Training Policymakers in Econometrics

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

January 2024

Training junior ministers in the school of thought associated with the credibility revolution increases demand for and responsiveness to causal evidence. Using a simplified Becker Degroot Marshak lottery, we randomize incoming policymakers into an econometrics training program. Treated policymakers' stated valuation of quantitative evidence and commissioning of RCTs in policymaking increases. One year after the training, treated policymakers are more likely to recommend funding for policies for which there is RCT evidence. Overall, our findings suggest econometrics training may provide a foundation for enhancing the appreciation and utilization of causal knowledge among policymakers. (*JEL D72, D78, O17*)

Keywords: randomized clinical trials, policy, credibility revolution, paradigm shifts.

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

"*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).[2] We study the causal effects of a paradigm shift in the social sciences (Kuhn 1962) on practitioners–policymakers–using the training of the paradigm as its instrument. There is growing academic interest in estimating the value of evidence-based decision-making (Abadie et al., 2023). Policymakers demand and may even respond to evidence (Hjort et al. 2021). Still, 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; Metzger et al., 2020; Vivalt and, Coville 2021; Lu and Chen 2021). Sticking to priors and being inattentive to evidence may stymie implementing good policies that might otherwise spur economic development (Kremer et al., 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

[2] It does this by comparing the world to a counterfactual scenario in absence of the intervention: "… while probabilities encode our beliefs about a static world, causality tells us whether and how probabilities change when the world changes, be it by intervention or by act of imagination." (Pearl and Mackenzie, 2018).

of Pakistan the “key wheels on which the entire engine of the state runs” (Government of Pakistan, 2019). They are almost identical to the elite bureaucrats of India who are called the “steel frame of India” (Bertrand et al., 2020, p. 627). These elite bureaucrats in Pakistan are recruited, trained, and incentivized in a manner similar to many developing countries, especially those countries that inherited these bureaucratic institutions during British colonial rule. Pakistan, India, and Bangladesh alone consist of more than a quarter of the world's population, making this study potentially relevant for many people.[3] We also studied the impact of training causal thinking in a policy decision involving deworming, a policy that shares many essential characteristics with other development policies, with the policymakers aiming to decide 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 econometrics training, then randomized the econometrics training, and measured consequences on attitudes, behavior, and officials’ policy decisions. The demand for the Mastering ’Metrics book yielded a proxy for potential compliers to the treatment assignment. We controlled for whether individuals chose the metrics book and assessed whether similar effects were observed for both high- and low-demanders, that is, we 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 the impact of schooling on wages. The other book was Mindsight, a self-help book that focuses on developing a positive outlook toward life and serves as our placebo (Siegel 2010). We isolated the effects of causal thinking separate from the demand for causal thinking with a simplified Becker Degroot Marshak mechanism. More specifically,

[3] For instance, India, Pakistan and Bangladesh all recruit these elite bureaucrats through a highly competitive exam and use scores in these entrance exams and assessment scores in the training academy as one key metric for promoting and evaluating these bureaucrats. Nevertheless, recent evidence of these elite policymakers suggests their limited ability to interpret numerical information (Callen et al., 2017; Metzger et al., 2020). Analysis of similar bureaucrats in India is conducted in Bertrand et al (2020).

deputy ministers chose a high or low probability of receiving one of the two books. The outcome of the lottery served as the instrument for estimating the causal effects of causal thinking, controlling for the probability of receiving the metrics or placebo book. The lottery completely determined the random assignment, allowing us to estimate the causal effects of receiving Mastering ’Metrics for those more likely vs. less likely to comply with the treatment.

The meat of our intervention 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, Joshua Angrist and Daniel Siegel, along with competitive writing assignments. We offer deputy ministers both sets of lecture videos and track their click behavior. The econometrician typically cannot observe defiers in an IV framework, so we use the click behavior to proxy for defiers (those who click on the video not assigned to them) and never-takers (those who choose not to click but have training assigned to them). Moving beyond lectures, we designed writing assignments inspired by theory and empirical evidence on the efficacy of social-emotional learning. Income deputy ministers were assigned to write two essays as part of the training program. The first essay summarized *every* chapter of their assigned book, while the second essay discussed how the materials would apply to their career. The essays were graded and rated competitively. Writers of the top essays were given monetary vouchers and received peer recognition from their colleagues (via commemorative shields, a presentation, and a 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 and 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 deputy ministers' attitudes, behavior, and policy decisions. We had essentially no attrition because we embedded our analyses in administrative data. We observed a balance on pretreatment quantitative ability as measured from quantitative scores in the entry examinations of the deputy ministers obtained from the Federal Public Service Commission (FPSC), an independent government arm that administers the entry examinations of these elite policymakers. Likewise, we observed balance on demographics, pretreatment writing and interview assessments. We also obtained data on the ministers’ regular policy assessments from the training academy. These were conducted 4–6 months following our workshop, and deputy ministers were scored independently in research methods and policy assessments. Our first main finding is that training causal thinking shifts policy attitudes. We surveyed policy attitudes on the importance of causal inference several months after the treatment assignment and performed a textual analysis of the high-stakes writing assignment. We find substantial effects. While attitudes on the importance of qualitative evidence are unaffected, treated individuals' beliefs about the importance of quantitative evidence in making policy decisions increase from 35% after reading the book and completing the writing assignment and grows to 50% after attending the lecture, presenting, discussing and participating in the workshop. We also find that deputy ministers randomly assigned to causal training have higher perceived value of causal inference, quantitative data, and randomized control trials. 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 causal inference and desire to run randomized trials. We also observe substantial performance improvements in scores on regular national research methods and public policy assessments. However, we find no effect of our metrics training treatment in policy assessments unrelated to econometrics or quantitative analysis. The research team did not specially request these regular assessments, so performance improvements are unlikely due to experimenter demand.

Our second main result emerges from the policy decisions of junior ministers. We provided a signal—an email summary of a recently published randomized evaluation on the long-run impacts of deworming to all deputy ministers (Kremer et al. 2021). We find that treated deputy ministers were twice as likely to demand fiscal support from the Federal Government to support deworming policy and –in annual budgetary requests made to their respective government divisions– recommend over three times the amount of funding for the deworming policy relative to the placebo group. Two alternative policies —renovating orphanages and schools— for which no RCT evidence was provided were unaffected by the metrics training. Renovating orphanages and schools was under spending review in the same budget cycle, so it is a natural control. These budgetary requests are made by the junior ministers, independent from the experimenter and the training academy, making them particularly helpful for the interpretation of our results: when faced with policy choices having real reputational costs, implementation challenges, and public budgetary constraints, treated policymakers choose the policy for which there is causal evidence.

Our third main result comes from a framed field experiment, where consistent with the results on actual policy impact —official letters sent and funds recommended— policy beliefs and hypothetical decisions directly relating to deworming signal are also affected. First, we elicited initial beliefs about the efficacy of deworming on long-run labor market outcomes. Then, the ministers are asked to choose between implementing a deworming policy versus a policy to build computer labs in schools. After the signal about the impact of deworming, we asked the same deputy ministers about their post-signal beliefs and to make the policy choice again. From this experiment, we observe that only those assigned to receive training in causal thinking showed a shift in their beliefs about the efficacy of deworming: treated ministers assigned to the treatment group updated their beliefs in the direction of the signal. After receiving the RCT evidence signal, the treated ministers also became more likely to choose deworming as a policy. The magnitudes are substantial—trained deputy ministers doubled the likelihood of choosing deworming, from 40% to 80%. An effect size similar to the doubling of deworming recommendation letters sent to the respective governmental divisions. Notably, this shift occurs only for those ministers who previously believed the impacts of deworming were lower than the effects found in the RCT study.

We also measured ministers' *stated* willingness-to-pay for three sources of information: RCTs, correlational data, and expert bureaucrat advice. We elicited willingness-to-pay for correlational data and senior bureaucrats' advice because these two alternative sources of information are the status quo that deputy ministers use to inform their policy decisions. We observed that treated deputy ministers were much more willing to spend out of pocket (50% more) and from public funds (300% more) for RCTs and less willing to pay for correlational data (50% less). In other words, the treatment did not increase willingness to seek *any* data and evidence but rather shifted policymakers' beliefs towards the paradigm associated with the credibility revolution. The tripling of stated willingness-to-pay for RCT is also close to the actual funds recommended for deworming to the Federal Government of Pakistan.

Econometrics training likely induced deputy ministers to choose deworming and rate quantitative evidence differently because they learned causal inference concepts, a fact suggested 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. We also find that the effects of metrics training are similar for those who express high versus low demand for econometrics training and even those with high versus low pretreatment quantitative test scores. This suggests that the initial demand for learning about metrics and math ability may not be crucial for policymakers' receptiveness to econometrics training.[4]

Our results are robust to a series of sensitivity checks. First, we show that in addition to the randomly assigned groups being balanced across individual characteristics and in pretreatment quantitative ability as measured by their entry mathematics assessment scores, the results are also robust to randomization inference and multiple hypothesis tests. Together, these robustness checks suggest that small or idiosyncratic samples assigned to treatment or control are unlikely to explain our results. Second, since the study was conducted online during the COVID-19 pandemic, spillovers were likely minimized because the deputy ministers were not together on-site at the regular training facility. Third, we show only roughly ten percent of deputy ministers attrited, which is also balanced across treatment and control.

Fourth, 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 of deworming changed their project choices. Experimenter demand is also not reflected in alternative policies that were also underspending review during the same budget cycle: funding requests for orphanage and school renovation —policies for which no causal evidence was provided— are unaffected by the metrics training. More importantly, however, the budgetary requests to the government are made independently by the experimenter and the academy and, hence, unlikely to be driven by experimenter demand effects.[5] Last, we show that those who ex-ante demanded the metrics versus placebo training are similarly impacted by the metrics training, suggesting that the demand for metrics training also had little bearing on its effectiveness.

[4] This may be due to the fact that our econometrics training's focus was on concepts and intuition rather than mathematical formulae.

[5] Especially relevant is the fact these budgetary requests are made more than 6 months after these ministers have received their final assessment scores from the academy and have already graduated. Upon graduation, the academy effectively loses all power to transfer these bureaucrats by virtue of their final assessment scores being already determined. These bureaucrats are then only next trained 10 years later, by a different and independent institute (called National School of Public Policy).

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 and the utilitarian cost-benefit perspective associated with data science (Frank et al. 1993; Rubinstein 2006; Bonnefon et al. 2016; 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 science: the credibility revolution (Angrist and Pischke, 2010; Pearl and Mackenzie, 2018). To our knowledge, we 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) and numerical literacy and problem-solving (Deming 2022). 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 impacted the high-stakes decisions of policymakers and explained up to 30% of the recent shift towards economic conservatism in the American judiciary (Ash, Chen, and Naidu 2021). Our study is closest to an RCT of eight hours of financial literacy training that impacts the economic preferences of adolescents (Sutter, Weyland, Untertrifaller, Froitzheim 2020) and an RCT that included two hours of financial literacy training that impacted those who had low levels of financial literacy (Cole, Sampson, and Zia 2011). However, we study the impact of econometric literacy training — in causal thinking — on the 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 is an understanding and appreciation of causal evidence. This, in turn, may promote implementing good policies that might otherwise have high rates of return for economic growth. By shaping deputy ministers' causal thinking with 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. In contrast to the predominant focus of numerous training studies on lay populations, our investigation delves into the dynamics of high-stakes decision-makers. Specifically, earlier research scrutinizes cohorts that include central bankers (as discussed by Malmendier et al., 2017), senior deputy ministers (Mehmood, 2022 in Pakistan and Bertrand et al., 2020 in India), and judges (as highlighted in the work of Ash et al., 2016). We trained junior deputy ministers' causal thinking as they joined civil service and estimated the impact on attitudes and subsequent demand for evidence and policy choices. We *cautiously* interpret our findings as suggesting that causal thinking not only increases responsiveness to causal evidence but *may* also correct for some mistakes in the belief-updating process in response to evidence. The rest of the paper is organized as follows. Section II provides the background and details on the experimental setup. Section III describes the data and empirical specification, while Section IV presents the main results. Section V conducts a heterogeneity analysis. Section VI discusses a series of sensitivity tests. A final section concludes.

II. Background Context and Study Design

A. *Background*

Study Context.— We conducted a randomized evaluation implemented through close collaboration with an elite civil service training academy. The Academy in Pakistan is one of the most prestigious training facilities that prepares top brass policymakers—*junior* deputy ministers—for their jobs (Mehmood et al., 2021). 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 senior deputy ministers entail policymaking, policy recommendations, implementation, and advisory positions to the President, Prime Minister, and cabinet ministers. These deputy ministers attend training workshops at the Academy, taking part in several workshops and assessments designed to hone and assess their policy skills right after joining civil service. We obtained access to almost all the incoming deputy ministers entering service in a single year with our training embedded alongside their regular training workshops on professional etiquette, bureaucratic procedures, social skills, management, and public policy workshops.[6] We designed a "mastering metrics" training workshop for these deputy ministers and delivered it as they participated in the Academy's regular training program at the Academy. We obtained access to ministers' policy assessments for other workshops, alongside our mastering metrics training workshop and national policy exams.[7] *Treatment design details*

October: Baseline survey and book choice.— We conducted a baseline survey and asked the participants to choose one of two books (1) Mastering 'Metrics: The Path from Cause to Effect by Joshua Angrist and Jörn-Steffen Pischke or (2) Mindsight: The New Science of Personal Transformation by Daniel J. Siegel on 20th October 20XX. The first book is an accessible introduction to the fundamental problem in causal identification and summarizes key concepts associated with the credibility revolution. These include RCT as an experimental ideal, regressions as comparison of means, instrumental variables, difference-in-differences and regression discontinuity designs with a particular focus on public policy applications. The book is written for undergraduates and is particularly appropriate for our policymakers since all of

[6] 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.
[7] The director's letter of support is attached as Table B1 in Appendix B.

them at least holding a bachelor's degree. The second book is a popular self-help book emphasizing “personal transformation” and serves as our placebo.[8]

November Assignment of treatment.— In 10th November, the director of the elite Civil Service. The Academy administration sent a request 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 would complete the assignment within the deadline. The close collaboration with the partner organization 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.[9] 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 after a month (the set deadline). The full detailed transcript of the message by the director detailing their assignment tasks can be found in Table A1 of Appendix A.

March: Attitude Survey, lecture, presentation, discussion and workshop.— On 10 March, in collaboration with the Academy, we organized two Zoom sessions, one for a randomly

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

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

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). We distributed commemorative shields, often accompanied by peer recognition, and monetary gift vouchers to the top 6 performers. After conducting the endline survey on attitudes, we announced the first three positions for both groups and distributed the commemorative shields and gift vouchers to a luxury departmental store. The 1st position received a monetary voucher of USD 150, the 2nd position received a USD 100 voucher, and the 3rd position received a USD 80 voucher. The placebo group also received the vouchers, and hence we had 6 winners. These winners also gave a 30-minute presentation summarizing key lessons of the respective books and how the training will inform their policymaking within the treatment arm. This was followed by 30-minute video lectures delivered by the authors of the books to the respective randomly assigned groups. The group assigned Mastering 'Metrics attended the video lecture by Joshua Angrist and the group assigned Mindsight attended the video lecture by Daniel Siegel. A structured discussion of 30 minutes for both arms followed. In particular, we asked participants the following two questions: (a) What do you think is the main point of the lecture? (b) How can you apply the concepts learned in this lecture to your job? In the end, this part of our engagement with the ministers concluded by asking the same questions on attitudes towards quantitative and qualitative evidence, randomized evaluations and causal inference. This allowed us to assess the short-run impact of our complete metrics training, i.e., essays summarizing the book, essays applying the lessons to policy, attending the video lecture, receiving commemorative shields and gift vouchers, presentation and discussion of key lessons learned. Table B2 in Appendix B presents screenshots of commemorative shields and gift vouchers distributed to the deputy ministers.

May: Initial Beliefs, Post-Signal Beliefs, Willingness-to-Pay and Project Choice.— In May, about 6 months following the book assignments, we elicited policymakers' beliefs on the impact of a policy and asked them to make policy choices. Specifically, we elicited their beliefs on the impact of deworming of children in schools on earnings 20 years later. We also elicited a response on a policy choice between choosing to implement a deworming initiative or an ICT initiative to establish computer labs in a small city of Kuchlak in Balochistan. The decision between the deworming and ICT policies was used because it was an actual policy choice faced by similarly ranked public officials at the time. In particular, we asked them to choose either to implement a computer lab policy that involved installing a computer lab in each school of Kuchlak or implement a deworming policy where they launched deworming campaigns in all schools of the same town.[10] To minimize experimental demand effect, we nudged policymakers to choose an alternative computer lab project. That is, before the project choice the policymakers see the following prompt:

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

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

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

Finally, we collected post-signal beliefs on deworming's impact and updated policy choices between the deworming and computer policies following this signal.[11] For more details on the willingness-to-pay transcripts of questions and project choice, see Table A3 in Appendix

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

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

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 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. So, the geographical dispersion of the ministers due to the pandemic at the time of the training and the non-shareability of the link reduced treatment spillovers. However, it should be noted that treatment spillovers would likely suggest that our estimates are underestimated.

September: Actual Policy Choices.— About one year following the training, the deputy ministers made policy decisions in the field. As part of their official duties, they wrote letters addressed to their respective government divisions recommending funding for certain policies. The first policy recommendation was for a deworming policy in schools, while two other alternative funding recommendations were for school and orphanage renovations. A senior official their respective divisions in Pakistan issued a call for funding recommendations from the deputy ministers for the next budgetary cycle. The difference between this and last year's call was two additional policy choices that we were able to embed within the call. The first was the possibility of recommending funding for a deworming policy in schools for which we had provided a signal of causal impact on wages. The second policy choice, we also added, was the possibility of recommending funding for orphanage renovations, for which no signal or causal evidence was provided. The third recommendation in the call was for school renovations which was included based on the ministry's internal needs assessment. The renovation of orphanages and schools was under spending review in the same budget cycle, so it serves as a natural control. Therefore, the call for funding recommendations from the respective government ministry included policies for which no RCT evidence of its efficacy was provided, i.e. renovating orphanages and schools, while one option was to recommend funds for the deworming policy for which we did provide causal evidence. Using this administrative data, we are able to ascertain who recommended which policy and the exact amount of funds each deputy minister actually recommended for the respective policy. This policy choice is high-stakes because every letter has a reputational cost since it is signed with the full name of the minister and they are typically charged with overseeing the implementation of the recommended policy in the following fiscal year.

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) which administers the entry examinations for these elite policymakers.[12] Data on WTP, attitudes and beliefs were collected by our research team under the auspices of the Federal Government of Pakistan.

Outcome Variable on Policy Decision.— We obtain data for fiscal support or budgetary requests of deputy ministers from the respective government ministries where these civil servants serve in Pakistan. These annual budgetary requests for fiscal support, made months after graduation from the Academy, are independent from the potential experimental demand effects of the experimenter and the Academy. The data for fiscal support of the deputy ministers is available for three policies: one related to our signal of RCT evidence (deworming policy) and two placebo policies (school and orphanage renovations) unrelated to our signal. The funding requests are made roughly a month before the federal budget for the next fiscal year is announced every year.

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

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

[12] 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.

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

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 suggests 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 suggests our sample is also balanced in quantitative ability. The close collaboration with the training Academy and the director resulted in our intervention to have a 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 metrics training has no significant effect on attrition.

IV. Main results

A. *Treatment Effect on Attitudes*

Treatment effects on the 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 policymakers' attitudes just before the lecture, presentation and discussion of the material related to the assigned book—i.e., we measure the causal effect of partial training via book assignment of summarizing and application of concepts to policy. We find that policymakers treated with this training rate the importance of quantitative evidence in policymaking by about an additional point. This is a substantial effect and equivalent to a nearly 35% increase over the average rating of the placebo group. In Column 2 of Table 2, we present results of our full training where policymakers attend respective video lectures from the authors of the book, receive a gift voucher, and commemorative shields, present the key lessons of the respective book, and partake in a structured discussion of the material in a workshop (full metrics training). The evidence is consistent with metrics training being reinforced: effect sizes increase by about 50% and are statistically different from those obtained prior to the reinforcement training. In particular, full metrics training (book assignments, lectures, presentations and discussions) increases rating on the importance of quantitative analysis in policymaking by about 1.5 points on a 5-point scale. This is a 50% increase in ratings over the placebo mean.

Following this, we conducted a falsification test on a similar question that helps mitigate potential concerns that individuals rate *any* evidence higher regardless of its quantitative nature. The dependent variable in this case is the rating on the question "How important do you think *qualitative* analysis is in public policy making?" Columns 3 and 4 of Table 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 the 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 policy scenario:

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

One of the options is to "Run a randomized control trial", while other options are unrelated or inconsistent with the main message of the book, such as "Compare two groups of people who had previously benefited most from the policy with those that did not?" and "Survey if there is demand for the policy". These options appear in random order. The dependent variable takes the value of one if the policymaker answered "Run a randomized trial" and zero for all other options. The results of estimating equation (1) with this dependent variable is reported in Table 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 suggest that months after the training, treated policymakers' perceived importance of quantitative evidence and randomized evaluations increased, while we observe no effect on importance of qualitative evidence.

Why did the policymakers demand randomized evaluations?— What explains policymakers attaching greater importance to quantitative analysis and randomized evaluations? Here we present some evidence that the results may be explained by the fact that policymakers learn about causal inference and selection issues. In the last two columns of Table 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 these results with partial and full metrics training. Ministers assigned metrics training are 15 percentage points more likely to answer that randomized evaluations are "Apple to apple comparisons," suggesting that they understand that random assignment of subjects in the control and treatment groups solves the selection problem by "comparing apples to apples". This is one likely explanation why our treated group may have higher perceived importance of quantitative evidence and randomized evaluations.

In Figure 1, we report all results of Table 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 request for randomized trial to evaluate effectiveness of public policies, and about 0.30 standard deviation increase in answering that randomized evaluations allow for apple to apple comparisons relative to the placebo group. We find no effect of metrics training on beliefs about qualitative evidence, however.

These results are consistent with textual analysis of the ministers' high-stakes assignments. Though, suggestive, the analysis of their writings suggests that the metrics assigned group likely learned many causal inference concepts. Specifically, in Figure 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". The metrics training appears to affect policymakers' attitudes towards the paradigm associated with the credibility revolution.

B. *Treatment Effect on Policy Assessments*

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

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

The responses were scored by a panel of experts (former Supreme Court judges, prominent academics, and former senior deputy ministers) and the assessment was high-stakes since it determined their future career trajectories. Table 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 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).

B. *Treatment Effect on Policy*

While the results so far are suggestive of the potential impact of the metrics training on policy, we are yet to provide direct evidence of metrics training impacting actual policy. About 12 months after the book assignment, in September 2021 we observed deputy ministers' actual policy decisions: letters written to the government ministeries for funding recommendations of alternative policies. We embed within a call for funding recommendations requested by the government two policy options: deworming and orphanage renovation policies. For the former policy, we provide causal evidence for its impact, for the latter, no such evidence is provided. These funding recommendations sent to their respective government division have a reputational element and deputy ministers are typically charged with implementing policies they recommend. The first policy choice was for a deworming policy in schools, while two other alternative policy choices were to obtain funds for school and orphanage renovations. Using this administrative data on these policy recommendations of the deputy ministers, we ascertain both their choice of policy and the amount of funds recommended for each policy. This can be viewed as both an extensive and intensive margin measurement of the policy decision.

Table 4 reports these results. We find that metrics-trained ministers are about 30 percentage points more likely to write letters to recommending funds to implement the deworming policy. This is a substantial effect and equivalent to about doubling of letters written for the deworming policy (Table 4, Column 1). Likewise, we also find that metrics trained policymakers are likely to recommend about Pakistan Rupees 400, 000 (USD 2500) more, for the deworming policy, than the placebo policymakers who request about 171, 000 (Table 4, Column 2). This is more than double the funds requested relative to the placebo group. Taken together, the metrics training appears to impact deputy ministers in their official duties: issuing policy recommendations on budget allocations that have costly reputational consequences.

Notably, we find no effect of metrics training on other alternative policy choices for which no RCT evidence was provided. Metrics-assigned policymakers are neither more likely to write letters or recommend additional funding for school and orphanage renovations (Table 4, Columns 3-6). The point estimates are small and statistically insignificant, suggesting an experimental demand effect of writing more letters and advising additional funding in all policies is unlikely to be behind our results. In the appendix, we also show that the effects of metrics training are similar for those who express high or low demand for econometrics training (see Table B4 in Appendix B). The treatment effects of metrics assignment are not significantly different for those who chose the econometrics book or chose the placebo book. Taken together, the evidence suggests that when faced with policy choices having real reputational costs, implementation challenges and public budgetary constraints, treated policymakers choose policy for which there is causal evidence.

D. *Treatment Effect shifts beliefs*

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

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

Interestingly, we found that the policymakers underestimated the long-run impact of deworming relative to the 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 assigned the mastering metrics training is substantially more likely to respond to the signal by shifting their prior beliefs towards the signal value of 13%, while the placebo-trained group is unaffected by the signal. This strongly suggests that the metrics training shifted policymakers' beliefs on the long-run impact of deworming.

To investigate individual-level shifts in beliefs, we compare the distribution in belief updating following the signal with policymakers' initial beliefs. Figure B4 reports this shift in beliefs after receiving the signal for the two different groups, the treatment group in the diagonal solid line and the placebo group in the more horizontal dotted line. The top panel shows the treatment and control groups for those who chose the metrics book, and the bottom panel shows the same for those who chose the placebo book. Three features of the graphs are particularly noteworthy. First, the results are similar regardless of the book choice. Second, the shift in beliefs for the treatment group falls along a diagonal that intersects the x-axis roughly at the signal. That is, treated individuals whose prior beliefs are 13% update very little on average. Moving to the left along the x-axis, treated individuals update their beliefs in a positive direction, though do not necessarily jump to the signal. Likewise, moving along the right, treated individuals update their beliefs in a negative direction. Third, in contrast, the placebo individuals almost never update their beliefs. Some placebo individuals even do not react to the signals in a Bayesian manner. This can be seen from two features of the data. First, placebo individuals whose beliefs are below the signal occasionally update their beliefs in a direction opposite of the signal. Second, a moderate number of placebo individuals whose initial beliefs are near the signal seem to overreact to the signal by updating past the signal, i.e., in a positive or negative direction by an amount almost equal to the original signal in absolute value. This finding suggests that in the absence of econometrics training, policymakers may misinterpret quantitative evidence.

Next, we use the differential updating of beliefs around the signal to offer a mechanism for the impact of metrics training on policy (results in Section C). We conduct an analysis where we consider two endogenous variables: the update in beliefs after receiving the signal and the update in beliefs interacted with whether the initial belief is above the signal. The update in beliefs is instrumented for by the metrics assignment and the update in beliefs interacted with whether the initial belief is above the signal is instrumented for by the metrics assignment interacted with the initial belief being above the signal. Table 5 reports these results. For individuals whose prior beliefs are below the signal, using the variation in belief updating associated with econometrics training, we see an increase in the beliefs about the 20-year impact of deworming on wages by one percentage point effect is associated with 4% greater likelihood to recommend deworming as a policy. However, for those whose prior beliefs are above the signal, a decrease in one percentage point effect is associated with about 5 percentage points greater likelihood to recommend deworming. These results suggest that the effects of econometrics training is possibly explained by those who shift their beliefs towards the signal in response to causal evidence on the efficacy of deworming.

E. *Treatment Effect on Demand for Evidence*

Effect of Metrics Training on Stated WTP. — In May 2021, we also elicited policymakers' demand for evidence. Specifically, we elicited the stated WTP from both private and public funds for three pieces of information for: (1) results from a RCT on the impact of deworming on long-run income; (2) correlational data on incomes of schools with and without deworming program; (3) advice from senior public officials on the impact of deworming policy. The latter two choices are status quo sources of information available to the ministers. In Table 6 (Column 1), we find that metrics trained policymakers' stated willingness-to-pay is about PKR 2000 (USD 13) from their own pocket (private funds) for causal evidence. This is equivalent to about 1% of their monthly salary. It is, however, unclear whether these metrics-trained policymakers would also state they will pay for causal evidence from public funds, especially when their budget is constrained and other available sources of information, such as correlational data and advice from senior officials, are readily available. Therefore, in Column 4 of Table 6, we elicit WTP for the same information from public funds.[13] The effects from public funds are substantially larger and statistically significant, with metrics assigned to policymakers willing to pay about 1400000 (USD 8500) for the information from a randomized evaluation. This is about twice as the stated

[13] We measure both public and private WTP as Hjort 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." (Hjort et al., 2021, p. 24).

WTP for the placebo group. This is similar to doubling the actual funding recommendations for deworming made to the government (Table 6, Column 4). The evidence, therefore, suggests that even when there are alternative 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 in Figure 4 where we plot the distributions of WTP from public and private funds: metrics assigned policymakers are significantly more likely to state they will pay large amounts from personal and public funds for randomized evaluations than those assigned the placebo training. Interestingly, however, we observe that metrics-trained policymakers *decrease* their stated WTP for correlational data. This finding is consistent with metrics-trained ministers becoming closer to the paradigm of credibility revolution that discounts simple correlations over well-identified studies. In particular, we find that those assigned the metrics training pay about PKR 1000 (USD 6.40) less from their private funds for correlational data comparing incomes of pupils in schools with and without deworming relative to those assigned the placebo training and about PKR 55000 (USD 350) less for this information from public funds (Table 6, Columns 2 and 5, respectively). This is equivalent to a 50% decrease in WTP for correlational data from private funds and a 33% decrease from public funds.

Figure 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 to policymakers willing to pay large amounts for correlational data relative to the placebo group. Finally, we find that metrics training has essentially no effect on policymakers' stated WTP for the advice of senior bureaucrats, another competing source of information available to the ministers. These results are reported in Columns 3 and 6 of Table 6 for stated private and public WTP, respectively. We observe that there is no statistically significant difference between WTP for advice from senior bureaucrats among the metrics trained and placebo group policymakers. In Figure 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.

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

Project choice and heterogeneity by initial beliefs. — In this subsection, we present evidence that the metrics training impacted the stated project choice of policymakers. Earlier, we observed from Figure 3 that metrics training shifted policymakers' beliefs on the impact of deworming on income after being presented with the signal value of 13%, while no such shifting of initial beliefs was observed for the group assigned the placebo training. This indicates that metrics training made the policymakers more responsive to RCT evidence relative to the placebo group. However, we observe interesting heterogeneity in these results. Metrics-assigned policymakers are *only* more likely to choose the deworming relative to the computer lab policy if their initial beliefs on the impact of deworming is below the signal value impact of 13%. In particular, metrics-trained policymakers whose priors are below the signal value of 13% impact are about twice as likely —40% to 80%— to choose to implement the deworming policy when causal evidence is provided to them (Figure 7 of Panel A). The effect is observed 6 months following the metrics training, suggesting a persistent effect of our intervention. We do not find much evidence of the metrics training impacting policymakers that had above the 13% prior belief on the impact of deworming on income. This indicates policymakers shifted their beliefs and chose a policy for which there was causal evidence only insofar as they initially believed that the deworming did not have much impact. As Cantoni et al. (2019) suggested, we further disentangle heterogeneity by prior beliefs via semiparametric 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 cautiously interpret these patterns as broadly consistent with metrics training causing individuals to be more receptive to causal evidence.

Heterogeneity by Defiers and Never-Takers. — We next studied heterogeneity by defiers and never-takers. The tracking of the click behavior of policymakers across metrics and placebo lecture videos allowed us to investigate whether individuals assigned the metrics training attempted to defy the treatment assignment and attempt to access the placebo training. Likewise, we were able to investigate whether there were individuals who were assigned the treatment but never clicked on the lecture videos (never-takers). Therefore, the unique setting allows us to proxy for defiers and never-takers to the metrics training treatment and examine the potential heterogeneous impacts of our treatment. By matching the individual code and official email of policymakers, which they had used to log in to access the assigned lecture, we could directly observe and track the click behavior of individual ministers using oTree (Chen, Schonger, and Wickens, 2016).[14] We also used the expertise of a computer scientist which made it nearly impossible for the policymakers to share, download, or access training to which they were not assigned. In Table 7 (Columns 1 and 3), we report the heterogeneous impact of metrics training assignment on defiers—those who were assigned the metrics training lecture but attempted to access the placebo training lecture. We found that defiers are significantly less likely to be impacted by our metrics training. For example, we observed that metrics trained defiers are willing to pay—from public funds—about PKR 2,532,000 (USD 19000) less for randomized evaluations (Table 7, Column 3). We found that never-takers are less likely to be impacted as well, but to a lesser extent (Table 7, Column 4). These results suggest that metrics training did not uniformly impact all policymakers and that behavioral data can be a potential method for detecting defiers and never-takers in estimating the average treatment effects in randomized trials.

Heterogeneity by High and Low Demanders for Econometrics Training.—Finally, we used our behavioral elicitation of potential compiler status to assess a typical concern of RCTs. Namely, in RCTs, the econometrician estimates the Local Average Treatment Effect (LATE) for the compliers that respond to treatment and the econometrician typically is unable to observe defiers. It is a plausible concern that people who demand to learn causal thinking may be more responsive to the treatment assignment. Thus estimates of the treatment impacts would be uninformative on those who are potential non-compliers. In our unique experimental set-up, we

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

developed a proxy for compliers through those who demanded the metrics book; we show that the effects are the same for both the high and low demanders. As can be seen from Table B5 of Appendix B, we observe no significant differences between the treatment effects for low and high demanders of metrics training. The interaction term is asking whether the metrics assignment has a differential impact for people who chose the book, which we do not find evidence for. The level term "Metrics Assigned" is the main varaible of interest and shows that the effect is statistically significant regardless of whether individuals choose the metrics book or not. Note that we do not show the coefficient on book choice since it does not have a causal interpretation. However, the interested reader can view Table B4 or the bar chart in Figure B5 which shows that those who demanded the metrics book and those who do not are equally likely to be impacted by our metrics training. For a disaggregated analysis of samples of metrics book versus placebo book chosen, please see Appendix C.

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 scores respond more to causal evidence. Table B6 presents these results. We find no evidence of the heterogeneous effect of metrics training — on policymakers who had high versus low quantitative ability. The effects of our treatment are very similar in terms of point estimates for those with above or below median quantitative scores. These results suggest that our instrument for the paradigm shift has a similar effect regardless of pretreatment quantitative ability.

VI. Robustness and Discussion

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

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

Sample Size and Statistical Power. — The focus on deputy ministers allows us to study an elite group of high-stakes decision-makers who can potentially impact long-run economic and political development. However, the selective nature of these policymakers necessitate that they are by design few in number. Therefore, our sample is restricted to about 200 deputy ministers, which raises concerns about lack of statistical power. Nevertheless, even with 200 individuals, our evidence complements several classical experimental studies that train individuals with less statistical power. For instance, the Abecedarian Program (n = 111), the Perry Preschool Program (n = 123), and the Jamaican Study (n = 129) (Muennig et al., 2011; Heckman and Karapukula, 2019; Walker and Himes, 1991). Our power calculation with statistical power of 80% and significance level of 5% reveals that even in our sample, the individual level randomization allows us to detect a minimum detectable effect equivalent to a change of 0.23 standard deviations. Still, 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 B7, B8, B9 and B10 of Appendix B with 1000 iterations of this process.[15] Even though the p-values slightly increase, the treatment effects are still statistically significant at conventional significance levels. These results strongly suggest that idiosyncratic small sample bias is unlikely to explain our results.

Experimental Demand. — It is also unlikely that experimental demand drives our results, i.e., deputy ministers in the metrics training treatment are providing responses in a manner they feel are simply expected by the experimenter. This is due to several reasons. First and foremost, the annual budgetary requests made by the deputy ministers to their respective divisions are independent from both the experimenter and the academy. Especially relevant is the fact these

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

budgetary requests are made months after these ministers have received their final assessment scores from the academy and have already graduated in a "passing out ceremony".[16] Second, two policies for which no causal evidence signal were also included in the budget documents and they are unaffected by the metrics training. Neither placebo policy was affected by metrics training. Third, in addition to increased WTP for commissioning RCTs, we find no evidence of metrics training on WTP for advice of senior bureaucrats, a potential source of experimenter demand. Fourth, we use a method inspired by the methodology of De Quidt et al. (2018) to mitigate concerns of experimenter demand. We requested the subjects to choose an opposite ICT project than the deworming policy for which causal evidence was provided. In other words, we demanded that a policy other than what we would expect based on RCT evidence presented should be chosen.[17] Last, national policy exams were administered separately from the experimental team. All of these patterns are inconsistent with experimental demand explaining our results.[18]

External Validity. — In this subsection, we discuss external validity. First, the elite bureaucrats of Pakistan, their selection procedures and training are similar to many other developing countries, especially India and Bangladesh who, like Pakistan, inherited these bureaucratic institutions during the British Colonial rule of the Indian subcontinent. Pakistan, India and Bangladesh alone consist of more than a quarter of world population making this study particularly relevant for a large number of people. Second, we follow List (2020)'s *Selection-Attrition-Naturalness-Scaling* (SANS) conditions in our discussion of generalizability of our results. First, in terms of selection, our sample consisted of almost all deputy ministers who entered service in Pakistan via competitive examinations in a given year (that we have anonymized). Considering the nature of the setting, time frame and choice task, we obtained natural measures such as policy assessments and simulations. The policymakers performed natural tasks in the field such as policy choices, teamwork assessments, national public policy assessments and field visits. Finally, in terms of scaling our intervention in other settings, the intervention was cheap to deliver since it was largely online. Since the training was delivered

[16] Upon graduation, the academy effectively loses all power to transfer these bureaucrats by virtue of their final assessment scores being already determined. These bureaucrats are then only next judged 10 years later, by a different institute (called National School of Public Policy).

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

[18] The results are also robust to an alternative specification of a saturated model that also includes an interaction between metrics assigned and metrics chosen (see Table B11 in Appendix B).

online it may also be scaled to other high-stakes decision-makers such as judges and CEOs in other settings. However, we view these results as a WAVE1 insight in the nomenclature of List (2020), and replications need to be completed to understand if the effect sizes can be applied to general populations as well as high-stakes decision-makers in other contexts.

Multiple Hypothesis Testing. — Given the fact that we are testing multiple hypotheses, we also examine if our results are driven by false rejections. Under the assumption that treatment has no effect on any of our outcomes (all our null hypotheses are true), then the probability of one or more false rejections when using a critical value of 0.05 is about 40%. As a result, in order to reduce the likelihood of false rejections, we adjust for the fact that we are testing for multiple hypotheses. Following the literature, we use sharpened False Discovery Rate (FDR) q-values suggested by Anderson et al., 2008 (see for instance Heckman et al., 2018 for an application). These sharpened q-values are presented in square brackets in Table B12 where we also show standard p-values from our regressions in parentheses for comparison. Our results remain robust at conventional significance levels.

Metrics Training and Prosocial Behavior. — A large body of evidence documents that economics training may make individuals less prosocial. Individuals trained in neoclassical economic concepts are more likely to free ride, less likely to donate or cooperate (see, e.g., Marwell and Ames, 1981; Frey and Meier, 2003; Bauman and Rose, 2011). When it comes to evidence, with its attention to utilitarian cost-benefit calculations, it has been suggested that practitioners of data science can become less prosocial (Bonnefon et al. 2016). We, however, present evidence that the metrics training program does not come at the expense of reduced prosociality. We measure prosocial behavior by field measures such as visits to orphanages and volunteering in impoverished schools, as well as language use in their writings. The field measures are obtained from the Academy on the policymakers' "syndicate field trip" workshops. The policymakers undertook two field trips, one 4 months and the other 6 months following the training. In the first, they were provided a choice to either visit a prominent orphanage (*Dar-ul-Aman*) or attend lectures on a specific government program from a senior bureaucrat. In the second syndicate field trip, the policymakers are asked to choose between volunteering to teach in any impoverished government school that falls under the government's Progressive Education Network (PEN) or once again choosing to attend a lecture on government programs from a senior public official. From the top of Figure 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. 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. [19] 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.

VII. Conclusion

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

One year after the training, in their official duties, treated policymakers are twice as likely to actually choose and triple the funding recommendation to the government for policies for which there is causal evidence. Training econometrics through an influential book that summarizes concepts associated with the credibility revolution yielded significant impacts on demand for and responsiveness to causal evidence. After six months, treated individuals' perceived importance of quantitative analysis increased by 50%. Their performance in national

[19] It is a continuous variable with 0 denoting no similarity and 1 indicating a 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)

research methods and public policy exams improve by 0.5–0.8 sigma. Text analyses of their writings suggest an understanding of concepts associated with the credibility revolution. We also found that treated individuals' stated WTP for commissioning RCTs using public funding increased by 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 policymakers respond to causal information. Econometrics training focusing on causal thinking likely corrected belief updating in how policymakers react to causal evidence. Deputy ministers assigned to the placebo group sometimes updated their beliefs in the opposite direction of the signal or updated their beliefs past the signal. These findings suggest that training causal thinking not only increases responsiveness to causal evidence but may even correct for mistakes in the belief-updating process.

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 the credibility revolution. We *cautiously* interpret the training program as welfare improving in light of the body of evidence that finds ambiguous effect of ICT policies on learning outcomes (Cristia, Ibarraran, Cueto, Santiago, and Severin 2012; Kremer, Brannen, and Glennerster 2013; Mbiti 2016), and positive welfare effects of deworming policies (Miguel and Kremer, 2004; Croke et al., 2016; Ahuja et al. 2017). Future research can follow the long-term effects of metrics training on these deputy ministers and study the welfare consequences of training in the school of thought associated with the credibility revolution.

REFERENCES

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

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

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

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

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

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

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

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

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

Bertrand, M., Burgess, R., Chawla, A. and Xu, G., 2020. The glittering prizes: Career incentives and bureaucrat performance. The Review of Economic Studies, 87(2), pp.626-655.

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

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

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

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

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

Chen, Daniel L., Martin Schonger, and Chris Wickens. 2016. "oTree—An open-source platform for laboratory, online, and field experiments." *Journal of Behavioral and Experimental Finance* 9: 88-97.

Cole, Shawn, Thomas Sampson, and Bilal Zia. 2011. "Prices or knowledge? What drives demand for financial services in emerging markets?." *The journal of finance* 66 (6): 1933-1967.

Conte, Elisabetta, Ilaria Grazzani, and Alessandro Pepe. 2018. "Social cognition, language, and prosocial behaviors: a multitrait mixed-methods study in early childhood." *Early Education and Development* 29 (6): 814-830.

Cristia, Julian, Pablo Ibarrarán, Santiago Cueto, Ana Santiago, and Eugenio Severín. 2017. "Technology and child development: Evidence from the one laptop per child program." *American Economic Journal: Applied Economics* 9 (3): 295-320.

Croke, Kevin, Joan Hamory Hicks, Eric Hsu, Michael Kremer, and Edward Miguel. 2016. *Does mass deworming affect child nutrition? Meta-analysis, cost-effectiveness, and statistical power* (No. w22382). National Bureau of Economic Research,

David J. Deming. and Weidmann, Ben. 2021. *Team Players: How Social Skills Improve Team Performance*. Forthcoming Econometrica.

De Quidt, Jonathan, Johannes Haushofer, and Christopher Roth. 2018. "Measuring and bounding experimenter demand." *American Economic Review* 108 (11): 3266-3302.

Durkin, Kevin, and Gina Conti‐Ramsden. 2007. "Language, social behavior, and the quality of friendships in adolescents with and without a history of specific language impairment." *Child development* 78 (5): 1441-1457.

Frank, Robert H., Thomas Gilovich, and Dennis T. Regan. 1993. "Does studying economics inhibit cooperation?." *Journal of economic perspectives* 7 (2): 159-171.

Frey, Bruno S., and Stephan Meier. 2003. "Are political economists selfish and indoctrinated? Evidence from a natural experiment." *Economic Inquiry* 41 (3): 448-462.

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

Gabaix, Xavier. 2014. "A sparsity-based model of bounded rationality." *The Quarterly Journal of Economics* 129 (4): 1661-1710.

Grantham-McGregor, Sally M., Christine A. Powell, Susan P. Walker, and John H. Himes. 1991. "Nutritional supplementation, psychosocial stimulation, and mental development of stunted children: the Jamaican Study." *The Lancet* 338 (8758): 1-5.

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

Heß, Simon. 2017. "Randomization inference with Stata: A guide and software." *The Stata Journal* 17 (3): 630-651.

Heckman, James J., and Ganesh Karapakula. 2019. *The Perry Preschoolers at late midlife: A study in design-specific inference* (No. w25888). National Bureau of Economic Research.

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

Kremer, Michael, Conner Brannen, and Rachel Glennerster. 2013. "The challenge of education and learning in the developing world." *Science* 340 (6130): 297-300.

Hamory, Joan, Edward Miguel, Michael Walker, Michael Kremer, and Sarah Baird. 2021. "Twenty-year economic impacts of deworming." *Proceedings of the National Academy of Sciences* 118 (14).

Kremer, Michael, Gautam Rao, and Frank Schilbach. 2019. "Behavioral development economics." In *Handbook of Behavioral Economics: Applications and Foundations 1* (Vol. 2, pp. 345-458). North-Holland.

Kuhn, T. S. 1962. "The Structure of Scientific Revolution. 1st Edn Chicago." *IL: University of Chicago Press*.

LaLonde, Robert J. 1986. "Evaluating the econometric evaluations of training programs with experimental data." *The American economic review*: 604-620.

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

Lu, Wei, Elliott Ash, and Daniel L. Chen. 2021." Motivated Reasoning in the Field: Polarization of Precedent, Prose, and Policy in U.S. Circuit Courts, 1930-2013." *Mimeo*.

Lusardi, Annamaria, and Olivia S. Mitchell. 2014. "The economic importance of financial literacy: Theory and evidence." *Journal of economic literature* 52 (1): 5-44.

Lusardi, Annamaria, Pierre-Carl Michaud, and Olivia S. Mitchell. 2017. "Optimal financial knowledge and wealth inequality." *Journal of Political Economy* 125 (2): 431-477.

Malmendier, Ulrike, Stefan Nagel, and Zhen Yan. 2017. *The making of hawks and doves: Inflation experiences on the FOMC* (No. w23228). National Bureau of Economic Research.

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

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

Miguel, Edward, and Michael Kremer. 2004. "Worms: identifying impacts on education and health in the presence of treatment externalities." *Econometrica* 72 (1): 159-217.

Mehmood, Sultan, Shaheen Naseer, and Daniel L. Chen. 2021. "Enhancing Public Officials' Altruistic Behavior through Training Programs." Mimeo.

Mehmood, S., Naseer, S. and Chen, D., 2022. AI as State Capacity. Submitted.

Metzger, L., Svoronos, T. and Khan, A., 2020. Policy decisions and evidence use among civil servants: A group decision experiment in Pakistan. CID Working Paper Series.

Muennig, Peter, Dylan Robertson, Gretchen Johnson. 2011. "The effect of an early education program on adult health: the Carolina Abecedarian Project randomized controlled trial." *American journal of public health* 101 (3): 512-516.

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

Pearl, J. and Mackenzie, D., 2018. The book of why: the new science of cause and effect. Basic books.

Rubinstein, Ariel. 2006. "Dilemmas of an economic theorist." *Revista de Economía Institucional* 8 (14): 191-213.

Shapin, Steven. 1982. "History of science and its sociological reconstructions." *History of science* 20 (3): 157-211.

Sidorov, Grigori, Alexander Gelbukh, Helena Gómez-Adorno, and David Pinto. 2014. "Soft similarity and soft cosine measure: Similarity of features in vector space model." *Computación y Sistemas* 18 (3): 491-504.

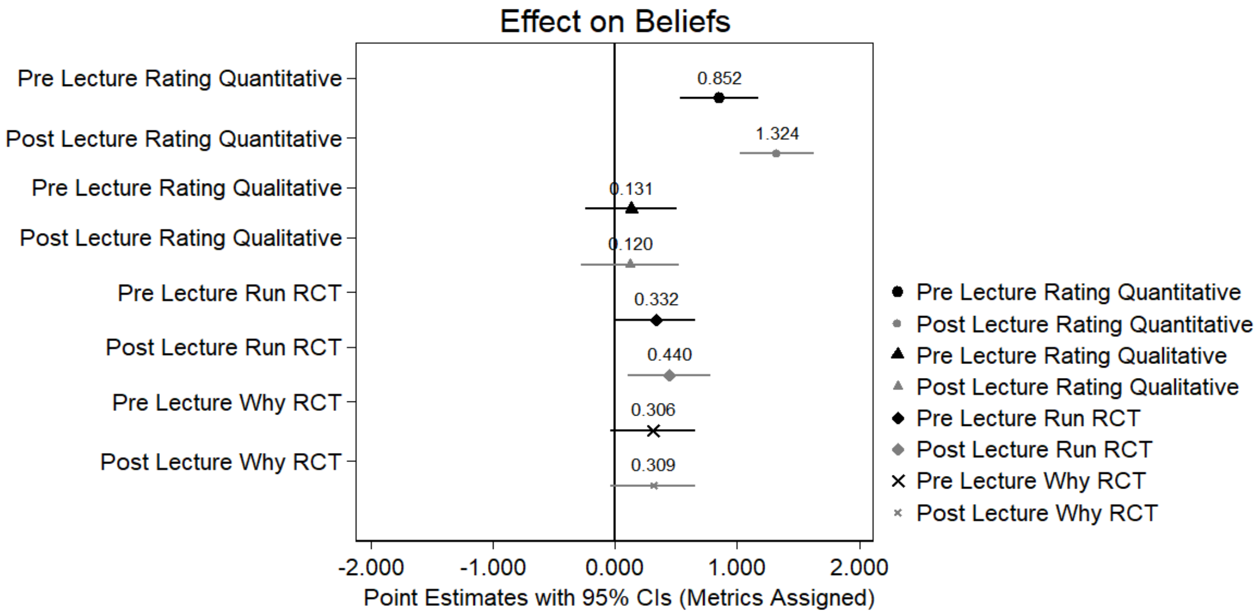
Siegel, Daniel J. 2010. *Mindsight: The new science of personal transformation*. Bantam.

Sutter, Matthias, Michael Weyland, Anna Untertrifaller, and Manuel Froitzheim. 2020. "Financial literacy, risk and time preferences–Results from a randomized educational intervention." *MPI Collective Goods Discussion Paper* (2020/17).

Vivalt, Eva, and Aidan Coville. 2021. "How Do Policymakers Update?" *Mimeo*.

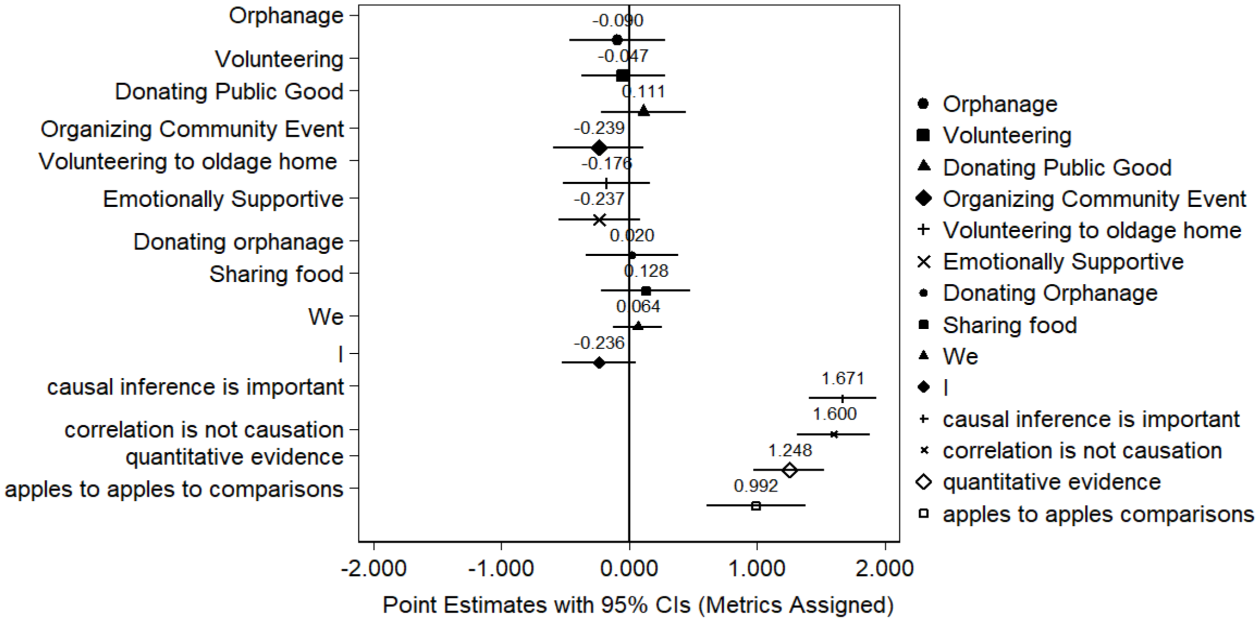
Figures and Tables

Figure 1: Impact of Metrics Training on Beliefs – Standardized



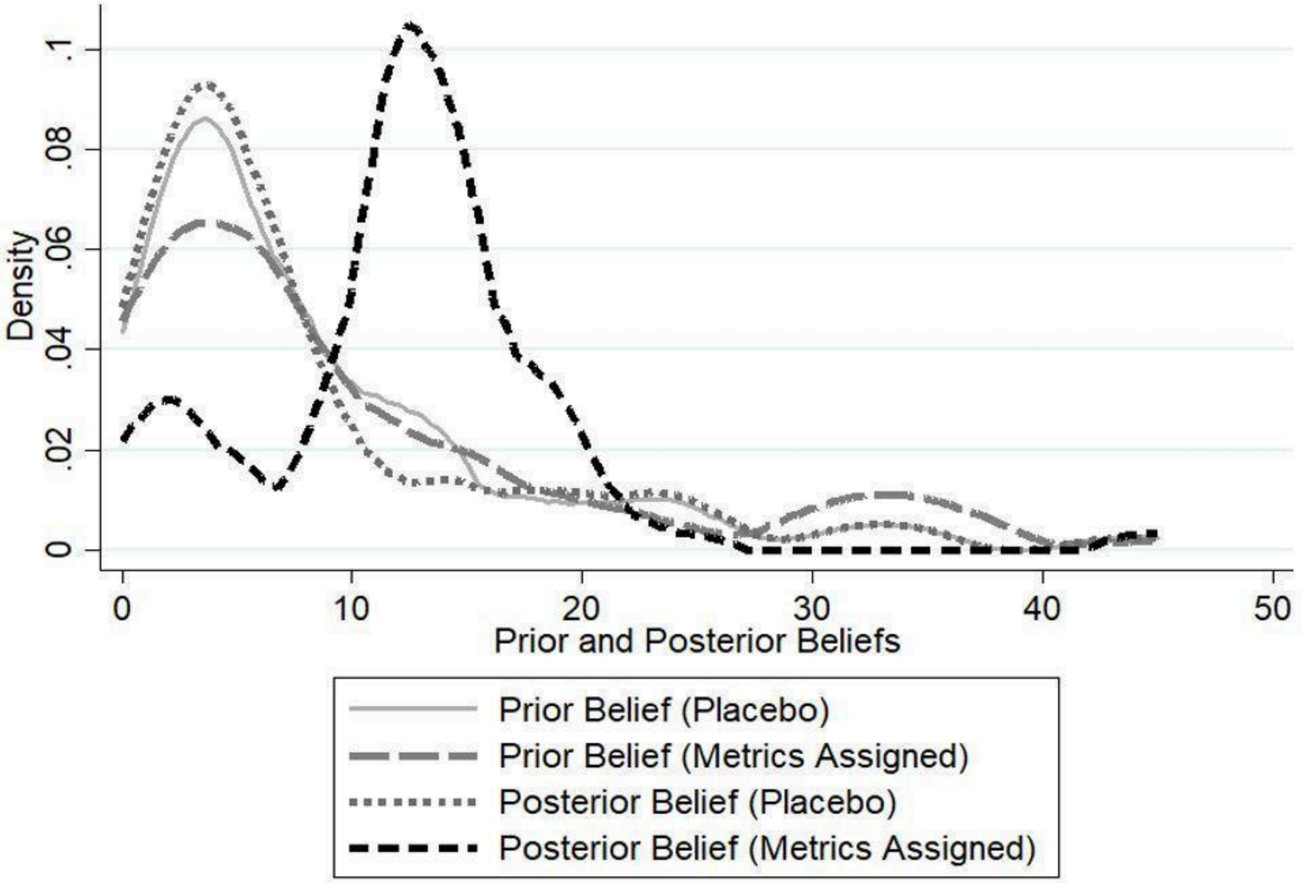
Note: The figure above estimates our main specification including choice of book and individual level controls with all dependent variables standardized to mean zero and standard deviation one. Point estimate and 95% confidence interval on the randomly assigned metrics training is presented for each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



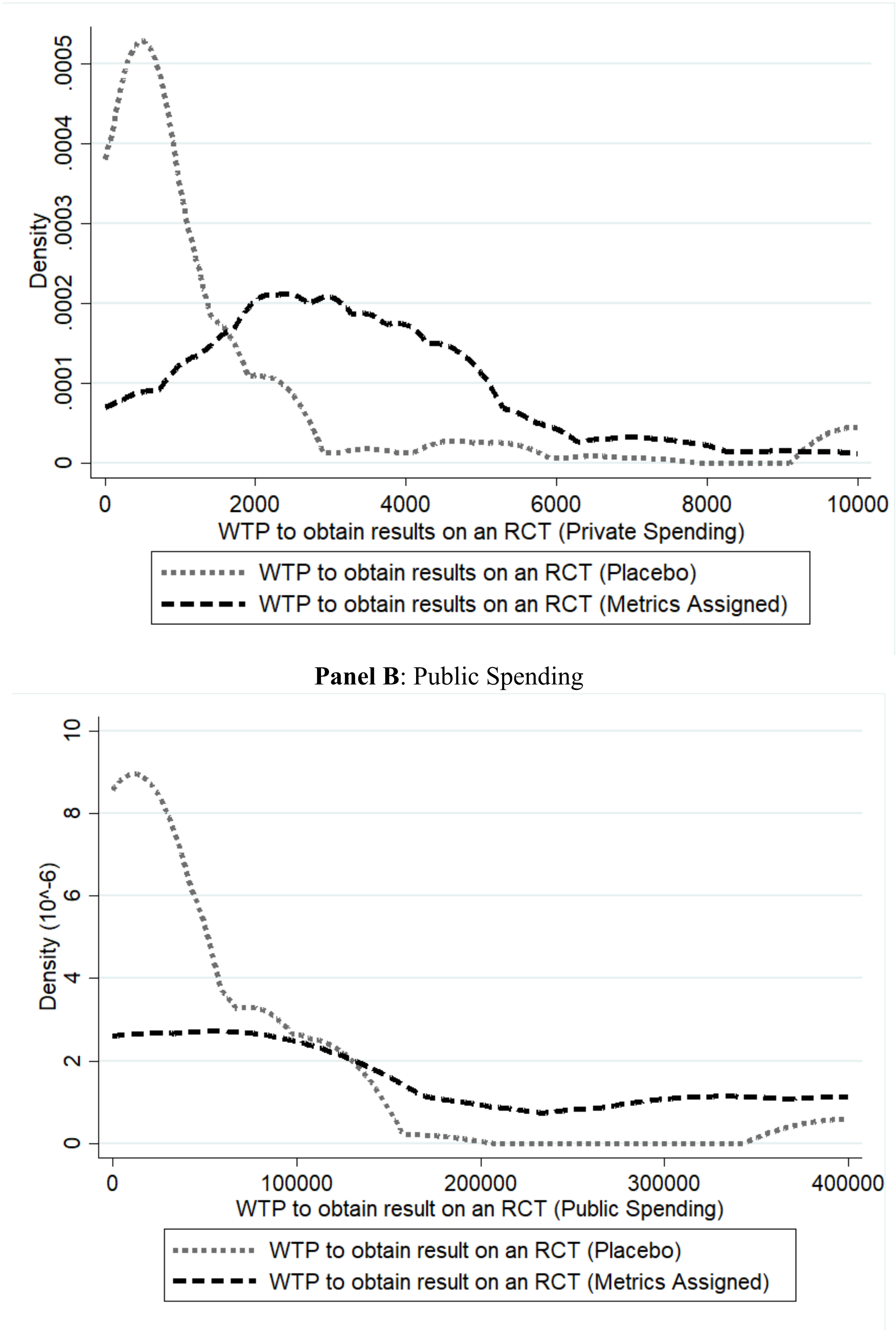
Note: The figure above estimates our main specification including choice of book and individual level controls on different dependent variables standardized to mean zero and standard deviation one. Point estimate and 95% confidence interval on the randomly assigned metrics training is presented for each 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 2: **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**

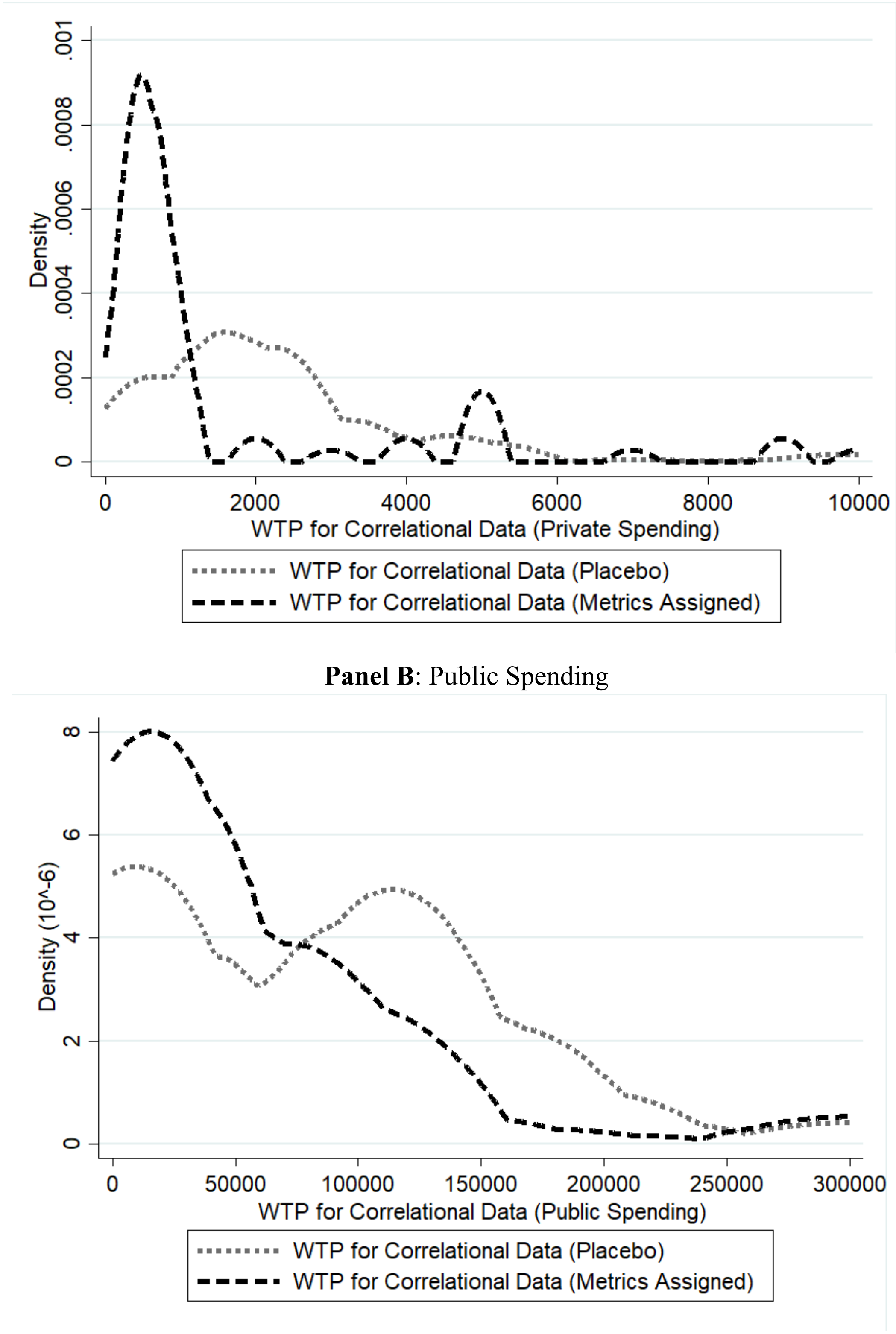


Panel A: Private Spending

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**



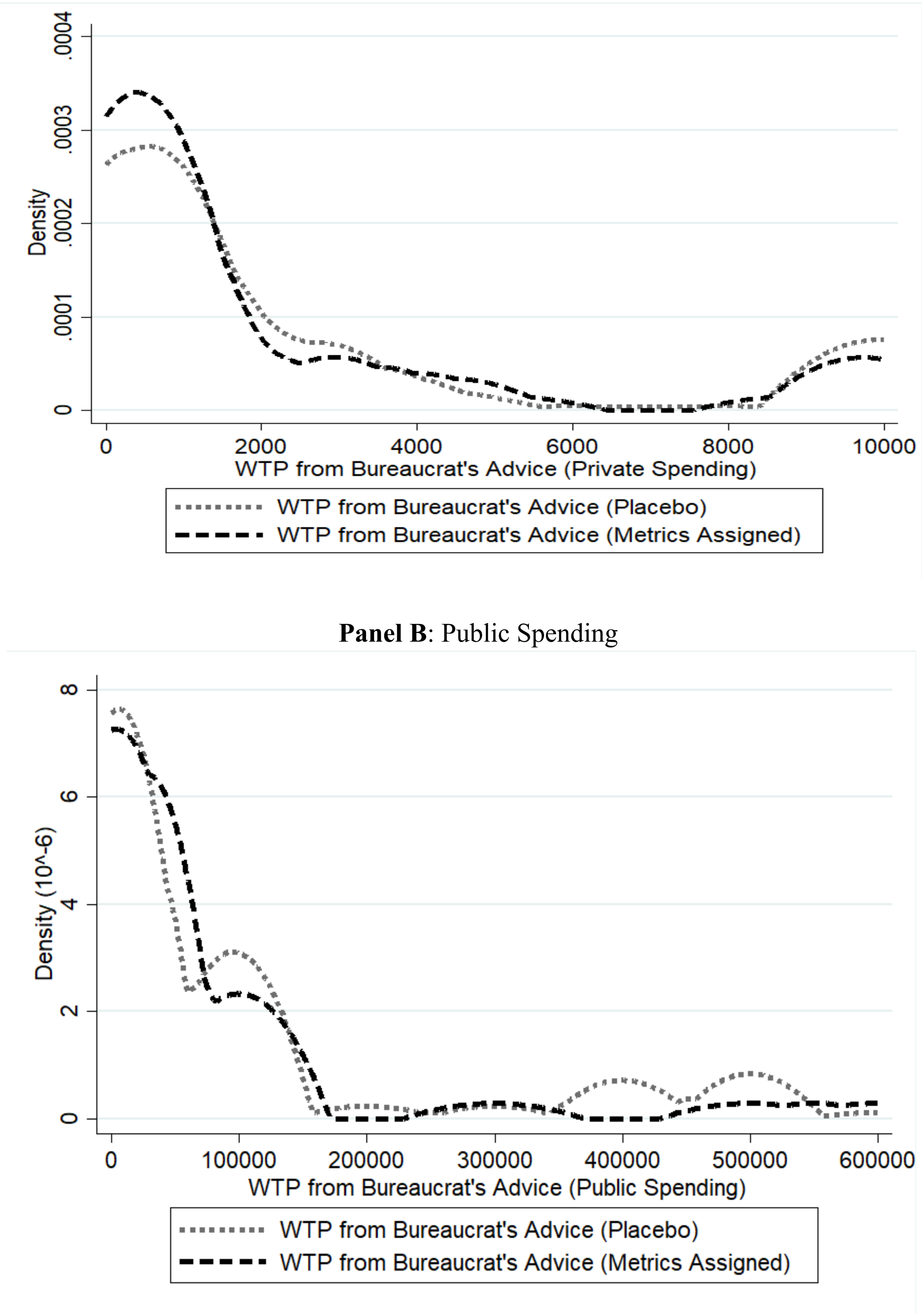
Panel A: Private Spending

Panel B: Public Spending

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

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

Panel A: Private Spending

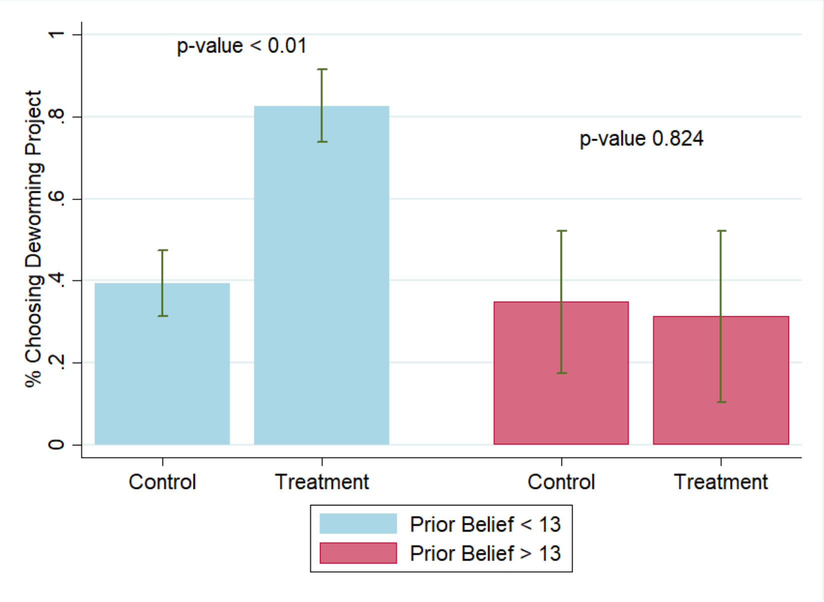


Panel B: Public Spending

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.

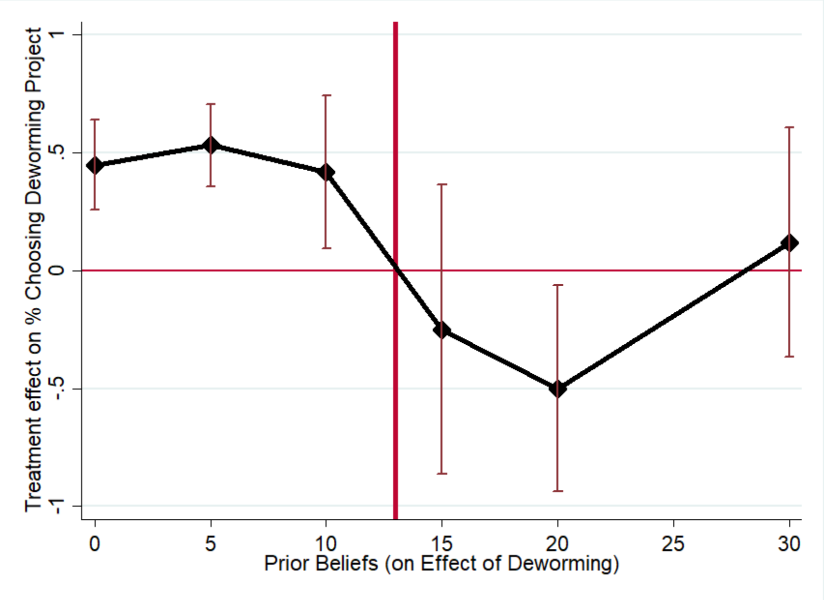
Figure 7: Effect of Metrics Training on Deworming Project Choice by Initial Beliefs

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: Deputy Minister - 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
Mean of dep. variable	0.324	34258.26	26.775	0.516	0.225	0.169	0.099	0.610	655.585	131.085	7.221

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

Table 2: The Effect of Metrics Training on Beliefs – Original Units

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

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

Table 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 3: Effect of Metrics Training on Policy

	Deworming Policy		***Orphanage Renovation Policy***		***School Renovation Policy***	
	Letter Sent	*Funds Recommended*	*Letter Sent*	*Funds Recommended*	*Letter Sent*	*Funds Recommended*
	(1)	(2)	(3)	(4)	(5)	(6)
Metrics Assigned	0.290***	401,888***	0.011	18,254	-0.053	-10,042
	(0.083)	(109,081)	(0.062)	(22,179)	(0.078)	(15,197)
Individual Controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	190	190	190	190	190	190
R-squared	0.164	0.206	0.120	0.103	0.089	0.100
Mean of dep. var. (placebo)	0.174	171812.1	0.174	51073.83	0.262	41744.97

Robust standard errors appear in brackets (clustered at the individual level). The dependent variables are letters sent and funds recommended to government division in Pakistan (in Pakistani Rupees) for budget allocation for Deworming, Orphanage and School renovations, respectively. Metrics Assigned is a dummy variable that switches on when a causal inference book is randomly assigned to participants. Consistent with all our earlier regressions, we always control for the metrics chosen. The causal inference book, Mastering Metrics, is randomly assigned conditional on the book being chosen. The estimations obtained from OLS regressions include the following controls: metrics chosen, written test scores, interview test scores, gender, birth in political capitals, asset ownership, income before joining public service, age, education, foreign visits and occupational group dummies. *** p<0.01, ** p<0.05, * p<0.1.

Table 4: Effect of Metrics Training on Willingness to Pay

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

Robust standard errors clustered at individual level appear in brackets. Dependent variables are the private and public stated willingness to pay for a randomized control trial, correlational comparison of means data, and suggestion from expert bureaucrat. 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***	
	Amount Randomized Trial	Amount Randomized Trial	Amount Randomized Trial	Amount Randomized Trial
	(1)	(2)	(3)	(4)
Defiers X Metrics Assigned	-888.410		-2531651*	
	(1301.347)		(1367816)	
Never Takers X Metrics Assigned		-1987.019		-1304302
		(1837.863)		(1558128)
Metrics Assigned	2179.581***	2165.335***	1775415***	15028283**
	(648.553)	(641.327)	(635063.5)	(702559.8)
Defiers X Metrics Chosen	-428.640		3326010	
	(1207.932)		(2437943)	
Never Takers X Metrics Chosen		-643.626		2190478*
		(1385.195)		(1303170)
Defiers	-463.802		-3101800	
	(857.643)		(2201009)	
Never Takers		1756.372		-1250269
		(1887.535)		(1201626)
p-values:	(1) = (2):	{0.590}	(3) = (4):	{0.532}
Individual Controls	Yes	Yes	Yes	Yes
Observations	180	180	180	180
R-squared	0.224	0.223	0.114	0.098
Mean of dep. var. (placebo)	1539.453	1539.453	928546.1	928546.1

Robust standard errors clustered at individual level appear in 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 chose to not to click any link. p-values testing equality of coefficients of Metrics Assigned X Defiers = Metrics Assigned X Never Takers are presented in curly brackets. The estimations obtained from OLS regressions include the following controls: metrics chosen, written test scores, interview test scores, gender, birth in political capitals, asset ownership, income before joining public service, age, education, foreign visits and occupational group dummies. All dependent variables are denominated in Pakistani Rupees *** p<0.01, ** p<0.05, * p<0.1.

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 Training Academy

Dear Officers,

With this email, we wanted to send you a mandatory assignment from the Civil Service 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 CSA ID and Book assigned to you on top of the document.

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

Best of Luck,

Civil Service Academy Director

Table A2: Structured Discussion Post-Lecture

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

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

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

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

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

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

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

[Enter Number]

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

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

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

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

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

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

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

[Enter Number PKR]

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

[Enter Number PKR]

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

[Enter Number PKR]

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

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

[Enter Number PKR]

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

[Enter Number PKR]

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

[Enter Number PKR]

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

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

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

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

[Enter Number]

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

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

[Enter Number PKR]

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

[Enter Number PKR]

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

[Enter Number PKR]

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

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

[Enter Number PKR]

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

[Enter Number PKR]

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

[Enter Number PKR]

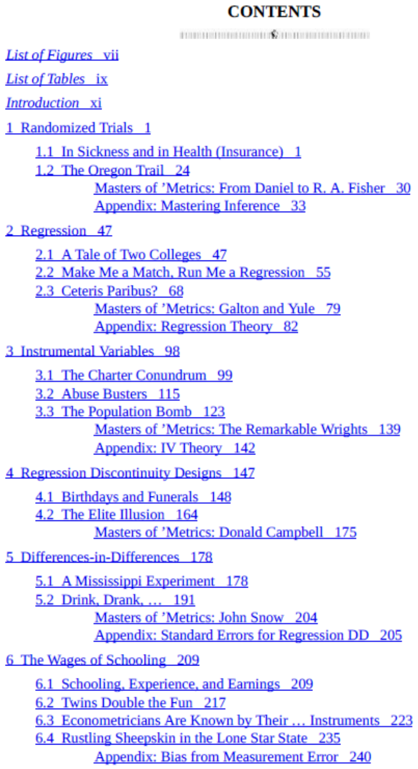
Question 8) Now, what project would you choose 1 or 2? One being deworming and two being computer lab.

[Enter Number]

B. Additional Figures and Tables

Figure B1: Contents of Treatment and Placebo Books

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



CONTENTS

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

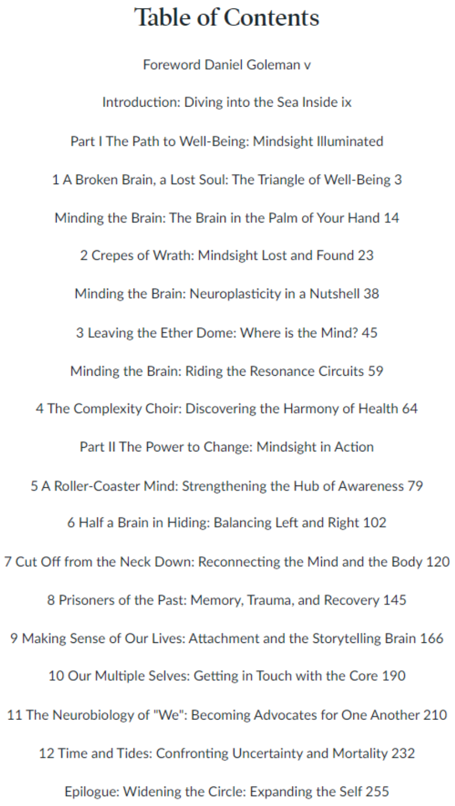
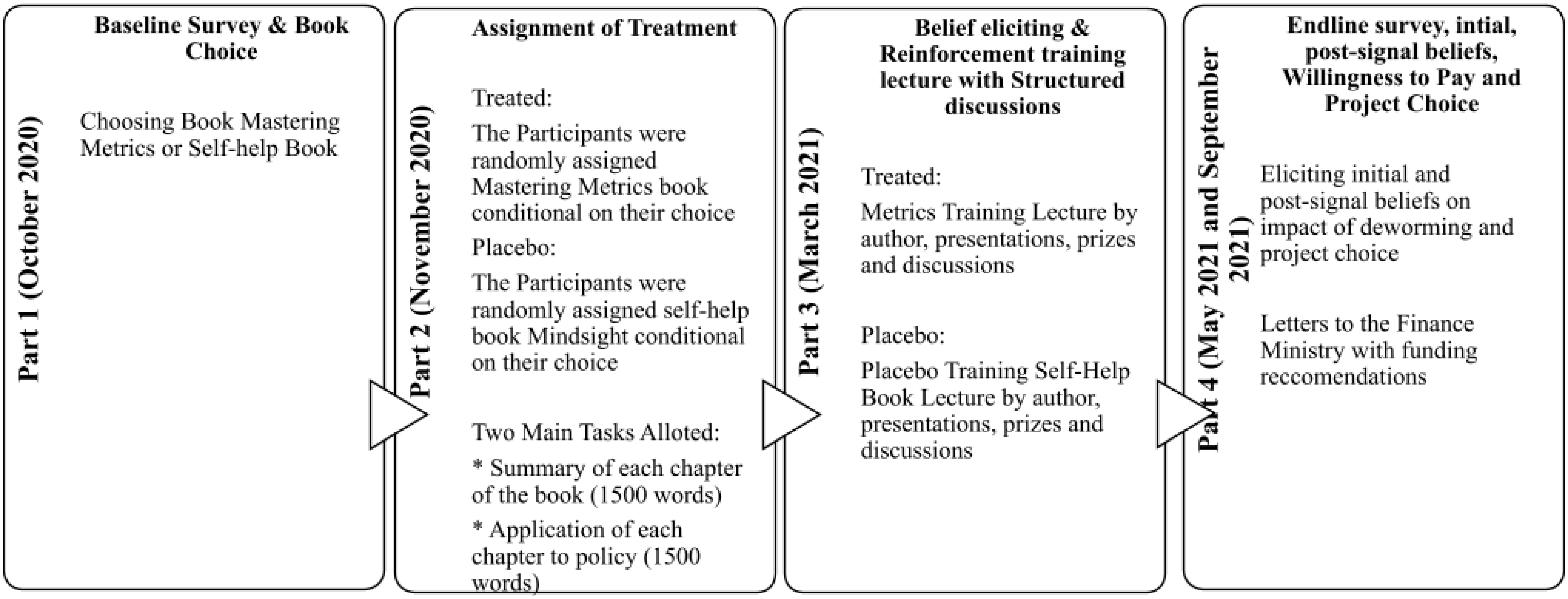


Table of Contents

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 and September 2021)

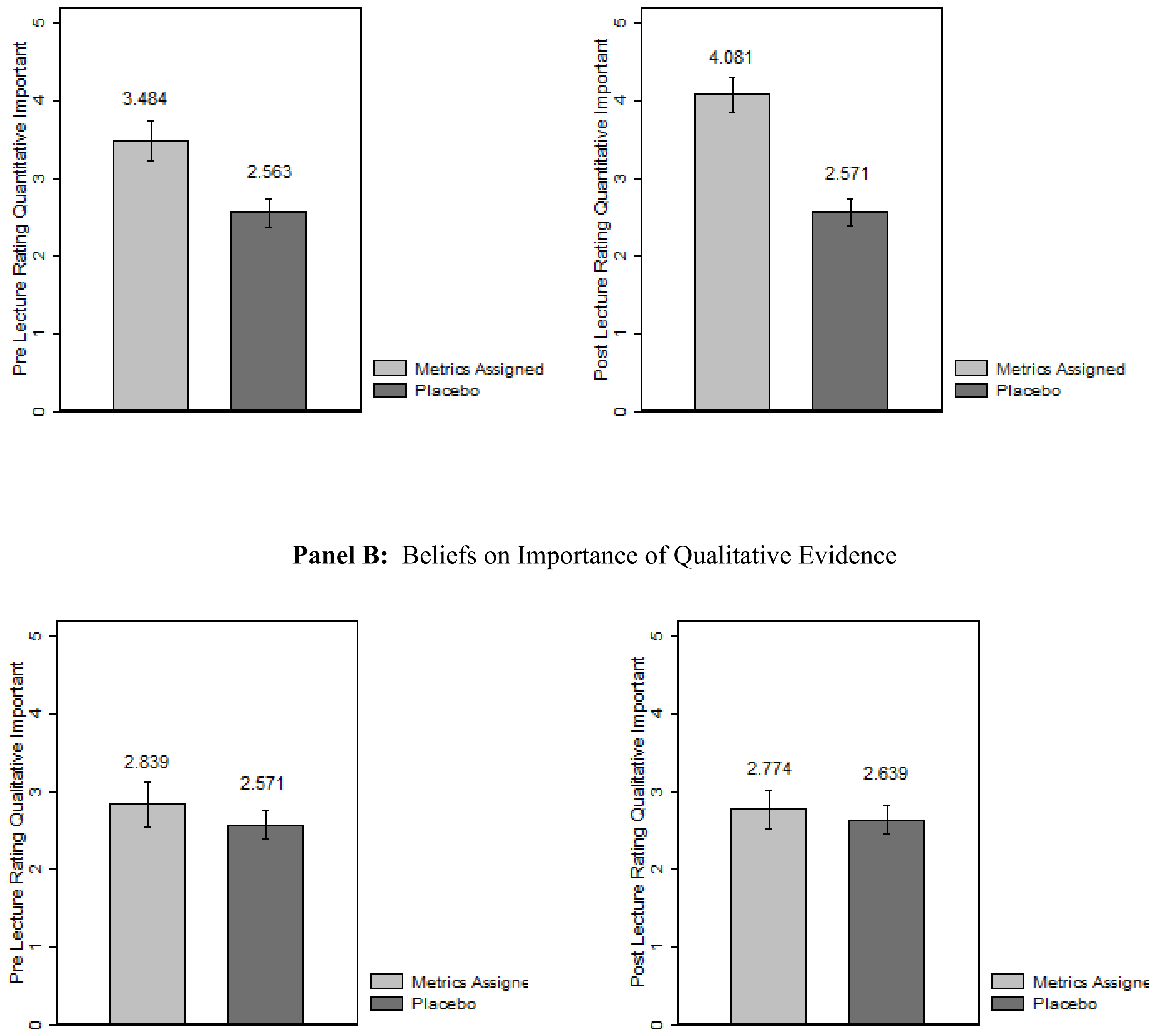
Endline survey, intial, post-signal beliefs, Willingness to Pay and Project Choice

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

Letters to the Finance Ministry with funding reccomendations

Figure B3: Impact of Metrics Training

Panel A: Beliefs on Importance of Quantitative Evidence

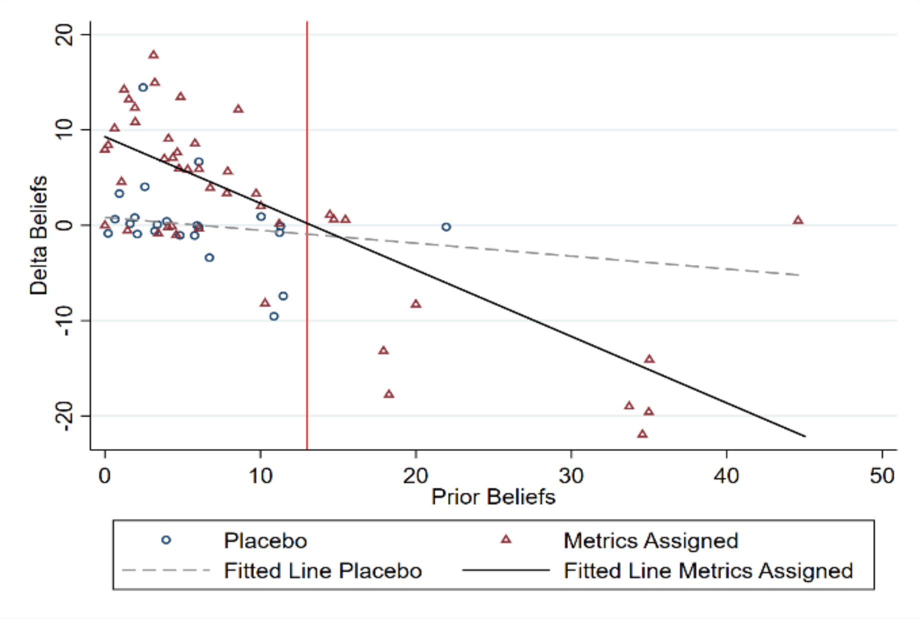


Panel B: Beliefs on Importance of Qualitative Evidence

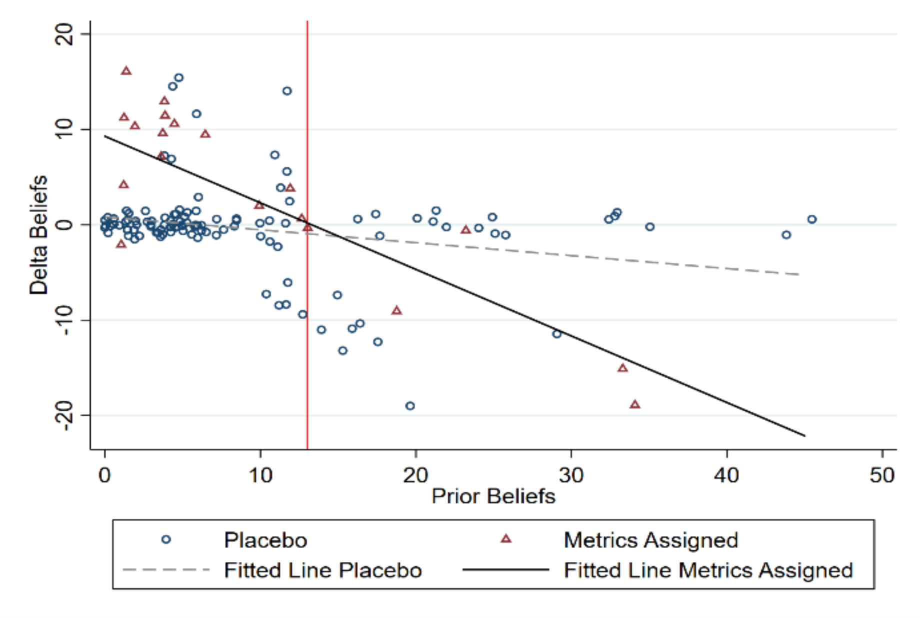
Note: The figures above compare beliefs on importance of quantitative (Panel A) and qualitative (Panel B) evidence for those before and after lecture and discussion for treated and control group. The left panels are those treated with metrics book and corresponding assignments, while figures on the right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

Figure B4: Shifts in Beliefs versus Initial Beliefs by Metrics Selection

Panel A: Metrics Selected

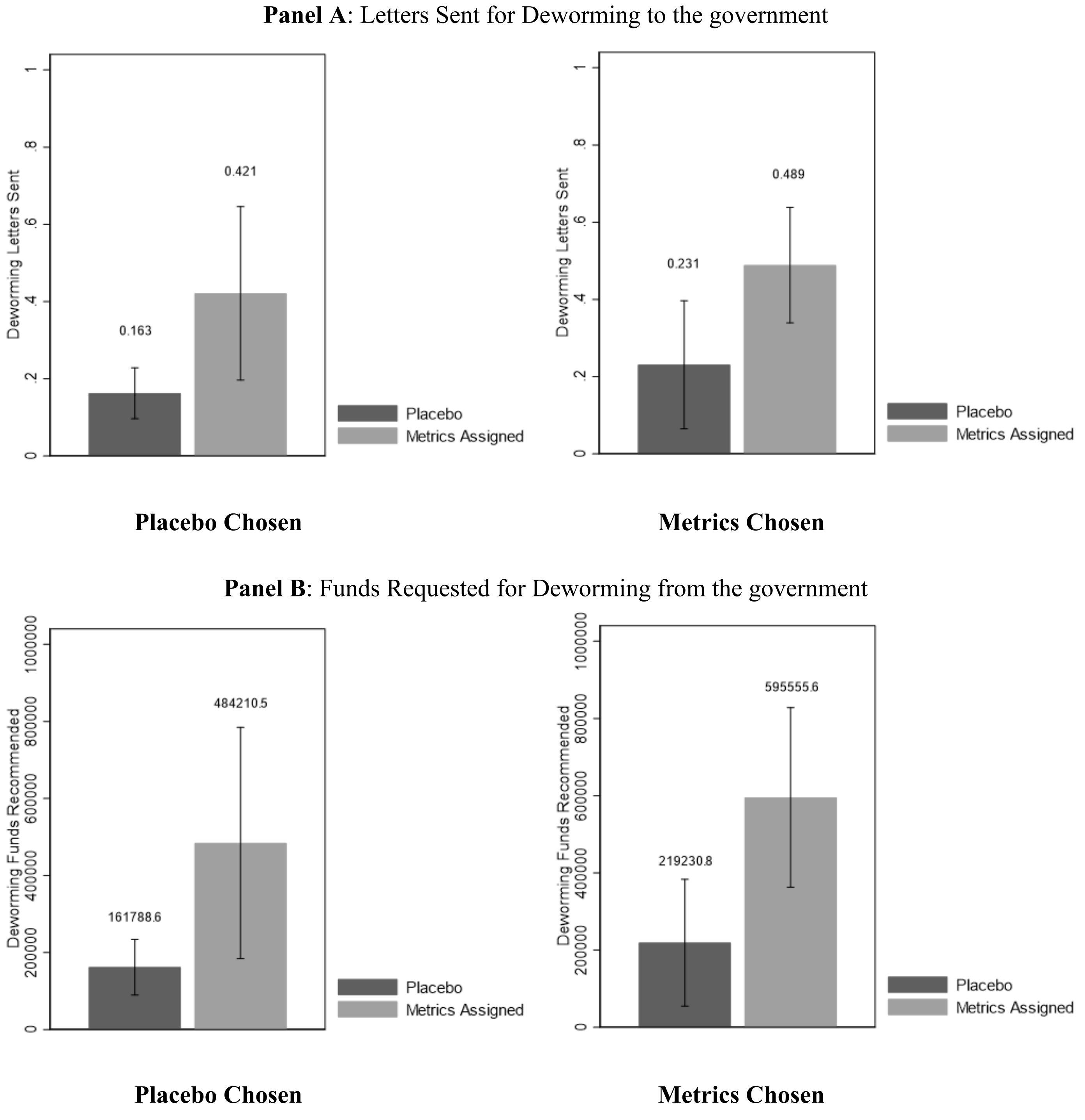


Panel B: Placebo Selected



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

Figure B5: Impact of Metrics Training on Actual Policy by Ex Ante Demand for Placebo and Metrics Book



Panel A: Letters Sent for Deworming to the government

Placebo Chosen

Metrics Chosen

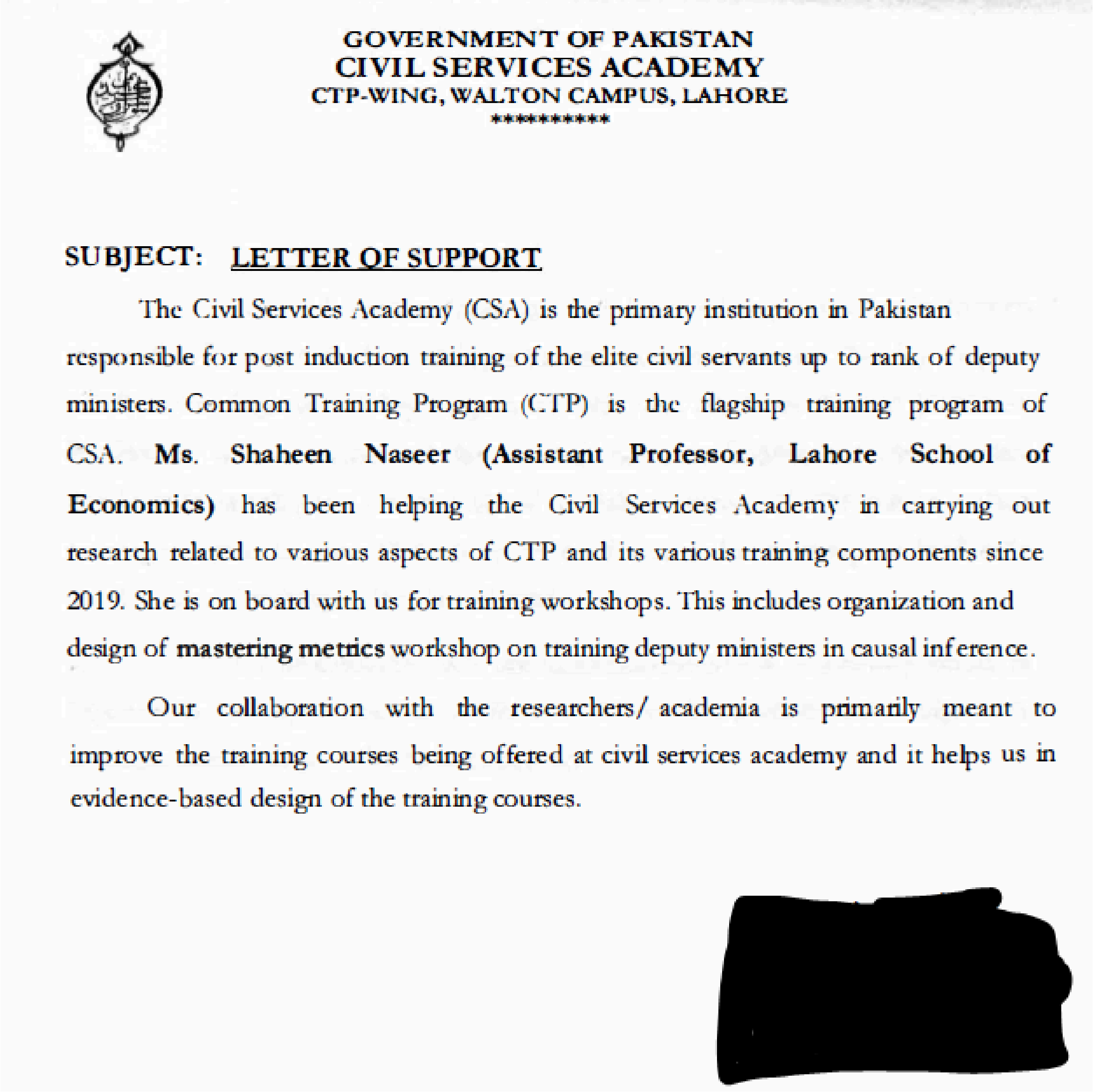
Panel B: Funds Requested for Deworming from the government

Placebo Chosen

Metrics Chosen

Note: The figure above plots bar charts for individuals assigned the metrics training and those assigned the placebo training. Panel A provides the bar charts for letters sent for deworming to the government, while Panel B provides the bar charts for ministers funding requests for deworming in Pakistan Rupees (PKR). The bar charts on the left represent ministers who selected the placebo training, while bar charts on the right represent those who selected the metrics training. 95% Confidence Intervals are also reported.

Table B1: Letter of Support from Director of Training Academy



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

SUBJECT: <u>LETTER OF SUPPORT</u>

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

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

[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: Effects of Metrics Training on Deworming Policy by Demand for Metrics

	Mastering Metrics Demanded		Placebo Demanded	
	Letter Sent	*Funds Requested*	*Letter Sent*	*Funds Requested*
	(1)	(2)	(3)	(4)
Metrics Assigned	0.221	352102*	0.324**	413664.9**
	(0.135)	(179872)	(0.127)	(160788.2)
Individual Controls	Yes	Yes	Yes	Yes
Observations	70	70	120	120
R-squared	0.238	0.315	0.128	0.178

Robust standard errors appear in brackets (clustered at the individual level). The dependent variables are letters sent and funds recommended to Pakistan's government (in Pakistani Rupees) for budget allocation for Deworming. Metrics Assigned is a dummy variable that switches on when a causal inference book is randomly assigned to participants. Consistent with all our earlier regressions, we always control for metrics chosen. The causal inference book, Mastering Metrics, is randomly assigned conditional on the book being chosen, hence we always control for metrics selected. The estimations obtained from OLS regressions include the following controls: metrics chosen, written test scores, interview test scores, gender, birth in political capitals, asset ownership, income before joining public service, age, education, foreign visits and occupational group dummies. *** p<0.01, ** p<0.05, * p<0.1.

Table B5: Impact of Treatment by Metrics Chosen

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

Robust standard errors clustered at individual level appear in 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, a dependent variable is constructed based on the statement "You are in charge of assigning people to a public policy program and before rolling it out, you want to learn if the policy is effective, what you would do?" These and other dependent variables are identical to those reported and explained in the accompanied notes of Table 2-4. Metrics assigned is a dummy variable that switches on when a causal inference book is randomly assigned to participants. The metrics book is randomly assigned conditional on it being chosen. The estimations obtained from OLS regressions include the following controls: metrics chosen, written test scores, interview test scores, gender, birth in political capitals, asset ownership, income before joining public service, age, education, foreign visits and occupational group dummies. *** p<0.01, ** p<0.05, * p<0.1

Table B6: Heterogeneity by Pretreatment Quantitative Assessment Scores

	Rating Quantitative		*Run RCT*		*Amount Randomized Trial (Private Spending)*		*Amount Randomized Trial (Public Spending)*		*Funds Requested for Deworming (in Pakistan Rupees)*	
Quantitative Scores	*Below Median*	*Above Median*	*Below Median*	*Above Median*	*Below Median*	*Above Median*	*Below Median*	*Above Median*	*Below Median*	*Above Median*
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Metrics Assigned	1.865***	0.942***	0.289**	0.259	447.309	843.157	-511279.2	-284572.3	393423.4***	270870.6**
	(0.244)	(0.282)	(0.145)	(0.119)	(1080.277)	(642.236)	(1312014)	(453235.8)	(142329)	(137300.9)
Metrics Assigned Equal (p-value):	(1) = (2):	[0.0135]	(3) = (4):	[0.8722]	(5) = (6):	[0.7528]	(7) = (8):	[0.8703]	(7) = (8):	[0.5355]
Individual Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	79	80	79	80	79	80	79	80	79	80
R-squared	0.469	0.594	0.246	0.168	0.310	0.257	0.167	0.285	0.222	0.324

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 the 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 include the following controls: metrics chosen, written test scores, interview test scores, gender, birth in political capitals, asset ownership, income before joining public service, age, education, foreign visits and occupational group dummies. *** p<0.01, ** p<0.05, * p<0.1.

Table B7: The Effect of Metrics Training on Beliefs – Randomization Inference

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

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

Table B8: Impact of Metrics Training on Policy Making Course Grades – Randomization Inference

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

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

Table B10: Effect of Metrics Training on Willingness to Pay – Randomization Inference

Panel A: Private Spending

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

Panel B: Public Spending

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

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

Table B11: 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.

Table B12: Robustness to Multiple Hypothesis Testing

	Lecture Rating Quantitative	*Deworming Letter*	*Deworming Funds*	*Amount Randomization Trial*	*Amount Correlational Data*	*Assessment Public Policy*	*Assessment Research Methods*
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Metrics assigned	1.541	0.290	401,888	1391308	-35,274	0.505	0.799
p-value	(0.001) ***	(0.001) ***	(0.001) ***	(0.038) **	(0.007) ***	(0.003) ***	(0.001) ***
Sharpened q-value	[0.002] ***	[0.002] ***	[0.002] ***	[0.009] ***	[0.004] ***	[0.002] ***	[0.002] ***
Observations	190	190	190	180	180	190	190
R-squared	0.425	0.164	0.206	0.094	0.148	0.071	0.218
Mean of dep. var. (Placebo)	2.745	0.174	171812.1	928546.1	94136.720	0.171	0.219

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

Appendix C: Main Results by Metrics Chosen and Placebo Chosen Books

Appendix C Tables

Table C1.1: Deputy Minister - Balance of Treatment on Individual Characteristics (with Metrics Chosen = 0)

	(1) Birth Political Capitals	(2) Income	(3) Age	(4) Education	(5) Visited Foreign Country	(6) PAS	(7) PSP	(8) Other groups	(9) Pre Treat. Written Asses.	(10) Pre Treat. Interv. Asses.	(11) Pre Treat. Math Asses.
Metrics Assigned	-0.0641	-3,628	0.0140	0.0709	-0.0332	-0.0203	-0.0785	0.00260	-0.290	-0.0348	-0.142
	[0.119]	[6,319]	[0.571]	[0.118]	[0.106]	[0.0648]	[0.0558]	[0.0973]	[6.654]	[3.604]	[0.238]
Observations	109	109	109	109	109	109	109	109	109	109	109
Mean of dep. variable	0.324	34258.26	26.775	0.516	0.225	0.169	0.099	0.610	655.585	131.085	7.221
R-squared	0.130	0.215	0.289	0.316	0.155	0.587	0.513	0.606	0.414	0.254	0.137

Robust standard errors appear in brackets (clustered at the individual level). Metrics assigned is a dummy variable that switches on when a causal inference book is randomly assigned to participants. The causal inference book is randomly assigned conditional on the book being chosen. The controls include Metrics Chosen (a dummy variable that switches on when causal inference book is chosen by the participants), and all other available individual characteristics obtained from administrative data (i.e. all remaining column dependent variable except the dependent variable used in the respective column). In this regression, the Metrics Chosen parameter was set to 0. *** p<0.01, ** p<0.05, * p<0.1.

Table C1.2: Deputy Minister - Balance of Treatment on Individual Characteristics (with Metrics Chosen = 1)

	(1) Birth Political Capitals	(2) Income	(3) Age	(4) Education	(5) Visited Foreign Country	(6) PAS	(7) PSP	(8) Other groups	(9) Pre Treat. Written Asses.	(10) Pre Treat. Interv. Asses.	(11) Pre Treat. Math Asses.
Metrics Assigned	0.207	-13,534	0.345	0.163	-0.0641	0.00252	-0.0245	0.0263	-3.259	5.303	0.610
	[0.165]	[8,325]	[0.582]	[0.158]	[0.108]	[0.0551]	[0.0476]	[0.0567]	[9.423]	[6.287]	[0.427]
Observations	75	75	75	75	75	75	75	75	75	75	75
Mean of dep. variable	0.324	34258.26	26.775	0.516	0.225	0.169	0.099	0.610	655.585	131.085	7.221
R-squared	0.195	0.241	0.351	0.300	0.259	0.717	0.560	0.766	0.608	0.303	0.098

Robust standard errors appear in brackets (clustered at the individual level). Metrics assigned is a dummy variable that switches on when a causal inference book is randomly assigned to participants. The causal inference book is randomly assigned conditional on the book being chosen. The controls include Metrics Chosen (a dummy variable that switches on when the participants choose causal inference book), and all other available individual characteristics obtained from administrative data (i.e. all remaining column dependent variable except the dependent variable used in the respective column). In this regression, the Metrics Chosen parameter was set to 1. *** p<0.01, ** p<0.05, * p<0.1.

Table C2.1: The Effect of Metrics Training on Beliefs – Original Units (with Metrics Chosen = 0)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Pre-Lecture Rating Quantitative	Post-Lecture Rating Quantitative	Pre-Lecture Rating Qualitative	Post-Lecture Rating Qualitative	Pre-Lecture Run RCT	Post-Lecture Run RCT	Pre-Lecture Why Run RCT	Post-Lecture Why Run RCT
Metrics Assigned	0.817***	1.432***	-0.0422	-0.109	0.163	0.151	0.0423	0.0404
	[0.263]	[0.254]	[0.224]	[0.297]	[0.108]	[0.122]	[0.132]	[0.129]
Observations	109	109	109	109	110	110	108	108
Mean of dep. variable	2.745	2.979	2.490	2.596	0.362	0.404	0.396	0.396
R-squared	0.313	0.489	0.242	0.114	0.260	0.165	0.132	0.149

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 on the statement “How important do you think quantitative analysis is in public policy making?” In Columns 3 and 4 the dependent variable is a rating on “How important do you think qualitative analysis is in public policy making? “While in Columns 5 and 6 dependent variable is constructed from “You are in charge of assigning people to a public policy program and before rolling it out, you want to learn if the policy is effective, what you would do?” One of the options is to “Run a randomized control trial” while the other options are “Survey feelings of respondents regarding the policy”, and “Survey if there is demand for policy” and option 4 is to “compare two groups of people who had previously benefited most from the policy with those that did not?” The dependent variable takes the value of 1 if the run a randomized control trial option is chosen and zero otherwise. In Columns 7 and 8 the dependent variable is constructed based on the question: “Continuing with previous example, why the previous answer makes sense?”: (1) Because people in a RCT are apples to apples comparisons (2) People feelings are important determinant whether the public policy will work (3) Survey methods are known to produce causal effects (4) Comparing two groups of non-randomly selected people allows us to infer causality. The dependent variable takes the value of one if option 1 is chosen and 0 otherwise. Beliefs or ratings are measured before and after the lecture. Metrics assigned is a dummy that switches on when a causal inference book is randomly assigned to participants. Consistent with all our earlier regressions, we always control for metrics chosen. 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. In this regression, the Metrics Chosen parameter was set to 0. *** p<0.01, ** p<0.05, * p<0.1

Table C2.2: The Effect of Metrics Training on Beliefs – Original Units (with Metrics Chosen = 1)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Pre-Lecture Rating Quantitative	Post-Lecture Rating Quantitative	Pre-Lecture Rating Qualitative	Post-Lecture Rating Qualitative	Pre-Lecture Run RCT	Post-Lecture Run RCT	Pre-Lecture Why Run RCT	Post-Lecture Why Run RCT
Metrics Assigned	0.943***	1.608***	0.354	0.333	0.0809	0.191	0.161	0.161
	[0.245]	[0.248]	[0.318]	[0.289]	[0.144]	[0.139]	[0.135]	[0.135]
Observations	75	75	75	75	74	74	76	76
Mean of dep. variable	2.745	2.979	2.490	2.596	0.362	0.404	0.396	0.396
R-squared	0.322	0.519	0.155	0.279	0.192	0.178	0.216	0.216

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 on the statement “How important do you think quantitative analysis is in public policy making?” In Columns 3 and 4 the dependent variable is a rating on “How important do you think qualitative analysis is in public policy making? “While in Columns 5 and 6 dependent variable is constructed from “You are in charge of assigning people to a public policy program and before rolling it out, you want to learn if the policy is effective, what you would do?” One of the options is to “Run a randomized control trial” while the other options are “Survey feelings of respondents regarding the policy”, and “Survey if there is demand for policy” and option 4 is to “compare two groups of people who had previously benefited most from the policy with those that did not?” The dependent variable takes the value of 1 if the run a randomized control trial option is chosen and zero otherwise. In Columns 7 and 8 the dependent variable is constructed based on the question: “Continuing with previous example, why the previous answer makes sense?”: (1) Because people in a RCT are apples to apples comparisons (2) People feelings are important determinant whether the public policy will work (3) Survey methods are known to produce causal effects (4) Comparing two groups of non-randomly selected people allows us to infer causality. The dependent variable takes the value of one if option 1 is chosen and 0 otherwise. Beliefs or ratings are measured before and after the lecture. Metrics assigned is a dummy that switches on when a causal inference book is randomly assigned to participants. Consistent with all our earlier regressions, we always control for metrics chosen. 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. In this regression, the Metrics Chosen parameter was set to 1. *** p<0.01, ** p<0.05, * p<0.1

Table C3.1: Impact on Policymaking Assessments – Standardized (with Metrics Chosen = 0)

	(1)	(2)	(3)	(4)	(5)	(6)
	Assessment Public Policy	Assessment Public Policy	Assessment Research Methods	Assessment Research Methods	Assessment Teamwork	Assessment Teamwork
Metrics Assigned	0.282	0.122	0.636**	0.510*	0.193	0.160
	[0.227]	[0.249]	[0.244]	[0.260]	[0.264]	[0.290]
Observations	115	115	115	115	115	115
R-squared	0.011	0.124	0.080	0.168	0.005	0.171

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. In this regression, the Metrics Chosen parameter was set to 0.*** p<0.01, ** p<0.05, * p<0.1.

Table C3.2: Impact on Policymaking Assessments – Standardized (with Metrics Chosen = 1)

	(1)	(2)	(3)	(4)	(5)	(6)
	Assessment Public Policy	Assessment Public Policy	Assessment Research Methods	Assessment Research Methods	Assessment Teamwork	Assessment Teamwork
Metrics Assigned	0.823***	0.729***	0.965***	1.063***	-0.173	-0.211
	[0.230]	[0.237]	[0.242]	[0.252]	[0.265]	[0.278]
Observations	75	75	75	75	75	75
R-squared	0.145	0.297	0.167	0.310	0.007	0.162

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. In this regression, the Metrics Chosen parameter was set to 1.*** p<0.01, ** p<0.05, * p<0.1.

Table C4.1: Effect of Metrics Training on Policy (with Metrics Chosen = 0)

	(1)	(2)	(3)	(4)	(5)	(6)
	Letter Sent	Funds Recommended	Letter Sent	Funds Recommended	Letter Sent	Funds Recommended
Metrics Assigned	0.299**	443,944***	0.0108	24,135	-0.000548	-3,282
	[0.122]	[149,370]	[0.0997]	[36,913]	[0.104]	[22,659]
Observations	109	109	109	109	109	109
Mean of dep. var. (placebo)	0.174	171812.1	0.174	51073.83	0.262	41744.97
R-squared	0.246	0.271	0.156	0.126	0.156	0.171

Robust standard errors appear in brackets (clustered at the individual level). The dependent variables are letters sent and funds recommended to government divisions in Pakistan (in Pakistani Rupees) for budget allocation for Deworming, Orphanage and School renovations, respectively. Metrics Assigned is a dummy variable that switches on when a causal inference book is randomly assigned to participants. Consistent with all our earlier regressions, we always control for metrics chosen. The causal inference book, Mastering Metrics, is randomly assigned conditional on the book being chosen. The estimations obtained from OLS regressions include the following controls: metrics chosen, written test scores, interview test scores, gender, birth in political capitals, asset ownership, income before joining public service, age, education, foreign visits and occupational group dummies. In this regression, the Metrics Chosen parameter was set to 0. *** p<0.01, ** p<0.05, * p<0.1.

Table C4.2: Effect of Metrics Training on Policy (with Metrics Chosen = 1)

	(1)	(2)	(3)	(4)	(5)	(6)
	Letter Sent	Funds Recommended	Letter Sent	Funds Recommended	Letter Sent	Funds Recommended
Metrics Assigned	0.305**	373,264**	0.0698	34,614	-0.010 1	-3,797
	[0.122]	[162,049]	[0.0858]	[23,766]	[0.136]	[19,970]
Observations	75	75	75	75	75	75
Mean of dep. var. (placebo)	0.174	171812.1	0.174	51073.83	0.262	41744.97
R-squared	0.231	0.274	0.276	0.290	0.179	0.160

Robust standard errors appear in brackets (clustered at the individual level). The dependent variables are letters sent and funds recommended to government division in Pakistan (in Pakistani Rupees) for budget allocation for Deworming, Orphanage and School renovations, respectively. Metrics Assigned is a dummy variable that switches on when a causal inference book is randomly assigned to participants. Consistent with all our earlier regressions, we always control for metrics chosen. The causal inference book, Mastering Metrics, is randomly assigned conditional on the book being chosen. The estimations obtained from OLS regressions include the following controls: metrics chosen, written test scores, interview test scores, gender, birth in political capitals, asset ownership, income before joining public service, age, education, foreign visits and occupational group dummies. In this regression, the Metrics Chosen parameter was set to 1. *** p<0.01, ** p<0.05, * p<0.1.

Table C5.1: Shifting from Initial Beliefs on Policy Decisions (with Metrics Chosen = 0)

	(1)	(2)	(3)
	Deworming Letter Sent	Deworming Letter Sent	Deworming Funds Recommended
Initial Beliefs X Delta Beliefs X Prior Beliefs	-0.0862**	-0.0740	-137,874**
	[0.0415]	[0.0457]	[66,747]
Delta Beliefs	0.0478**	0.0505*	67,013**
	[0.0236]	[0.0263]	[33,568]
Constant	0.251***	2.876*	214,622***
	[0.0703]	[1.535]	[78,819]
Observations	106	106	106
Mean of dep. variable	0.174	0.174	171812.1

Robust standard errors appear in brackets (clustered at individual level). The dependent variables in Columns (1) and (2) are the dummy variables that switch on if the deputy ministers send letters to recommend funding for deworming policy from the respective government divisions, while for Columns (3) and (4) are exact amounts of funds recommended (in Pakistani Rupees). Initial Belief >13% is a dummy variable that switches on if the minister had a pre-signal belief that deworming has a larger than 13% effect on income (the RCT signal amount). Delta belief is the update in beliefs calculated as the post-signal belief minus the initial belief. Delta belief interacted with Initial Belief > 13% is instrumented for by the interaction of metrics assigned and Initial Belief > 13%. Delta Belief is instrumented by metrics assigned. Initial Belief > 13% is included as a control in the first and second stage equation. The independent variables are the interaction of delta beliefs and prior dummy, metrics assigned, and prior belief. Consistent with all our earlier regressions, we also control for 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. In this regression, the Metrics Chosen parameter was set to 0. *** p<0.01, ** p<0.05, * p<0.1.

Table C5.2: Shifting from Initial Beliefs on Policy Decisions (with Metrics Chosen = 1)

	(1)	(2)	(3)	(4)
	Deworming Letter Sent	Deworming Letter Sent	Deworming Funds Recommended	Deworming Funds Recommended
Initial Beliefs X Delta Beliefs X Prior Beliefs	-0.0748	-0.0945	-92,721	-111,349
	[0.0459]	[0.0604]	[65,691]	[81,389]
Delta Beliefs	0.0410**	0.0325	48,126*	33,980
	[0.0205]	[0.0312]	[27,602]	[37,515]
Observations	68	68	68	68
Mean of dep. variable	0.174	0.174	171812.1	171812.1

Robust standard errors appear in brackets (clustered at individual level). The dependent variables in Columns (1) and (2) are the dummy variables that switch on if the deputy ministers send letters to recommend funding for deworming policy from the respective government division, while for Columns (3) and (4) are exact amounts of funds recommended (in Pakistani Rupees). Initial Belief >13% is a dummy variable that switches on if the minister had a pre-signal belief that deworming has a larger than 13% effect on income (the RCT signal amount). Delta belief is the update in beliefs calculated as the post-signal belief minus the initial belief. Delta belief interacted with Initial Belief > 13% is instrumented for by the interaction of metrics assigned and Initial Belief > 13%. Delta Belief is instrumented by metrics assigned. Initial Belief > 13% is included as a control in the first and second stage equation. The independent variables are the interaction of delta beliefs and prior dummy, metrics assigned, and prior belief. Consistent with all our earlier regressions, we also control for 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. In this regression, the Metrics Chosen parameter was set to 1. *** p<0.01, ** p<0.05, * p<0.1.

Table C6.1: Effect of metrics Training on Willingness to Pay (with Metrics Chosen = 0)

	(1)	(2)	(3)	(4)	(5)	(6)
	Amount Randomized Trial	Amount Correlational Data	Amount Expert Bureaucrat	Amount Randomized Trial	Amount Correlational Data	Amount Expert Bureaucrat
Metrics Assigned	1,572**	-884.9**	-2,202	118,928	-62,084***	-52,842
	[627.3]	[378.8]	[1,829]	[1.666e+06]	[19,649]	[43,975]
Observations	108	108	108	108	108	108
Mean of dep. variable	1539.453	2214.07	4490.008	928546.1	94136.72	115937.5
R-squared	0.169	0.218	0.157	0.136	0.213	0.167

Robust standard errors clustered at individual level appear in brackets. Dependent variables are the private and public stated willingness to pay for a randomized control trial, correlational comparison of means data, and suggestion from expert bureaucrat. 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. In this regression, the Metrics Chosen parameter was set to 0. *** p<0.01, ** p<0.05, * p<0.1.

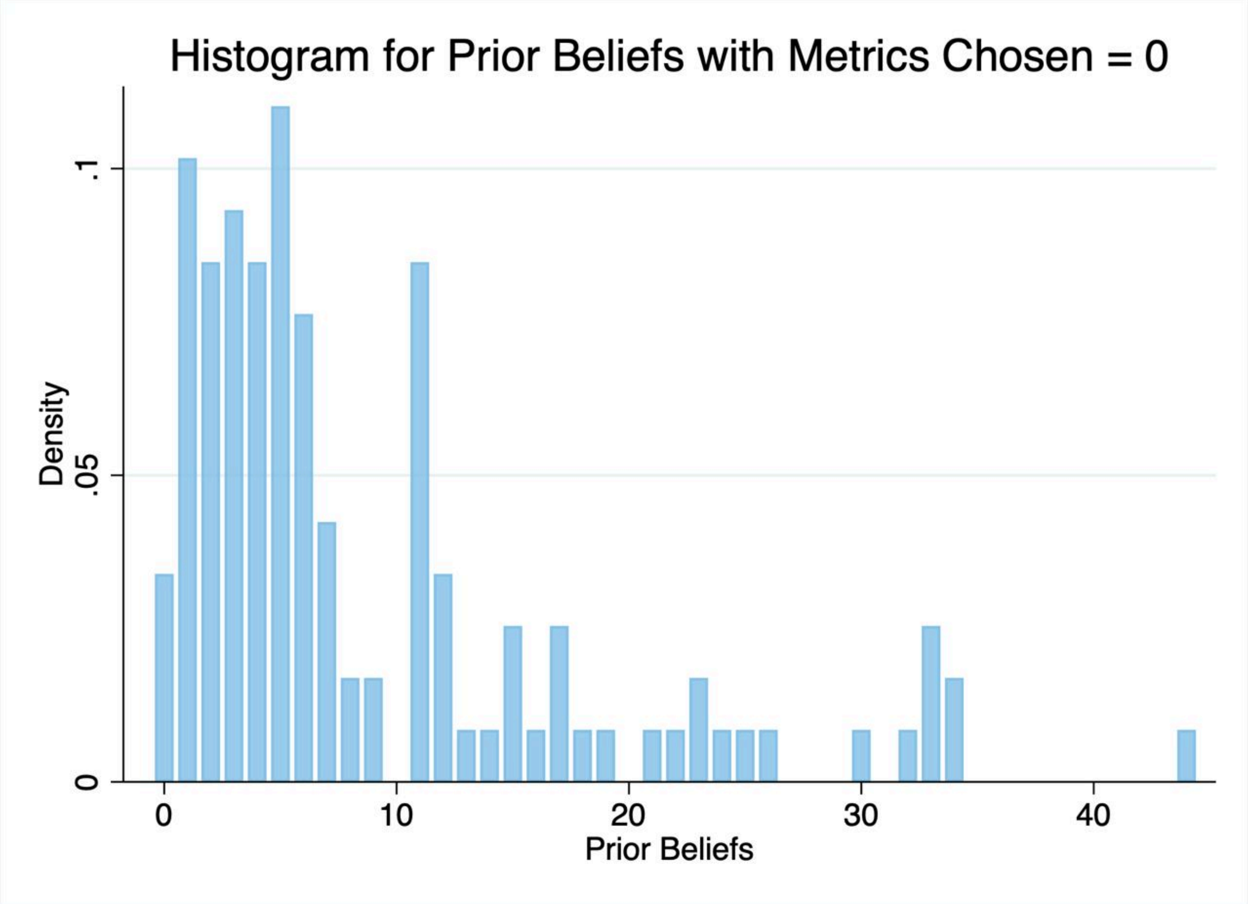
Table 6.2: Effect of metrics Training on Willingness to Pay (with Metrics Chosen = 1)

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	Amount Randomized Trial	Amount Correlational Data	Amount Expert Bureaucrat	Amount Randomized Trial	Amount Correlational Data	Amount Expert Bureaucrat
Metrics Assigned	2,881***	-1,512**	-309.7	1.172e+06	-8,779	84,222*
	[1,037]	[586.8]	[2,774]	[856,762]	[20,077]	[45,016]
Observations	72	72	72	72	72	72
Mean of dep. variable	1539.453	2214.07	4490.008	928546.1	94136.72	115937.5
R-squared	0.354	0.292	0.219	0.229	0.181	0.147

Robust standard errors clustered at individual level appear in brackets. Dependent variables are the private and public stated willingness to pay for a randomized control trial, correlational comparison of means data, and suggestion from expert bureaucrat. 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. In this regression, the Metrics Chosen parameter was set to 1. *** p<0.01, ** p<0.05, * p<0.1.

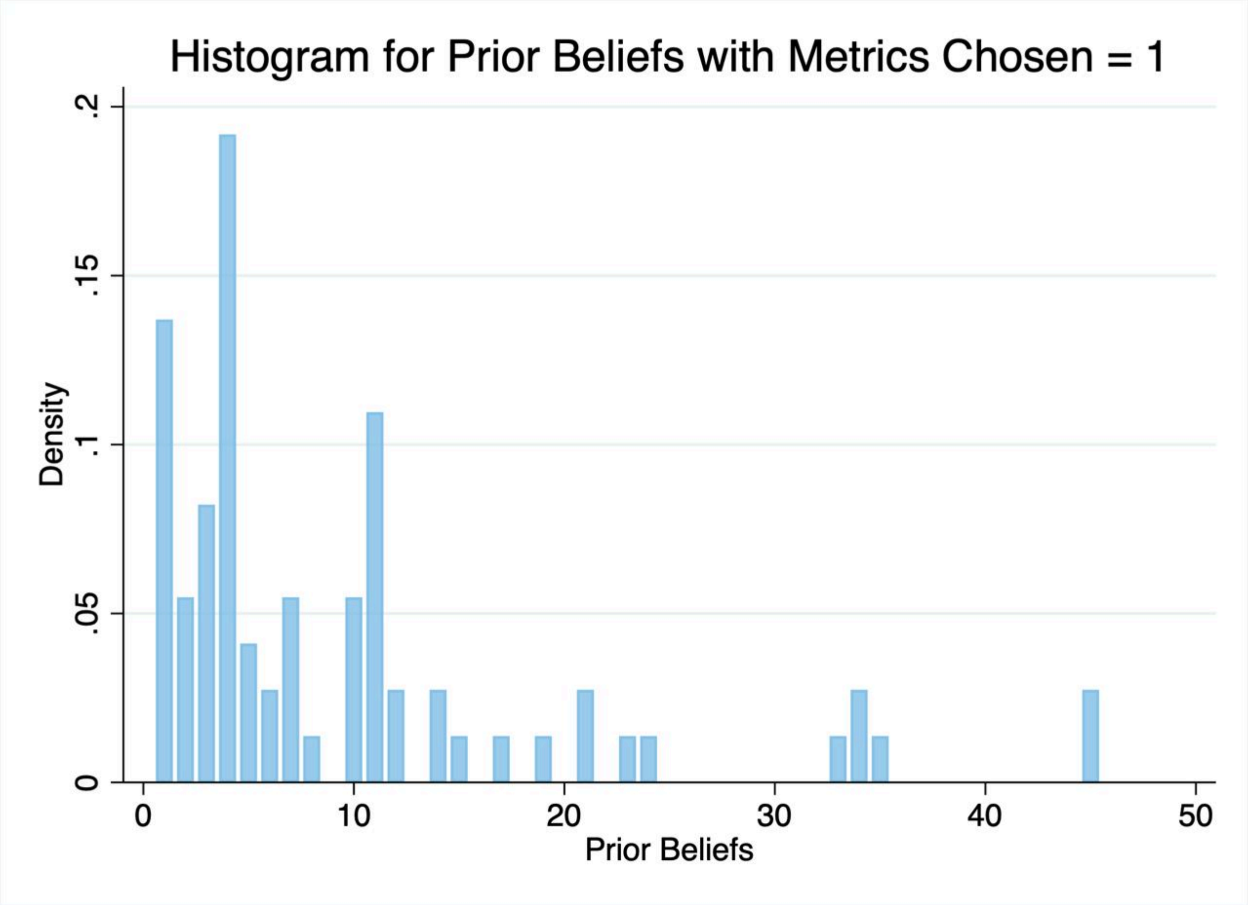
Appendix C Figures

Figure C1.1: Histogram for Prior Beliefs with Metrics Chosen = 0



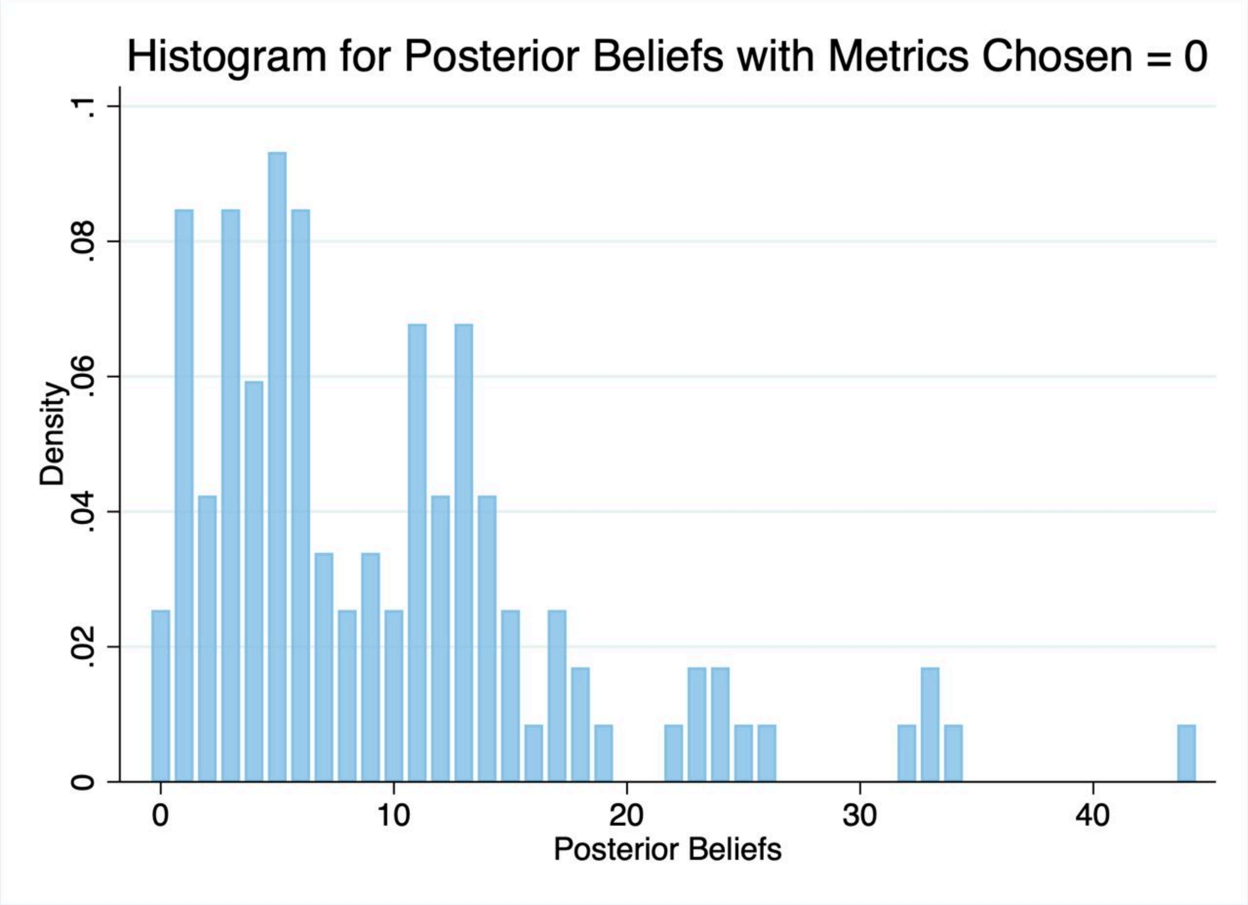
Note: The figure above plots histogram of Prior Beliefs by its absolute values. It shows Prior Beliefs histogram with Metrics Chosen parameter set to 0. The correct answer was 13% impact of deworming on wages.

Figure C1.2: Histogram for Prior Beliefs with Metrics Chosen set to 1



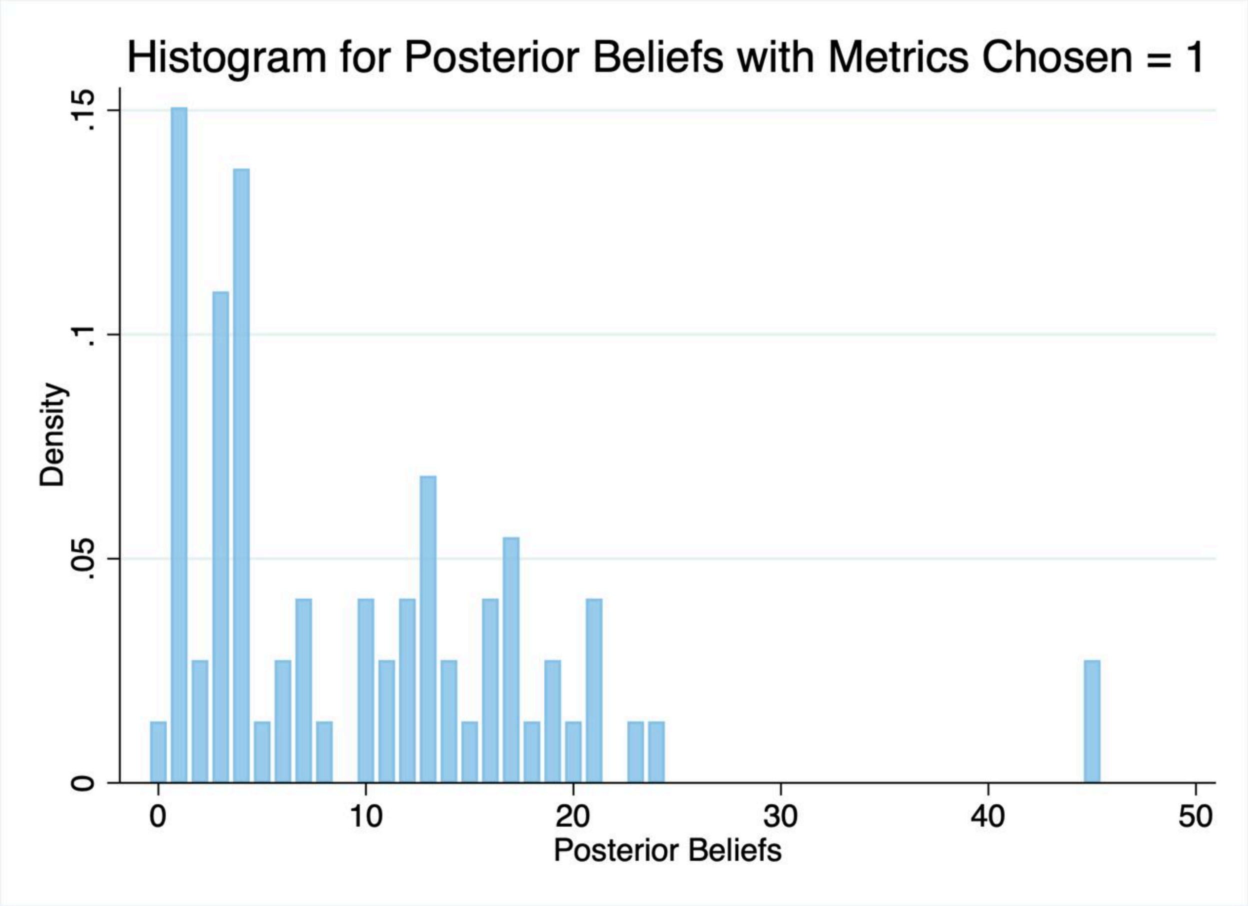
Note: The figure above plots histogram of Prior Beliefs by its absolute values. It shows Prior Beliefs histogram with Metrics Chosen parameter set to 1. The correct answer was 13% impact of deworming on wages.

Figure C2.1: Histogram for Posterior Beliefs with Metrics Chosen set to 0



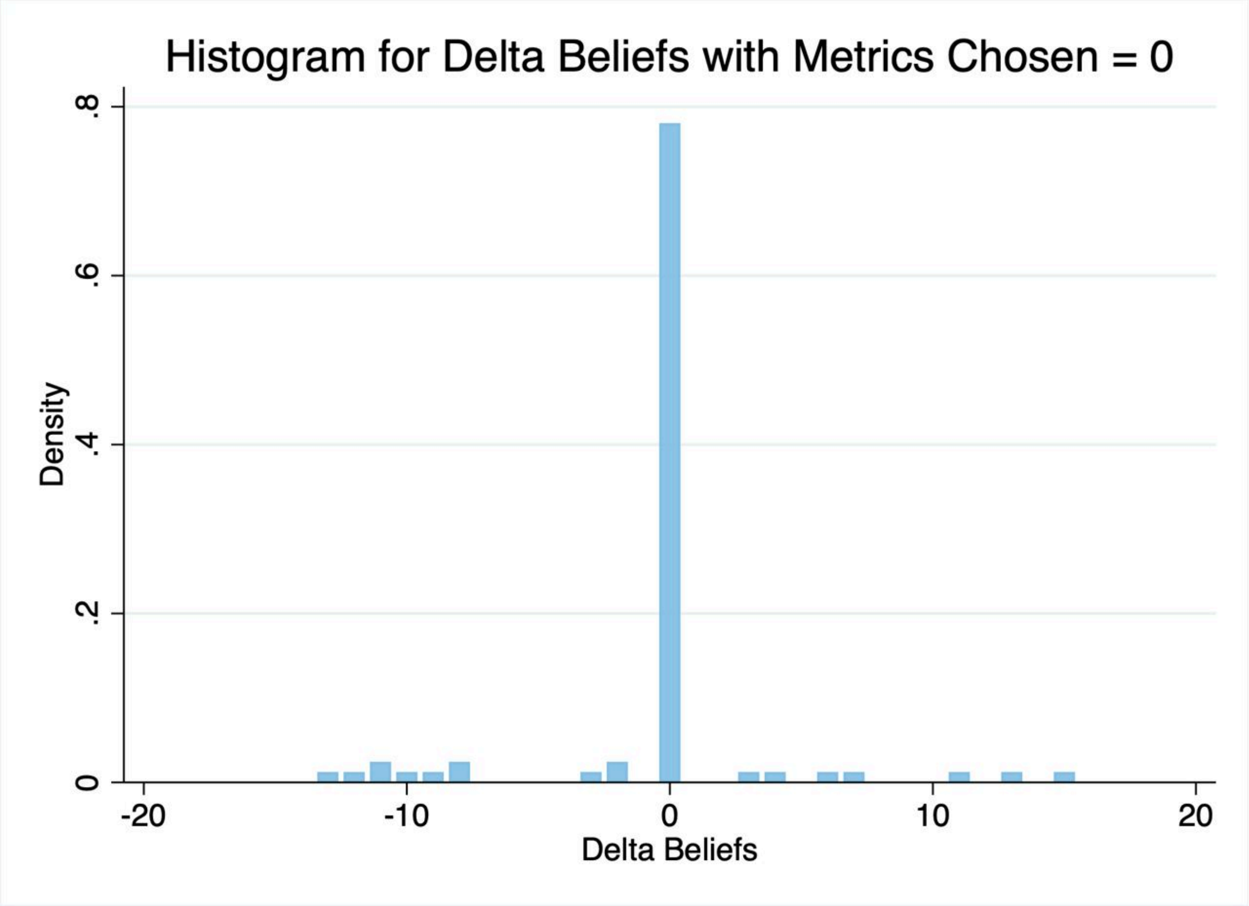
Note: The figure above plots histogram of Posterior Beliefs by its absolute values. It shows Posterior Beliefs histogram with Metrics Chosen parameter set to 0.

Figure C2.2: Histogram for Posterior Beliefs with Metrics Chosen = 1



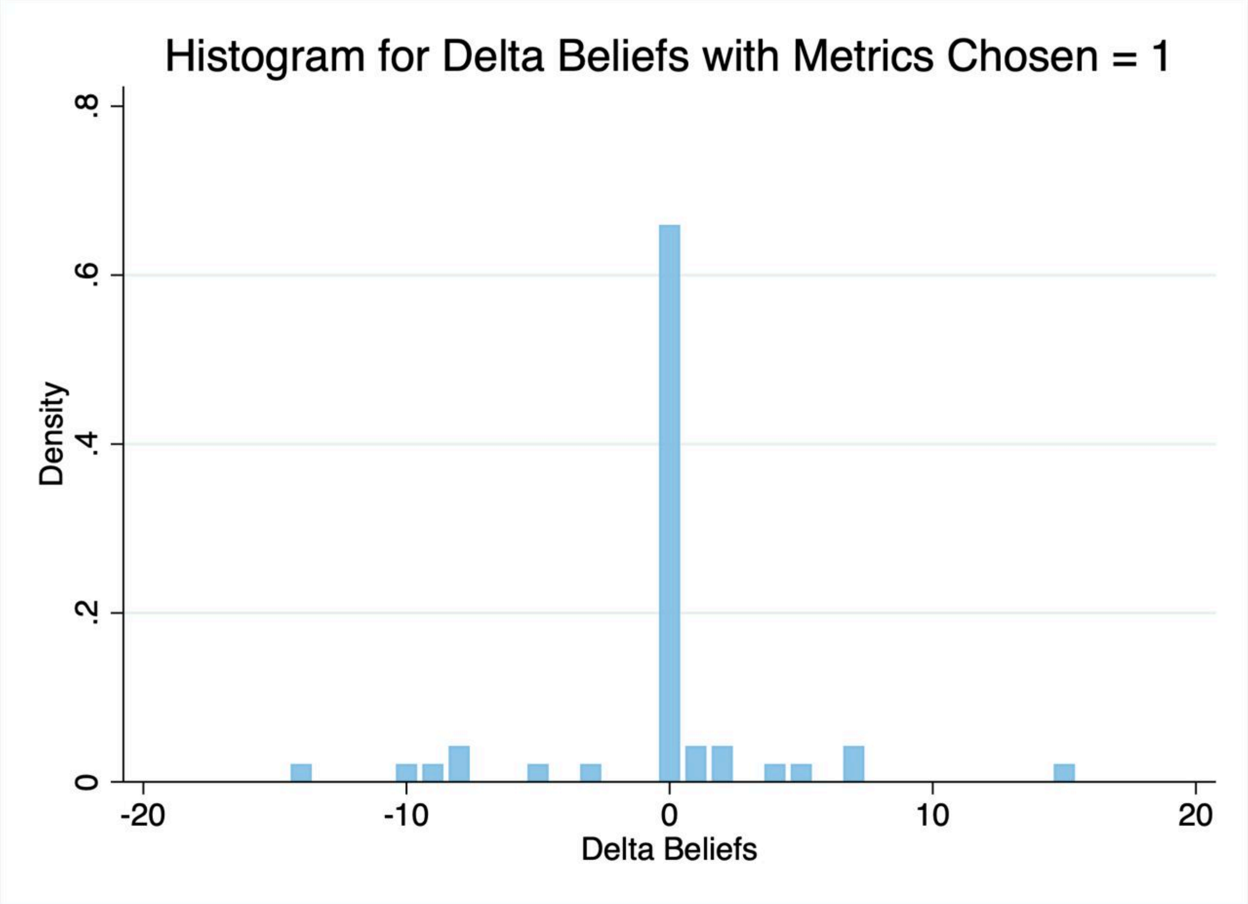
Note: The figure above plots histogram of Posterior Beliefs by its absolute values. It shows Posterior Beliefs histogram with Metrics Chosen parameter set to 1.

Figure C3.1: Histogram for Delta Beliefs with Metrics Chosen set to 0



Note: The figure above plots histogram of Delta Beliefs by its absolute values, which is a difference between Prior Beliefs and Posterior Beliefs. It shows Delta Beliefs histogram with Metrics Chosen parameter set to 0.

Figure C3.2: Histogram for Delta Beliefs with Metrics Chosen set to 1



Note: The figure above plots histogram of Delta Beliefs by its absolute values, which is a difference between Prior Beliefs and Posterior Beliefs. It shows Delta Beliefs histogram with Metrics Chosen parameter set to 1.