chooses the metrics book. 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 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. Most important to note is the balance on pretreatment scores on the mathematics assessment. This indicates 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 effect of metrics training on attrition.

IV. Main results

A. *Treatment effect on Attitudes*

Treatment effects on importance of quantitative analysis in policymaking. — In Table 2, 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 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 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 2, we present results of our full training where policymakers attend respective video lectures from the authors of the book, receive a gift voucher, commemorative shields, present the key lessons of the respective book, and partake in a structured discussion of the material in a workshop (full 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 2 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 B3 in Appendix B.

In Table 2, 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 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". 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 2 (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 indicate 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 evidence that the results are likely driven by the fact that policymakers learn new knowledge about causal inference and selection issues. In the last two columns of Table 2, 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 2 report the results on metrics training before the full and partial training. Subjects 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 a likely mechanism that why our treated group have higher perceived importance of quantitative evidence and randomized evaluations.

In Figure 1, we report all results of Table 2 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 requesting randomized evaluation to evaluate effectiveness of public policy, and about 0.30 standard deviation increase 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 corroborated by textual analysis of the ministers' high-stakes assignments. The analysis of their writings reveals that metrics assigned group likely learned causal inference concepts. In Figure 2, 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". Specifically, the group assigned the metrics book is about 40 percentage points more likely to use phrases associated with "causal inference is important", a doubling of phrase usage over the placebo mean and 10 percentage points more likely to use phrases similar to "Observational studies are not apple to apple comparisons", which is a 33% increase over the placebo mean. This suggests that metrics training shifted policymakers' beliefs towards the paradigm associated with the credibility revolution.

B. *Treatment effect on Policy Assessments*

Impact on high-stakes policy assessments. — The close collaboration with the training Academy provided us unique access to the administrative data on policy assessments of the deputy ministers in a high-stakes setting. The Academy administers assessments that in part determine the minister's transfers and postings. This allowed us to assess 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. The second was a research methods workshop. Its content included an introduction to hypothesis testing, multivariate regressions with several applications and case studies. 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 3 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 0.5σ higher in national public policy and 0.8σ higher in the research methods assessments. This strongly 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 3, 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).

C. *Does Metrics Training Crowd out Prosocial Behavior?*

Prosocial Behavior in the Field. — 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). We 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. Figure 2 presents these results. From the top of Figure 2, 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.

Prosocial Language. — These field results are corroborated by analyzing language use in the policymakers' writing assignments. We found that the metrics training program does not come at the expense of reduction in the use of prosocial language. 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.[7] 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 2, 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.

D. *Treatment Effect on WTP for Evidence*

Effect of Metrics Training on WTP. — In May 2021, 6 months following the metrics training workshop, all policymakers' initial beliefs on the effect of deworming were elicited. This is followed by a "signal" on causal evidence regarding the effect of deworming on various outcomes including income. We also elicited WTP from both private and public funds for three pieces of information for: (1) results from a RCT on 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. The WTP was elicited both before and after the signal. In particular, 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).*

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 3 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 3, the group

[7] 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)

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.

Metrics policymakers not only shifted their beliefs towards the information presented from a randomized evaluation, they were also more likely to pay for results from a randomized evaluation. In Panel A of Table 4 (Columns 1 and 4), we find that metrics trained policymakers are willing-to-pay about PKR 1000–2000 (USD 6.40–12.80) 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 demand 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 Panel B of Table 4 (Columns 1 and 4), we elicit WTP for the same information from public funds.[8] The effects from public funds are substantially larger and statistically significant with metrics assigned policymakers willing to pay about Rs 740000-1400000 (USD 5,000-8500) for the information from randomized evaluation. The evidence, therefore, strongly suggests that even when there are competing sources of information, such as obtaining correlational data or advice from senior public officials, treated policymakers are likely to spend substantial amounts to obtain evidence from randomized evaluation from public funds. These results can also be observed from Figure 4 where we plot the distributions of WTP from public and private funds: metrics assigned policymakers are significantly more likely to pay large amounts from personal and public funds for randomized evaluations than those assigned the placebo training.

Interestingly, however, we observe metrics trained policymakers to *decrease* their WTP for correlations. 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 personal funds for correlational data comparing incomes of pupils in schools with and without deworming relative to those assigned the placebo training, and about PKR 35000–55000 (USD 200–350) less for this information from public funds (Table 4,

[8] We measure both public and private WTP as Rao et al. (2021, p. 4) notes 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." (Rao et al., 2021, p. 24).

Panels A and B, Columns 2 and 5). This is equivalent to a 50% decrease in WTP for correlational data from private funds and a 33% decrease from public funds. Figure 5 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' WTP for advice of senior bureaucrats, another competing source of information available to the ministers. These results are reported in Table 4, Panels A and B, Columns 3 and 6. 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 6, 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. This is a reassuring falsification test since metrics training did not emphasize the importance of feedback or advice from senior bureaucrats. This also suggests that the metrics training does not crowd out potentially useful other sources of information such as senior bureaucrats' advice.

V. Heterogeneity by Initial Beliefs, Behavioral Proxies for 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 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 the signal value of 13%, while no such shifting 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, 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 to choose to implement the deworming policy when causal evidence is provided to them (Figure 7 of Panel A). The effect is observed 6 months following the metrics training, suggesting a persistent effect of our intervention. We find not 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 nonparametric estimation at different percentiles of prior beliefs. Panel B of Figure 7 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 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 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 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).[9] 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 5 (Columns 1 and 3) we report the heterogeneous impact of metrics training

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

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 5, Column 3). We found that never-takers are less likely to be impacted as well, but to a lesser extent (Table 5, 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 external validity within the study design. A typical concern in RCTs is that the compliers respond to treatment and we estimate Local Average Treatment Effect (LATE) since we do not 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 B4 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 pre-treatment quantitative respond more to causal evidence. Table B5 presents these results. We find no evidence of 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 or experimental demand effects. 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 decision-makers who can potentially impact long-run economic and political development. Nevertheless, 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, our power calculation with statistical power of 80% and significance level of 5% reveals that even with a sample of about 200 policymakers, the individual level randomization allows us to detect a minimum detectable effect equivalent to a change of 0.23 standard deviations. Our documented effect sizes are—at least—twice as large as this, providing us sufficient power to detect the effects with our sample. Nevertheless, Imbens and Rubin (2015) recommend, in relatively small sample randomized evaluations, to conduct 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 B6, B7 and B8 of Appendix B with 1000 iterations of this process.[10] 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.

[10] *ritest* in Stata is implemented to compute p-values corresponding to the permutation inference test by Heb (2017).

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, we bound experimental demand by using a methodology similar to De Quidt et al. (2018). 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.[11] Second, we do not find a change in perceived importance of *qualitative* evidence of our metrics training i.e. policymakers did not respond to any evidence. Third, in addition to increased WTP for commissioning RCTs, we find no evidence of metrics training on WTP for advice of bureaucrats, while there is an opposite effect on WTP for procuring correlational evidence. All of these patterns are inconsistent with experimental demand explaining our results.

External Validity. — In this subsection, we follow List (2020)'s *Selection-Attrition-Naturalness-Scaling* (SANS) conditions in our discussion of 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 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.

Correcting for Fallacy in Judgments. — We found that econometrics training not only increases demand for and receptiveness to causal information, it also appears to correct for fallacy in judgments that appear in the placebo group. To investigate if the econometrics treatment addresses fallacy in judgments of policymakers, we compared the distribution in belief updating following the signal with policymakers' initial beliefs. Figure B4 reports this

[11] Before the policy choice, all deputy ministers were told that ICT policy is important for the 21st century economy.

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. Two features of the graph are intuitive and show the results of Figure 7 in a less parametric manner. First, 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. Second, in contrast, the placebo individuals almost never update their beliefs. However, the data suggests some fallacy in judgments among the placebo individuals. The key feature of this figure is that those who do react to the signal do not necessarily react in a Bayesian manner. In fact, placebo individuals whose beliefs are below the signal occasionally update their beliefs in a direction opposite of the signal. Consistent with this anomaly is the moderate number of placebo individuals whose initial beliefs are near the signal, but 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.[12]

VII. Conclusion

In this paper, we present results from a field experiment with deputy ministers in Pakistan to explore the effect of econometrics training on the demand and the 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.

We found that 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

[12] Finally, our main results are also robust to alternate specifications, for instance, a saturated model that also includes an interaction between metrics assigned and metrics chosen (see Table B9 in Appendix B).

methods and public policy exams improve by 0.5–0.8 sigma. Text analyses of their writings suggest a shift in subjective value for causal inference. We also found that treated individuals' WTP for commissioning RCTs using public funding increased by 300% 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 people 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 among deputy ministers but may even correct for mistakes in the belief updating process.

The point estimates are sizable. We show that treated individuals are 40% more likely to choose policies when presented with RCT evidence, but only when their prior beliefs are less than the estimates from the RCT. Placebo individuals show either lack of responsiveness to or misinterpretation of causal evidence. Assessments on national policy exams of trained ministers improved by up to 1 standard deviation.

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. Future research can follow the long-term effects of metrics training on these deputy ministers as well as study the welfare consequences of training in the school of thought associated with the credibility revolution.

References

Ahuja, A., Baird, S., Hicks, J.H., Kremer, M. and Miguel, E., 2017. Economics of Mass Deworming Programs. Child and Adolescent Health and Development. 3rd edition.

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

Angrist, J.D. and Pischke, J.S., 2014. Mastering 'metrics: The path from cause to effect. Princeton University Press.

Ash, E., Chen, D.L. and Naidu, S., 2019. Ideas have consequences: The impact of law and economics on american justice. Center for Law & Economics Working Paper Series, 4.

Azoulay, P., Fons-Rosen, C. and Graff Zivin, J.S., 2019. Does science advance one funeral at a time?. American Economic Review, 109(8), pp.2889-2920.

Baekgaard Martin, Christensen Julian, Dahlmann Casper M., Mathiasen Asbjørn and Petersen Grund Niels B. 2017. “The role of evidence in politics: Motivated reasoning and persuasion among politicians.” British Journal of Political Science 08, 1–24.

Banuri, S., Dercon, S. and Gauri, V., 2019. Biased policy professionals. The World Bank Economic Review, 33(2), pp.310-327.

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

Berry, J., Fischer, G. and Guiteras, R., 2020. Eliciting and utilizing willingness to pay: Evidence from field trials in Northern Ghana. Journal of Political Economy, 128(4), pp.1436-1473.

Cantoni, D., Yang, D.Y., Yuchtman, N. and Zhang, Y.J., 2019. Protests as strategic games: experimental evidence from Hong Kong's antiauthoritarian movement. The Quarterly Journal of Economics, 134(2), pp.1021-1077.

Chen, D.L., Moskowitz, T.J. and Shue, K., 2016. Decision making under the gambler’s fallacy: Evidence from asylum judges, loan officers, and baseball umpires. The Quarterly Journal of Economics, 131(3), pp.1181-1242.

Chen D.L., Schonger M., Wickens C., 2016. oTree - An open-source platform for laboratory, online, and field experiments. Journal of Behavioral and Experimental Finance, 9, pp. 88-97.

Cole, S., Sampson, T. and Zia, B., 2011. Prices or knowledge? What drives demand for financial services in emerging markets?. The Journal of Finance, 66(6), pp.1933-1967.

Conte, E., Grazzani, I. and Pepe, A., 2018. Social cognition, language, and prosocial behaviors: a multitrait mixed-methods study in early childhood. Early Education and Development, 29(6), pp.814-830.

Cristia, J., Ibarrarán, P., Cueto, S., Santiago, A. and Severín, E., 2017. Technology and child development: Evidence from the one laptop per child program. American Economic Journal: Applied Economics, 9(3), pp.295-320.

Croke, K., Hicks, J.H., Hsu, E., Kremer, M. and Miguel, E., 2016. Does mass deworming affect child nutrition? Meta-analysis, cost-effectiveness, and statistical power (No. w22382). National Bureau of Economic Research.

Deming, D. and Weidmann, B., 2021. Team Players: How Social Skills Improve Team Performance. Forthcoming Econometrica.

De Quidt, J., Haushofer, J. and Roth, C., 2018. Measuring and bounding experimenter demand. American Economic Review, 108(11), pp.3266-3302.

Durkin, K. and Conti-Ramsden, G., 2007. Language, social behavior, and the quality of friendships in adolescents with and without a history of specific language impairment. Child development, 78(5), pp.1441-1457.

Finkbeiner, M., Nicol, J., Greth, D. and Nakamura, K., 2002. The role of language in memory for actions. Journal of Psycholinguistic Research, 31(5), pp.447-457.

Frank, R.H., Gilovich, T. and Regan, D.T., 1993. Does studying economics inhibit cooperation?. Journal of economic perspectives, 7(2), pp.159-171.

Frey, B.S. and Meier, S., 2003. Are political economists selfish and indoctrinated? Evidence from a natural experiment. Economic Inquiry, 41(3), pp.448-462.

Foucault, M., 1970. The order of things. Routledge.

Gabaix, X., 2014. A sparsity-based model of bounded rationality. The Quarterly Journal of Economics, 129(4), pp.1661-1710.

Hjort, J., Moreira, D., Rao, G. and Santini, J.F., 2021. How research affects policy: Experimental evidence from 2,150 brazilian municipalities. American Economic Review, 111(5), pp.1442-80.

Heb, S., 2017. Randomization inference with Stata: A guide and software. The Stata Journal, 17(3), pp.630-651.

Ifcher, J. and Zarghamee, H., 2018. The rapid evolution of homo economicus: Brief exposure to neoclassical assumptions increases self-interested behavior. Journal of Behavioral and Experimental Economics, 75, pp.55-65.

Kremer, M., Brannen, C. and Glennerster, R., 2013. The challenge of education and learning in the developing world. Science, 340(6130), pp.297-300.

Kremer, M., Hamory, J., Miguel, E., Walker, M., and Baird, S., 2021. Twenty-year economic impacts of deworming. Proceedings of the National Academy of Sciences, 118(14).

Kremer, M., Rao, G. and Schilbach, F., 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, R.J., 1986. Evaluating the econometric evaluations of training programs with experimental data. The American economic review, pp.604-620.

Leamer, E.E., 1983. Let's take the con out of econometrics. The American Economic Review, 73(1), pp.31-43.

Lu W. and Chen D.L., 2021. Motivated Reasoning in the Field: Polarization of Precedent, Prose, and Policy in U.S. Circuit Courts, 1930-2013. Mimeo.

Lusardi, A. and Mitchell, O.S., 2014. The economic importance of financial literacy: Theory and evidence. Journal of economic literature, 52(1), pp.5-44.

Lusardi, A., Michaud, P.C. and Mitchell, O.S., 2017. Optimal financial knowledge and wealth inequality. Journal of Political Economy, 125(2), pp.431-477.

Malmendier, U., Nagel, S. and Yan, Z., 2017. The making of hawks and doves: Inflation experiences on the FOMC (No. w23228). National Bureau of Economic Research.

Marwell, G. and Ames, R.E., 1981. Economists free ride, does anyone else?: Experiments on the provision of public goods, IV. Journal of public economics, 15(3), pp.295-310.

Mbiti, I.M., 2016. The need for accountability in education in developing countries. Journal of Economic Perspectives, 30(3), pp.109-32.

Miguel, E. and Kremer, M., 2004. Worms: identifying impacts on education and health in the presence of treatment externalities. Econometrica, 72(1), pp.159-217.

Mehmood, S., Naseer, S. and Chen, D.(2021) Training Effective Altruism. Mimeo

Merton, R.K., 1973. The sociology of science: Theoretical and empirical investigations. University of Chicago press.

Rubinstein, A., 2006. Dilemmas of an economic theorist. Revista de Economía Institucional, 8(14), pp.191-213.

Shapin, S., 1982. History of science and its sociological reconstructions. History of science, 20(3), pp.157-211.

Selten, R. and Ockenfels, A., 1998. An experimental solidarity game. Journal of economic behavior & organization, 34(4), pp.517-539.

Sidorov, G., Gelbukh, A., Gómez-Adorno, H. and Pinto, D., 2014. Soft similarity and soft cosine measure: Similarity of features in vector space model. Computación y Sistemas, 18(3), pp.491-504.

Siegel, D.J., 2010. Mindsight: The new science of personal transformation. Bantam.

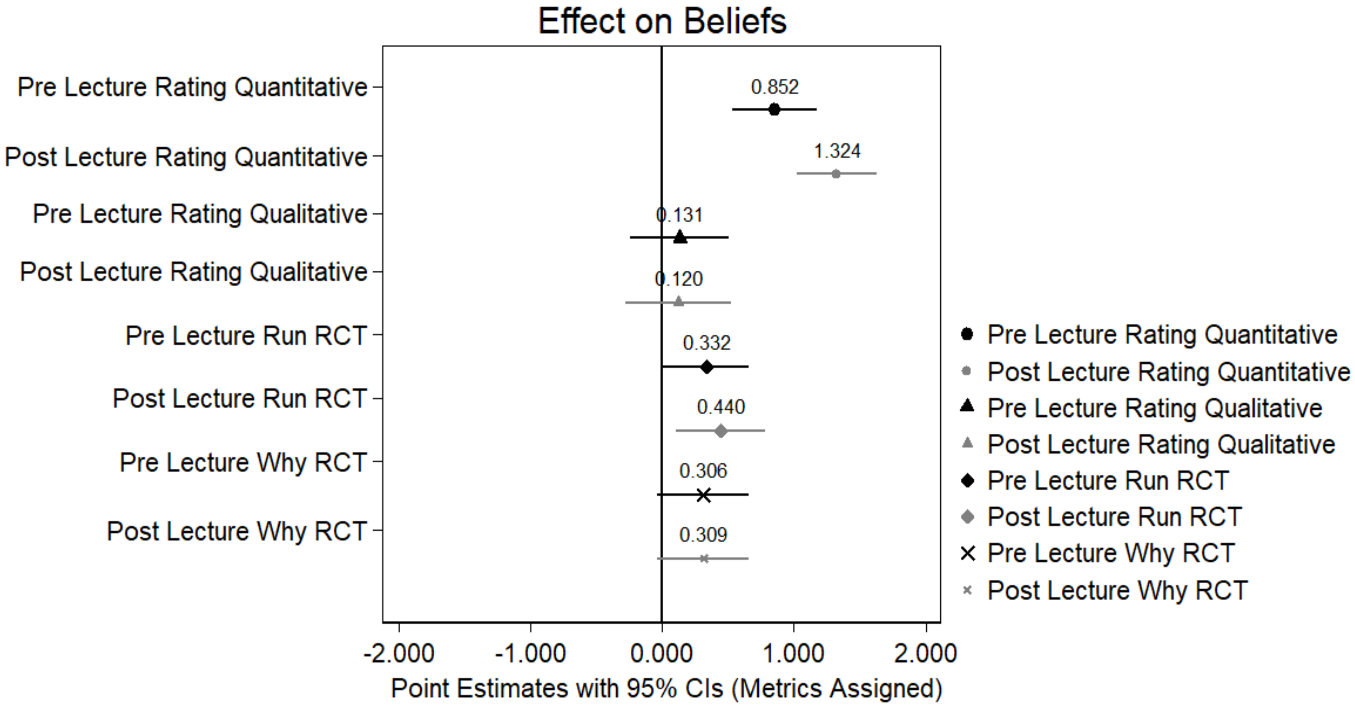
Sutter, M., Weyland, M., Untertrifaller, A. and Froitzheim, M., 2020. Financial literacy, risk and time preferences–Results from a randomized educational intervention. MPI Collective Goods Discussion Paper, (2020/17).

Thomas, K., 1962. The structure of scientific revolutions. International Encyclopedia of Unified Science, 2(2).

Vivalt, E. and Coville, A. 2021. How Do Policy-Makers Update Their Beliefs? Mimeo.

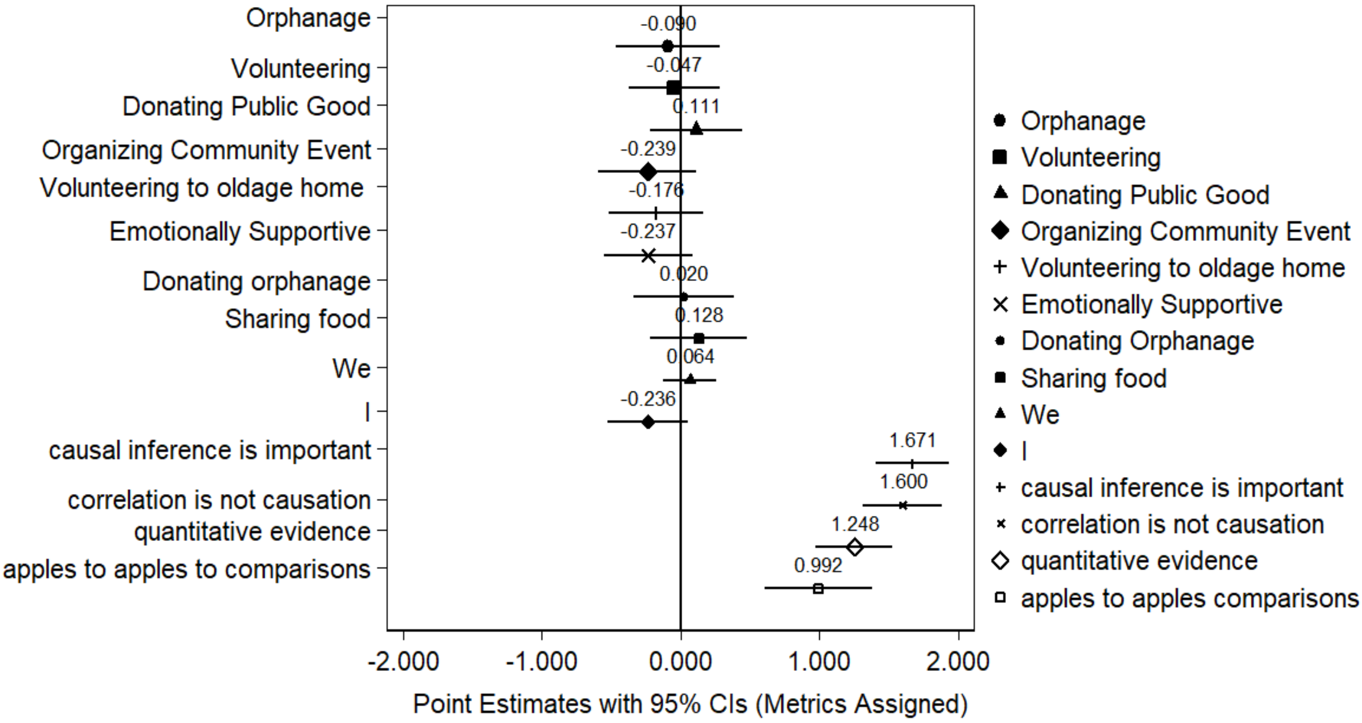
Figures and Tables

Figure 1: Impact of Metrics Training on Beliefs – Standardized



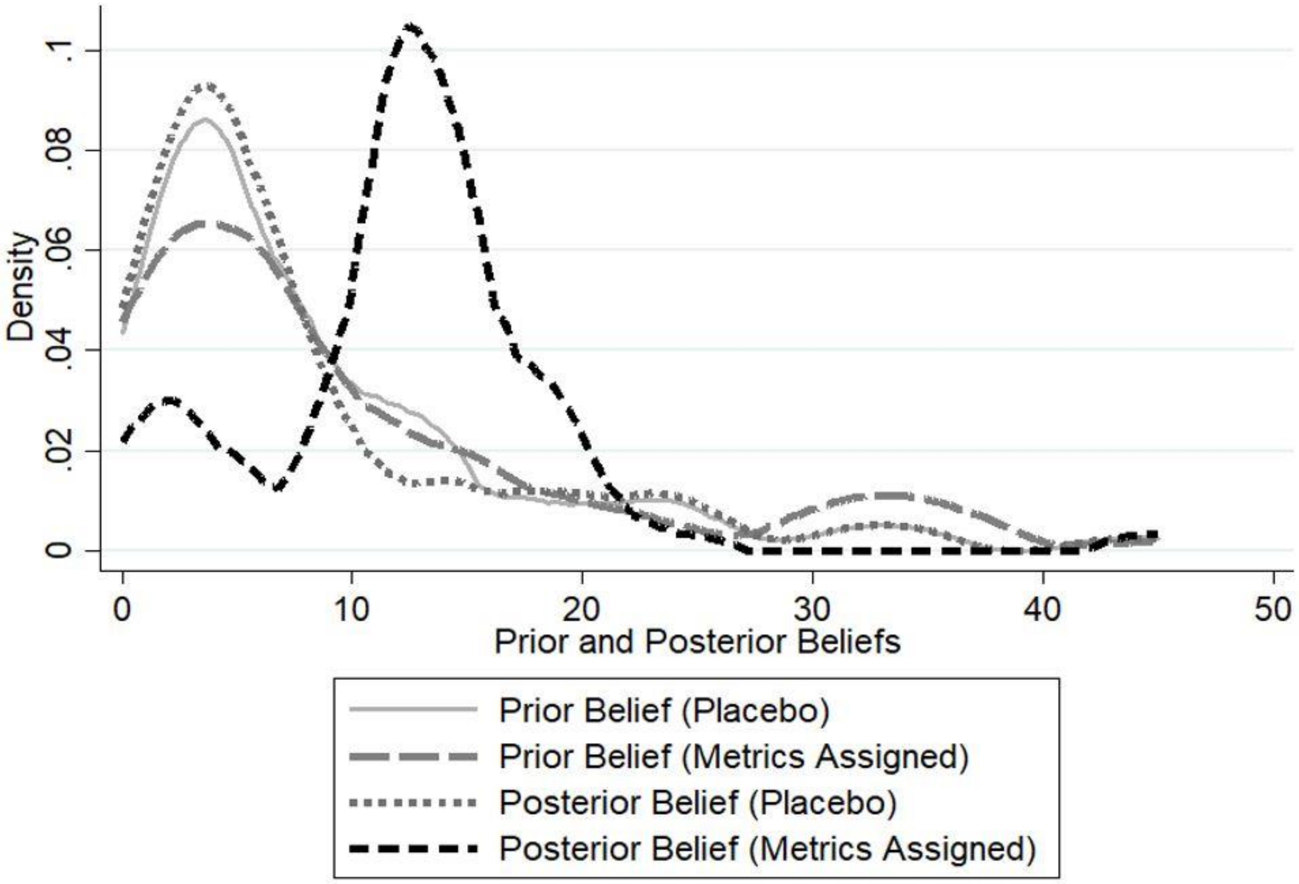
Note: The figure above estimates our main specification including choice of book and individual level controls with all dependent variables standardized to mean zero and standard deviation one. Point estimate and 95% confidence interval on the randomly assigned metrics training is presented for each 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: 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.

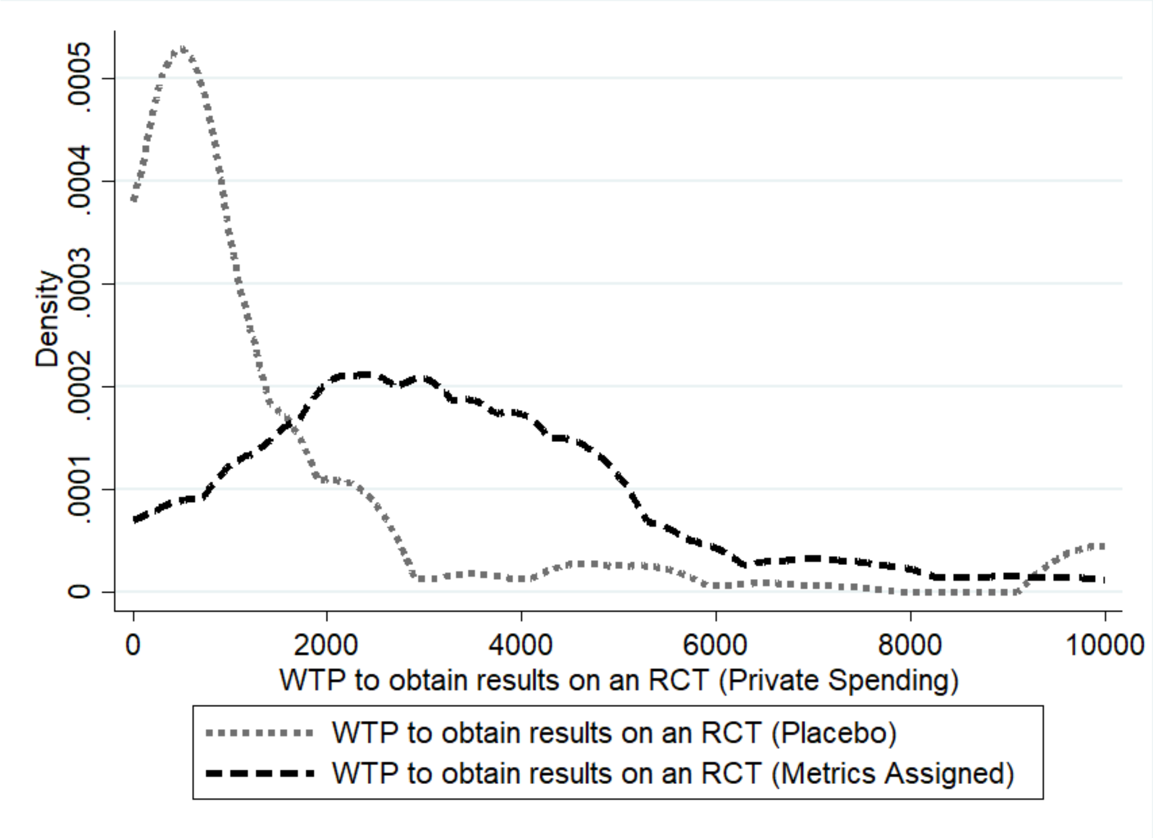
Figure 3: **Distribution of Initial Beliefs and Post-Signal Beliefs**



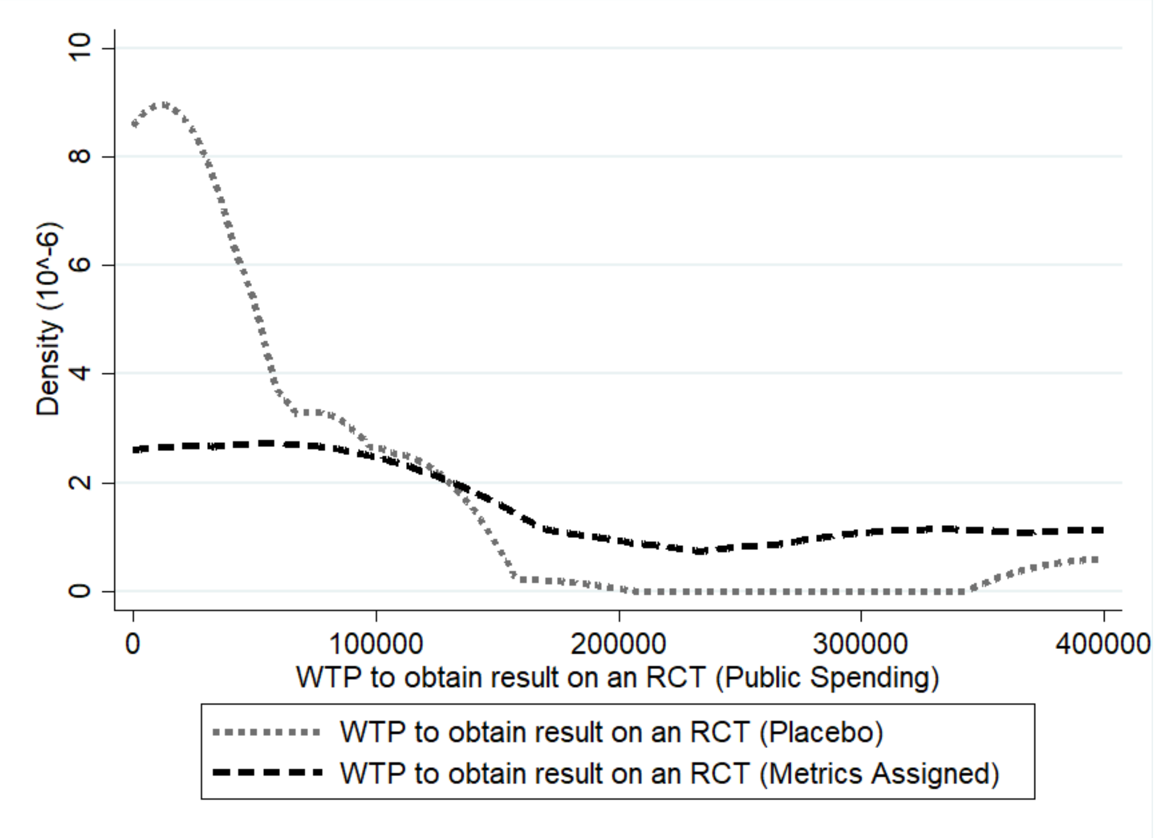
Note: The figure plots distribution of prior and posterior beliefs on the effect of deworming on income after information on the estimate from a randomized evaluation on the effect of deworming is revealed to all participants. The estimate for impact of deworming is taken from a 25 year long randomized 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 4: **Distribution of Post-Signal WTP for RCT**

Panel A: Private Spending



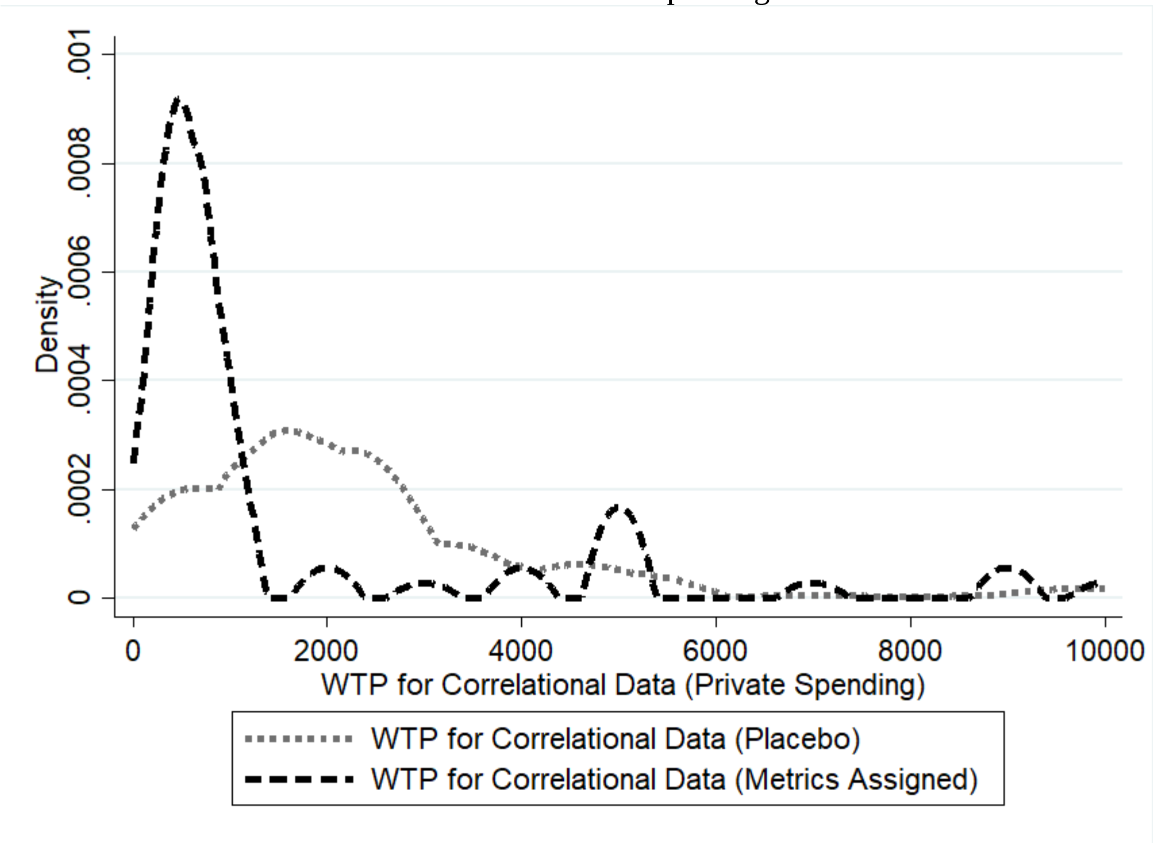
Panel B: Public Spending



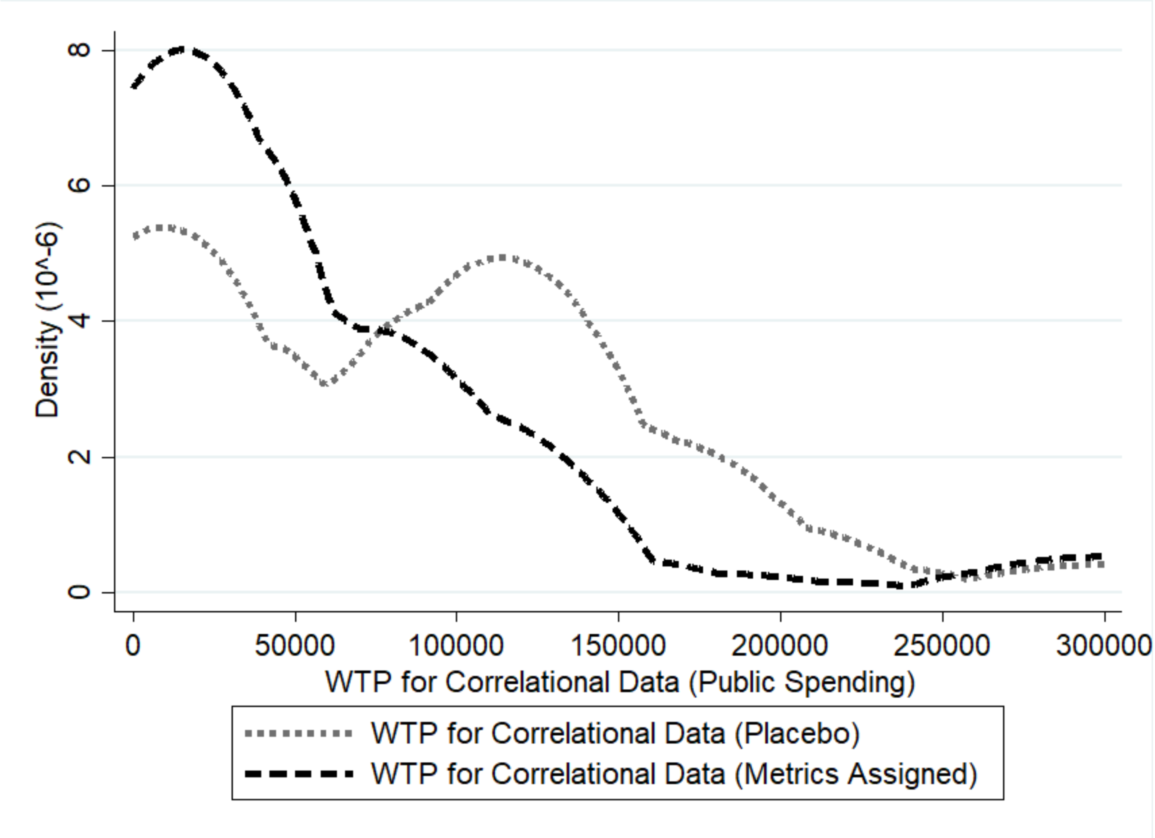
Note: The figure plots distribution of WTP in Pakistani Rupees for policymaker to obtain estimate from an RCT for a policy decision (choosing deworming or computer lab project). Panel A is for private spending whereas Panel B is the willingness to pay from the public exchequer.

Figure 5: **Distribution of Post-Signal WTP for Correlational Data**

Panel A: Private Spending



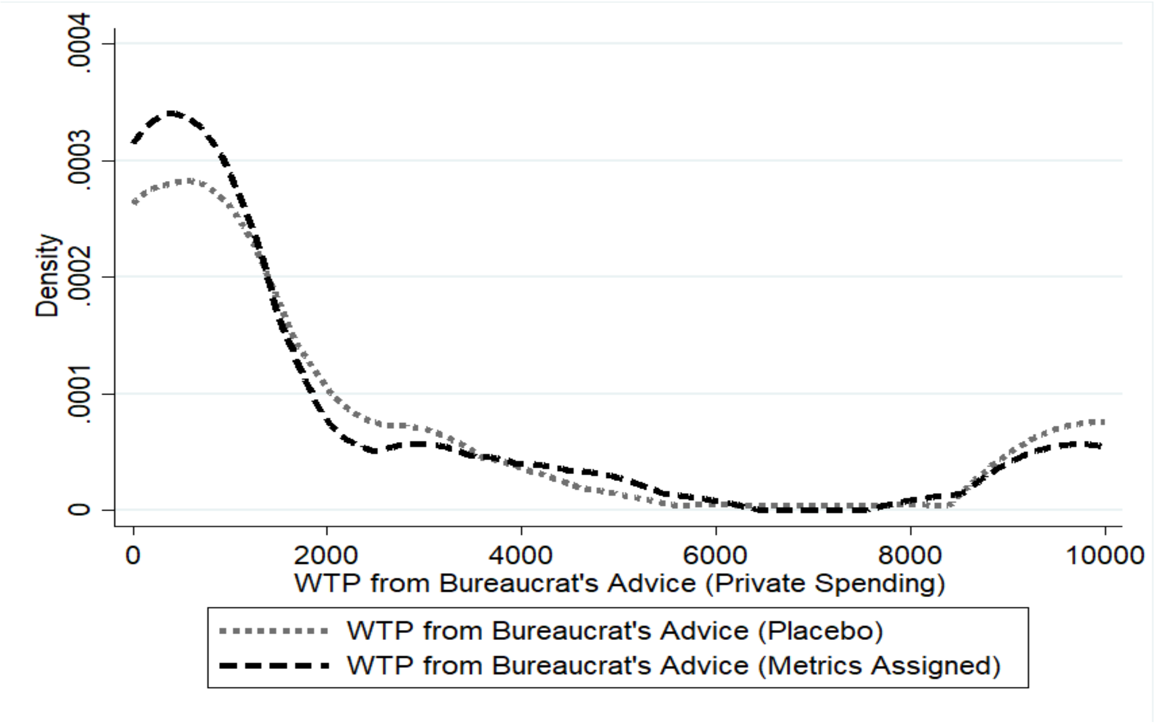
Panel B: Public Spending



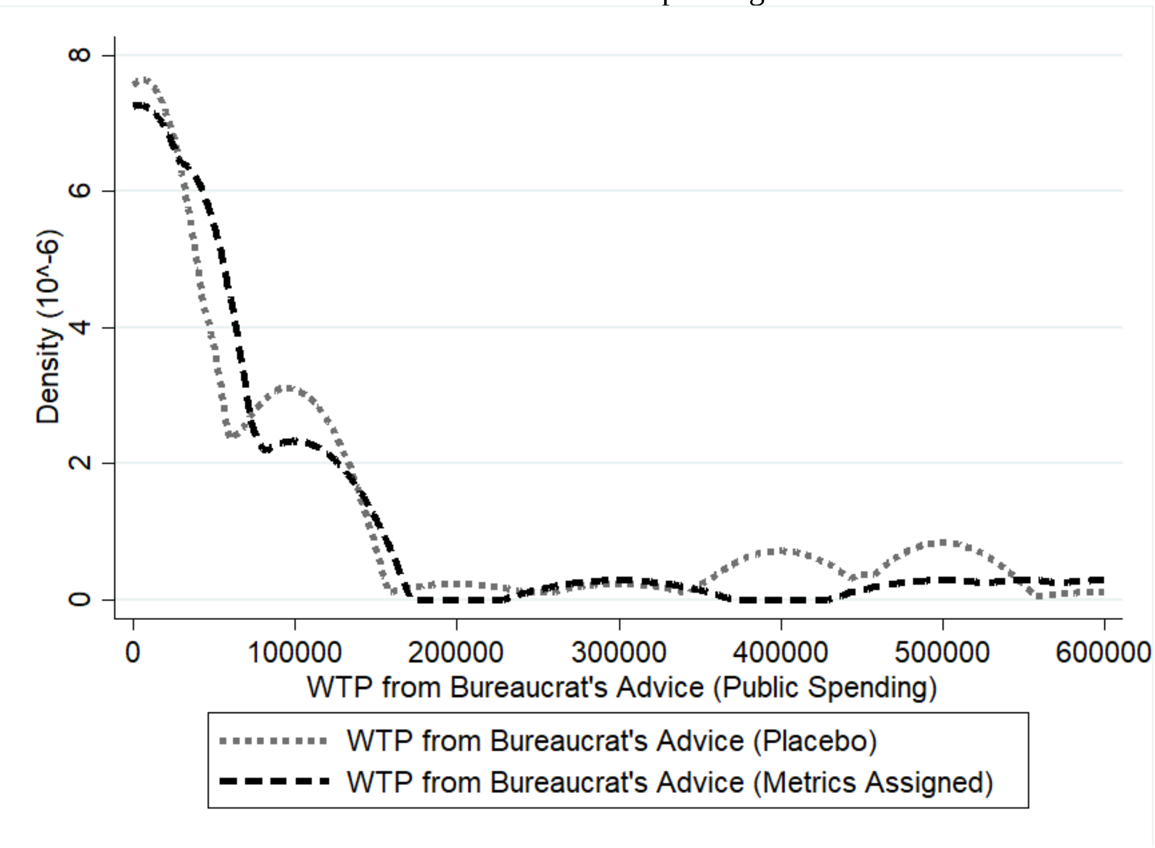
Note: The figure plots distribution of WTP in Pakistani Rupees for policymaker to obtain relevant correlation data for a policy decision (choosing deworming or computer lab project). Panel A is for private spending whereas Panel B is the willingness to pay from the public exchequer.

Figure 6: **Distribution of Post-Signal WTP for Bureaucrat's Advice**

Panel A: Private Spending



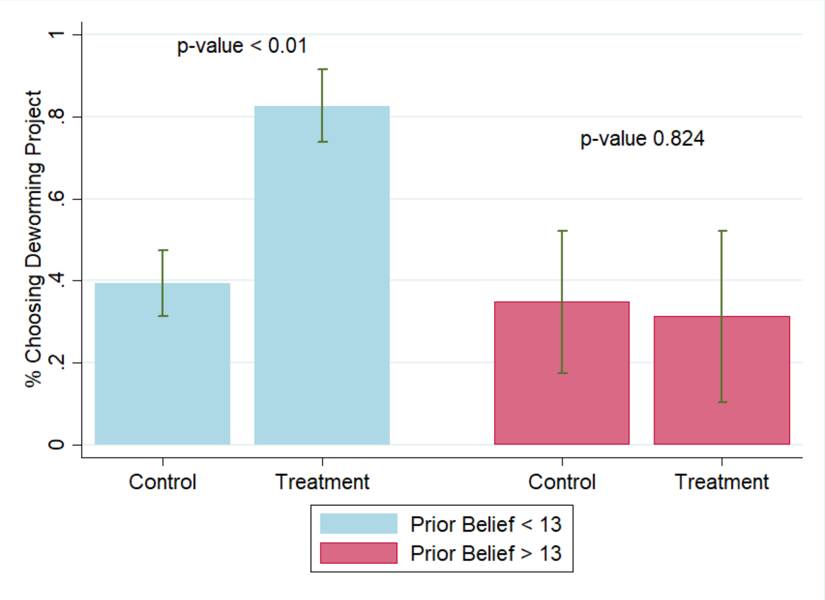
Panel B: Public Spending



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.

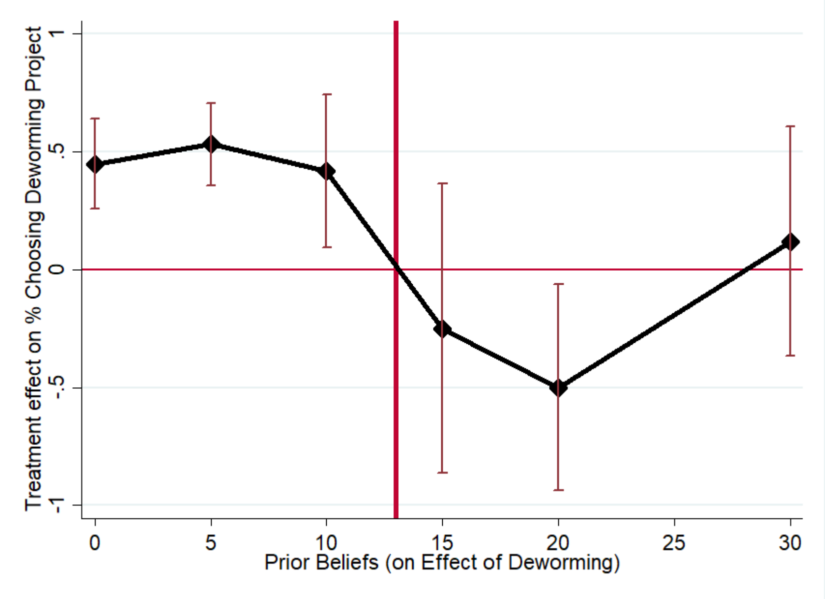
Figure 7: Effect of Metrics Training on Deworming Project Choice by Initial Beliefs

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.

Table 1: Balance of Treatment on Individual Characteristics

	(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
R-squared	0.047	0.165	0.253	0.225	0.070	0.630	0.453	0.645	0.484	0.227	0.017
Mean of dependent 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 brackets (clustered at the individual level). Metrics assigned is a dummy variable that switches on when 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 available at the Public Service Administration of Pakistan (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: The Effect of Metrics Training on Beliefs – Original Units

	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	0.167**	0 .220**	0.151*	0.153*
	(0.176)	(0.178)	(0.196)	(0.206)	(0.082)	(0.085)	(0.087)	(0.087)
Individual Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	190	190	190	190	190	190	190	190
Mean of dep. var. (placebo)	2.745	2.979	2.490	2.596	0.362	0.404	0.396	0.396

Robust standard errors clustered at individual level appear in 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? "While in Columns 5 and 6 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?" 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 one 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 zero otherwise. Every time the beliefs or ratings are measured before and after the lecture. Metrics assigned is a dummy variable that switches 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 3: 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 brackets (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. This assessment scores decisions pertaining to teamwork and group work these policymakers typically make in the field. Finally, in Columns (5) and (6) scores *Teams & Group Decisions* workshop. The workshop content included an introduction to hypothesis testing, multivariate regression, randomized controlled trials with particular focus on application to policy. 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 4: Effect of Metrics Training on Willingness to Pay

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	1486.622**	-963.938	46.442	2063.028***	-1020.318***	-1986.220
	(701.070)	(618.046)	(166.907)	(587.535)	(340.419)	(1389.955)
		p-value = 0.11				
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***	10256.680	1391308.300**	-35273.920***	-2047.595
	(228330.400)	(13832.040)	(49316.580)	(665160.300)	(13033.010)	(34090.020)
Individual Controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	180	180	180	180	180	180
R-squared	0.235	0.170	0.069	0.094	0.148	0.081
Mean of dep. var. (placebo)	224604.700	109746.900	168464.800	928546.100	94136.720	115937.500

Robust standard errors clustered at individual level appear in 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 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 5: Effect of Treatment on WTP by Defiers and Never Takers

	Private Spending		***Public Spending***	
	After Signal Amount Randomized Trial	*After Signal Amount Randomized Trial*	*After Signal Amount Randomized Trial*	*After Signal 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 brackets. 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 choose 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 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.

Online Appendix to:

Training Policymakers in Econometrics

By **Sultan Mehmood, Shaheen Naseer and Daniel Chen**

Contents

A. Additional Experimental Detail

Table A1: Transcript of Email sent by the Director

Dear Officers,

With this email, we wanted to send you a mandatory assignment from the Academy on the workshop delivered by Dr Shaheen Naseer. 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 2020 to our office. Please also cc the email to Dr. Shaheen Naseer. You should write your answer in a word document, convert it in PDF, put your 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 e-Books 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: 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]

Almost always same for metrics and placebo group as prior as no effect of metrics treatment…posterior = prior

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 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 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]

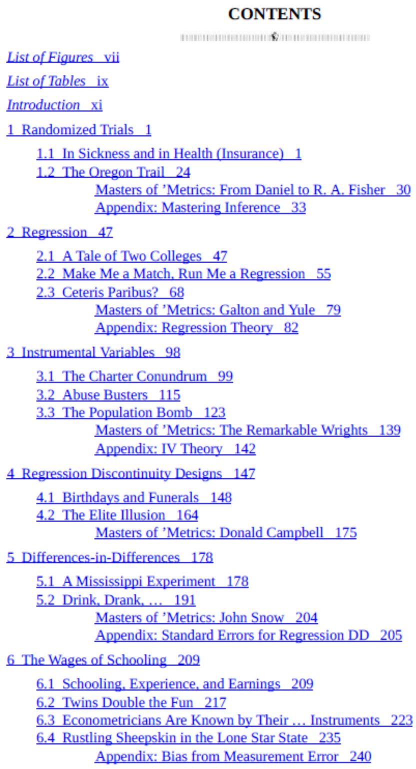
Question 8) Now, what project would you choose 1 or 2? One being deworming and two being computer lab.

[Enter Number]

B. Additional Figures and Tables

Figure B1: Table of Contents of Treatment and Placebo Books

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



CONTENTS

Panel B: “Mindsight” by Daniel J. Siegel.

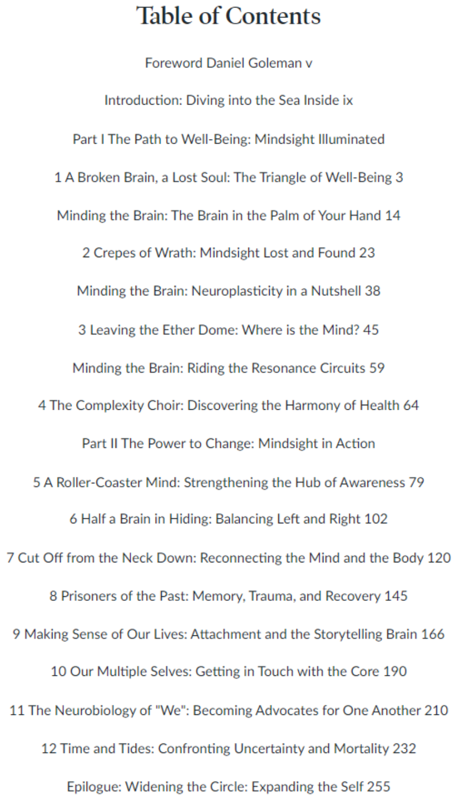
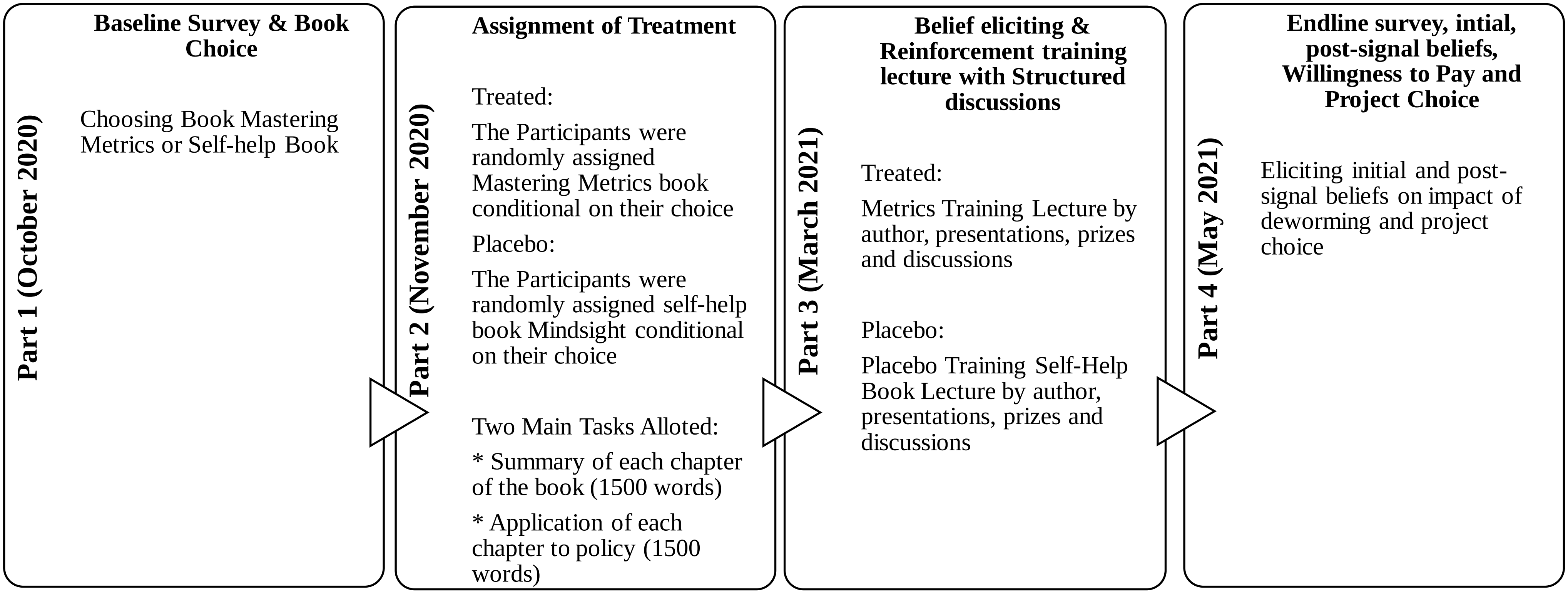


Table of Contents

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



Part 1 (October 2020)

Baseline Survey & Book Choice

Choosing Book Mastering Metrics or Self-help Book

Part 2 (November 2020)

Assignment of Treatment

Treated:

The Participants were randomly assigned Mastering Metrics book conditional on their choice

Placebo:

The Participants were randomly assigned self-help book Mindsight conditional on their choice

Two Main Tasks Alloted:

* Summary of each chapter of the book (1500 words)

* Application of each chapter to policy (1500 words)

Part 3 (March 2021)

Belief eliciting & Reinforcement training lecture with Structured discussions

Treated:

Metrics Training Lecture by author, presentations, prizes and discussions

Placebo:

Placebo Training Self-Help Book Lecture by author, presentations, prizes and discussions

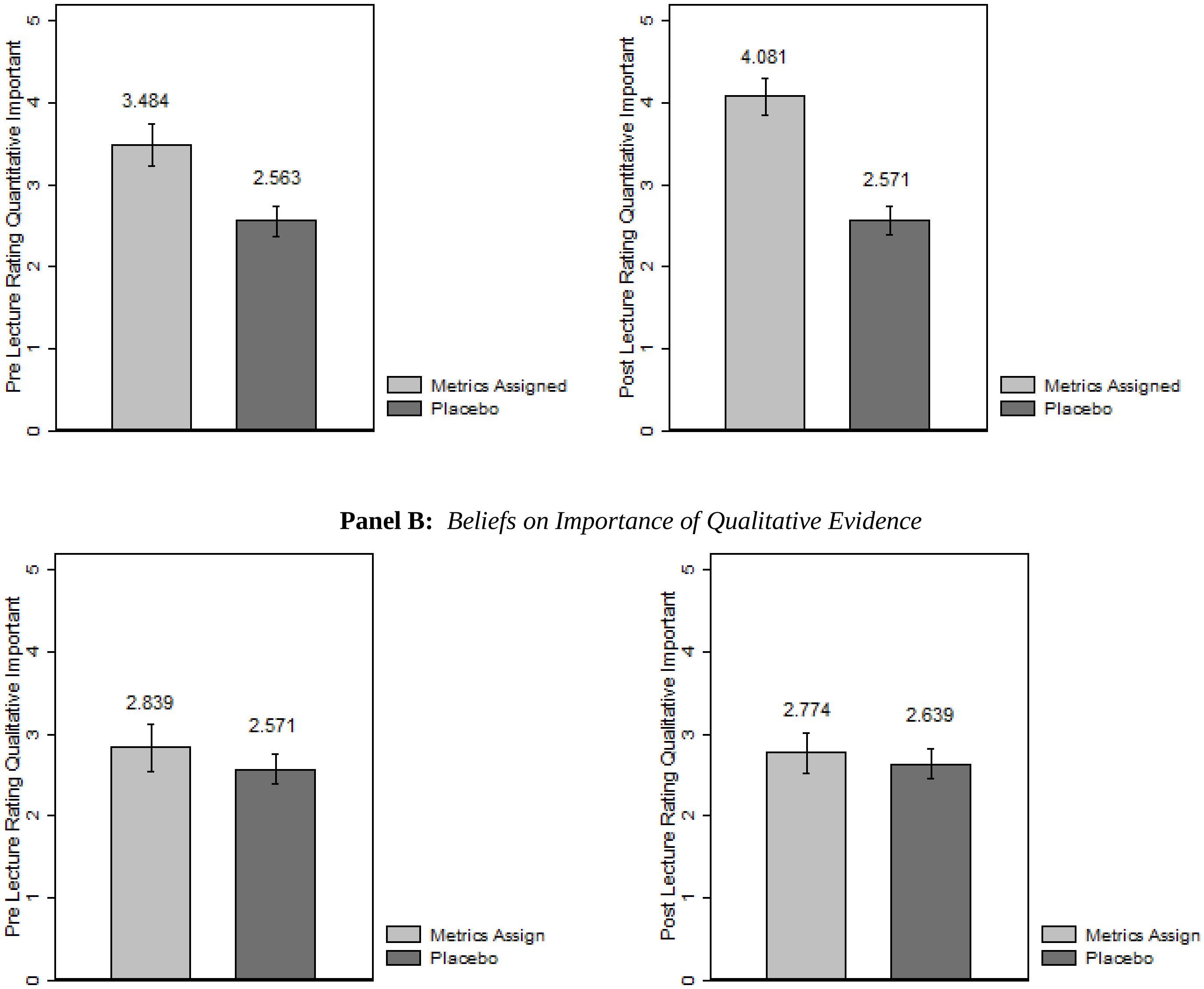
Part 4 (May 2021)

Endline survey, intial, post-signal beliefs, Willingness to Pay and Project Choice

Eliciting initial and post-signal beliefs on impact of deworming and project choice

Figure B3: Impact of Metrics Training

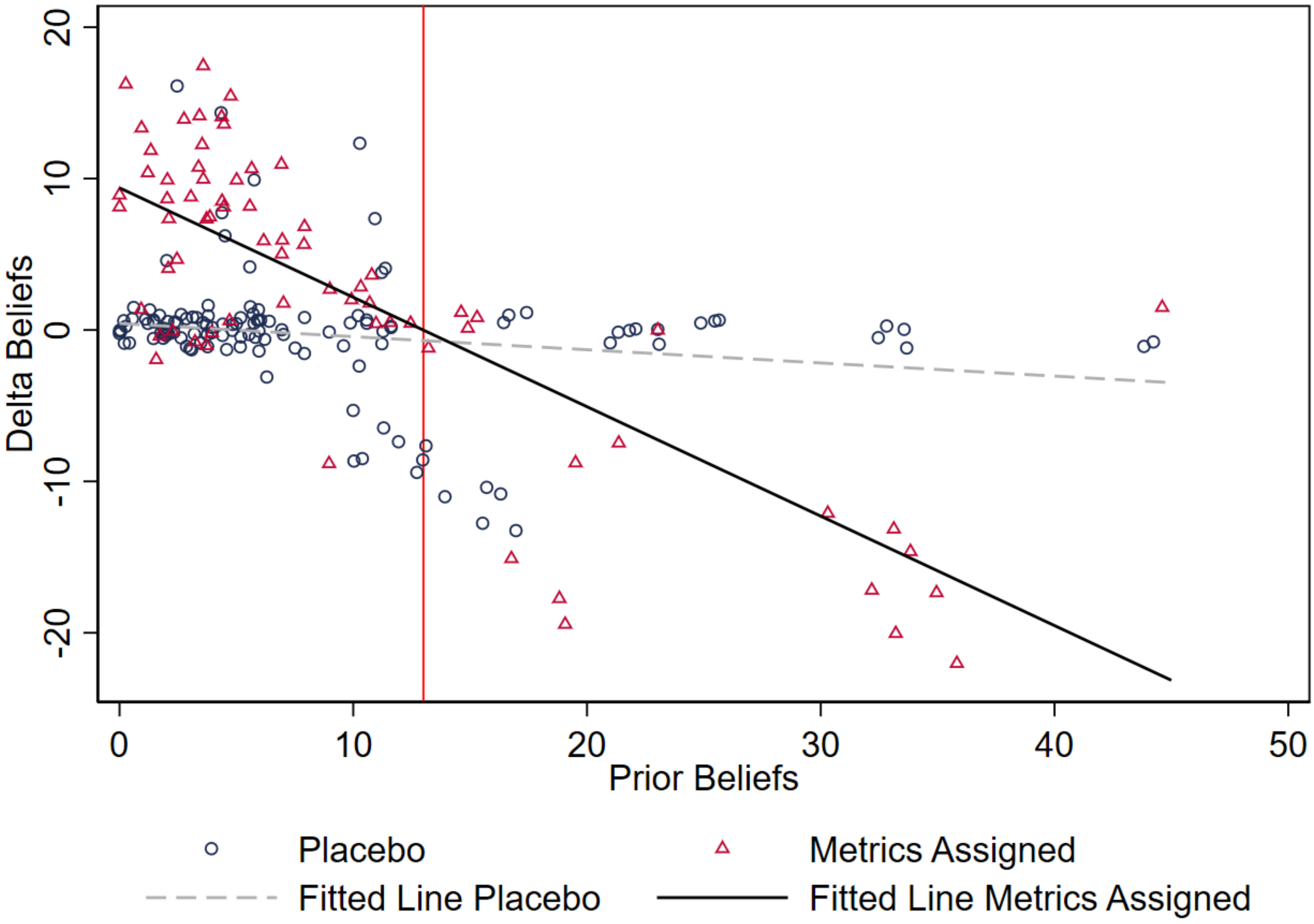
Panel A: *Beliefs on Importance of Quantitative Evidence*



Panel B: *Beliefs on Importance of Qualitative Evidence*

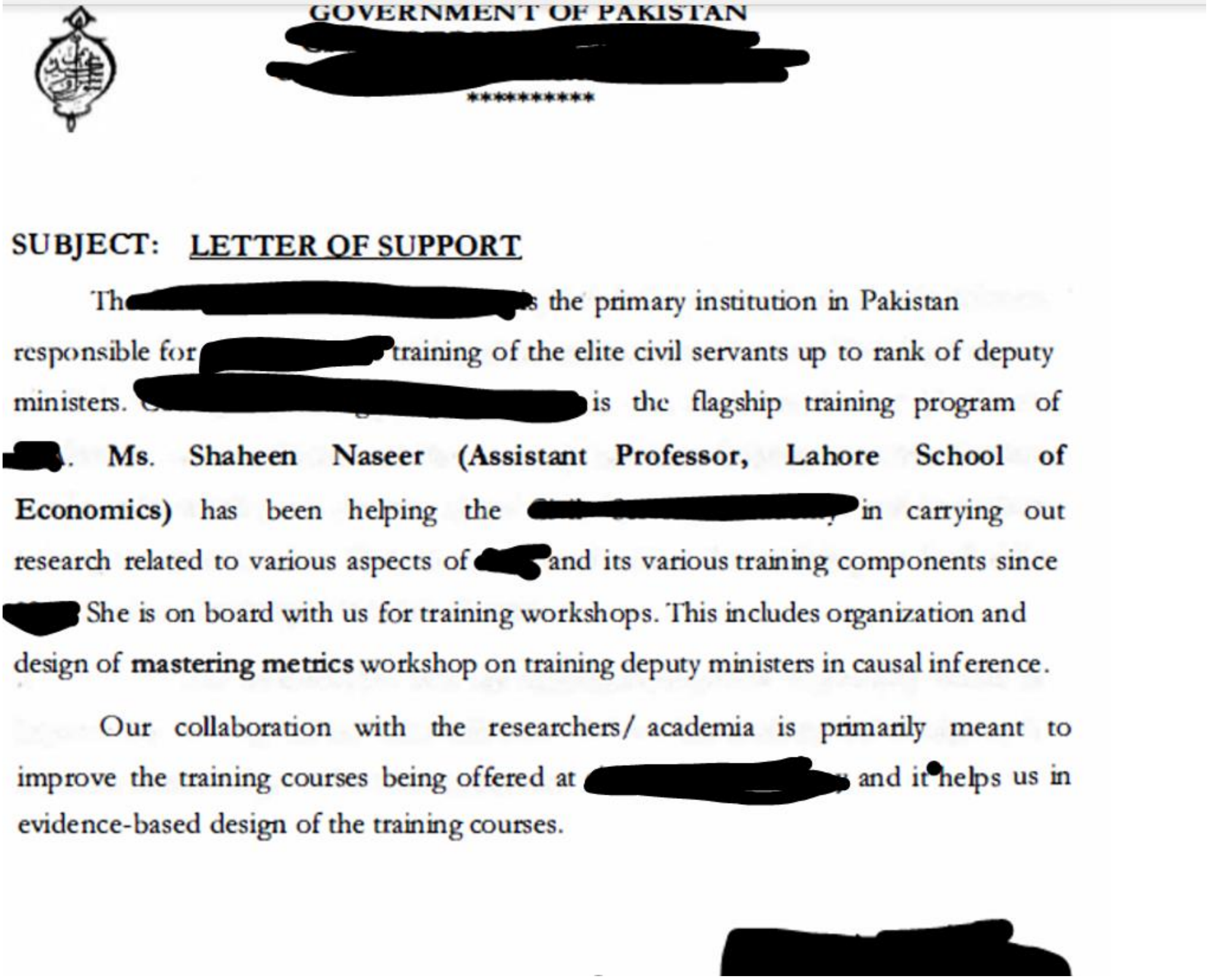
Note: The figures above compare beliefs on importance of quantitative (Panel A) and qualitative (Panel B) evidence for those before and after lecture and discussion for treated and control group. The left panels are those treated with metrics book and corresponding assignments, while figures on the right 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 present results on rating of quantitative evidence and Panel B present results on rating of importance of qualitative evidence for our metrics trained relative to the placebo group.

Figure B4: Shifts in Beliefs versus Initial Beliefs



Note: The figure above plots shifts in beliefs from prior to posterior relative to prior beliefs for individuals assigned metrics training and those assigned placebo training. Fitted lines are also shown for both metrics trained treated and placebo group. Vertical line represents the signal value of 13% impact of deworming on income.

Table B1: Letter of Support



GOVERNMENT OF PAKISTAN

[redacted]

****

SUBJECT: LETTER OF SUPPORT

Th[redacted]s the primary institution in Pakistan responsible for [redacted] training of the elite civil servants up to rank of deputy ministers. [redacted] is the flagship training program of [redacted]. **Ms. Shaheen Naseer (Assistant Professor, Lahore School of Economics)** has been helping the [redacted] in carrying out research related to various aspects of [redacted] and its various training components since [redacted] 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 [redacted] and it helps us in evidence-based design of the training courses.

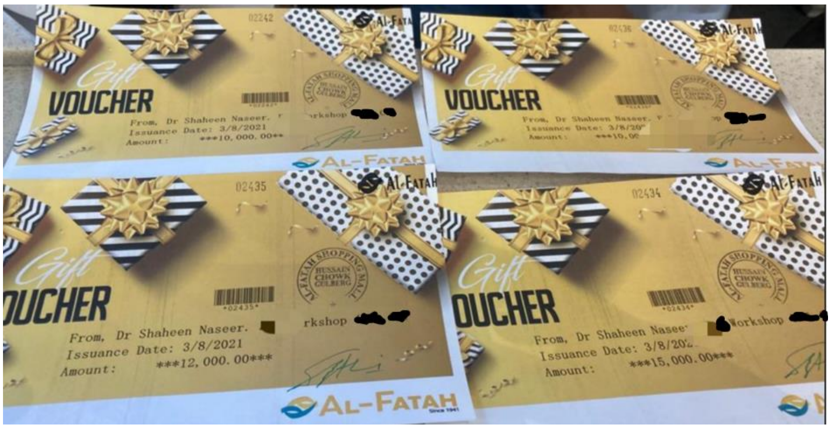
[redacted]

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 brackets. 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: Impact of Treatment by Compliers

	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 brackets. 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, 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 switch on when 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 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 B5: 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	1.735***	1.348***	0.166	0.168	2262.053**	2701.089**	1668811**	2883912**
	(0.220)	(0.422)	(0.114)	(0.153)	(1028.618)	(1114.788)	(637829)	(1186649)
Metrics Assigned Equal (p-value):	(1) = (2):	[0.501]	(3) = (4):	0.964	(5) = (6):	[0.555]	(7) = (8):	[0.482]
Individual Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	84	84	84	84	84	84	84	84
R-squared	0.592	0.448	0.212	0.510	0.402	0.295	0.167	0.285

Robust standard errors clustered at individual level appear in brackets. 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 metrics book is 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 B6: 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
	(0.000) *** {0.000} ***	(0.000) *** {0.000} ***	(0.487) {0.491}	(0.554) {0.533}	(0.046) ** {0.075} *	(0.011) ** {0.012} **	(0.085) * {0.104}	(0.080) * {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	2.657	2.657	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 B7: 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) ***	(0.0001) ***	(0.931)
	{0.007} ***	{0.0001} ***	{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 similarly policymakers. The course content cover scenarios that apply concepts of public goods, externalities, the use of data in policymaking. Columns (2) presents score 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 bureaucrat. Finally, in Columns (3) score on Research Method is reported. The workshop content included 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 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 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 B8: 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	0.081
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 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 B9: 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 brackets. 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.