Expectations and Adaptation to Environmental Threats*

Husnain F. Ahmad, Matthew Gibson, Fatiq Nadeem,

Sanval Nasim, Arman Rezaee

July 2025

Abstract

Scarce information and human capital may make it difficult to form accurate expectations, limiting responses to uncertain environmental threats like air pollution. We study two cross-randomized interventions in Lahore, Pakistan: 1) general training in forecasting; 2) provision of air pollution forecasts. Both reduced subjects' own air pollution forecast errors; the training effect suggests that modest educational interventions can durably improve forecasting skills. Forecast receipt increased demand for protective masks and the responsiveness of outdoor time to pollution. Forecast recipients were willing to pay 41 percent of their total mobile phone costs for continued access, consistent with welfare gains from adaptation.

Keywords: expectations, forecasting training, environmental threats, adaptation

JEL: Q56, Q53, D84, D90

---

*Corresponding author: Arman Rezaee, University of California, Davis; Email: abrezaee@ucdavis.edu. Ahmad: Department of Economics, Trinity University. Gibson: Department of Economics, Williams College, and IZA. Nadeem: Bren School of Environmental Science & Management, University of California, Santa Barbara. Nasim: Department of Economics, Colby College. Rezaee: Department of Economics, University of California, Davis. We are grateful to Clare Balboni, Chris Blattman, Kelsey Jack, Asim Khwaja, Asad Liaqat, Craig McIntosh, Rohini Pande, Marc St-Pierre, Chris Udry, and Catherine Wolfram for their helpful comments. We thank Yasir Iqbal, Mahnoor Kashif and Shotaro Nakamura for outstanding research assistance; and Mohammad Malick and Ahsan Tariq for excellent management of field work and data collection. This study is funded by grants from the International Growth Center (IGC); Lahore University of Management Sciences (LUMS); Abdul Latif Jameel Poverty Action Lab (JPAL); the Williams College Center for Development Economics and the University of the South. We would like to thank AirVisual (IQAir), AirNow, MeteoBlue, SPRINTARS, and the Pakistan Air Quality Initiative for sharing their data. This study has IRB approval in the US (UC Davis IRB-1336133-1) and Pakistan (LUMS). Our pre-analysis plan is registered with the AEA (RCTR-0004343). A previous version of this paper was titled "Forecasts: Consumption, Production, and Behavioral Responses."

1 Introduction

Contemporary people face risky and uncertain environmental threats, from climate change to water and air pollution. In order to adapt effectively, they must form expectations [forward-looking beliefs or forecasts; Bakkensen and Barrage, 2022]. For example, a farmer in India must form an expectation over the arrival of increasingly erratic monsoon rains in order to choose the planting date for rice. Forming accurate and precise expectations may be difficult, however, particularly in an information-scarce developing-country context [Stiglitz, 2000].[1] Underlying distributions may be changing over time [Kala, 2019]. Past outcomes of random processes may not be readily observable, and correlated environmental threats may be impossible to disentangle [Patel, 2023]. Third-party forecasts may be scarce or of low quality [Rosenzweig and Udry, 2013, 2014]. Levels of relevant human capital may be low [Jacoby and Skoufias, 1997, Beine et al., 2008], and ecumenical cognitive biases may compound the problem of producing useful expectations [Tetlock, 2017]. The resulting errors are plausibly consequential and create scope for beneficial interventions.

Air pollution provides a suitable domain to study expectations over environmental threats, particularly in the developing world [Chang et al., 2019]. It varies at high frequency, with large changes occurring from one day to the next. This allows a subject to produce or consume multiple forecasts over the course of an experiment, and creates scope for changes in a subject's forecasting process. Uncertainty over air pollution matters, as air pollution affects mortality and health [Knittel et al., 2016, Barrows et al., 2019, Barreca et al., 2021, Garg et al., 2023], as well as labour productivity [Chang et al., 2016a, Neidell, 2017, Chang et al., 2019, Adhvaryu et al., 2022] and crime [Ayesh, 2023].[2] Because of these consequences, one can reasonably expect that subjects take air pollution forecasting seriously. Air pollution has also become an ubiquitous part of life in developing cities [IQAir, 2023], making it a more natural forecasting domain than those sometimes employed in lab studies (e.g. stock prices).

In this paper, we exploit uncertain air pollution to study how people form expectations and make adaptation choices in the presence of limited information and human capital. We concern ourselves with the following broad questions. Can developing-country residents form useful forecasts, and can their forecasting ability be improved? How do changes in forecast inputs influence adaptation, especially avoidance of environmental harm? Do people

[1]Stiglitz [2000] writes, "One of the central aspects of less developed countries is that markets work less efficiently, including 'markets for information.' "

[2]Other important work in this area includes: Alberini et al. [1997], Cropper et al. [1997], Jeuland et al. [2015], He et al. [2019], Bishop et al. [2022], Gilraine and Zheng [2022], Persico [2022]. For reviews, see Graff Zivin and Neidell [2018] and Aguilar-Gomez et al. [2022].

exhibit positive demand for forecast products, consistent with welfare gains from changes in adaptation? The answers to these questions shed light on forward-looking human decision making. They are also important inputs to benefit-cost analyses of policies concerning air pollution monitoring and abatement.

To address these research questions we implemented a randomized controlled trial, which included two cross-randomized treatments: 1) in-person training designed to improve general forecasting performance across all domains, e.g. by avoiding base-rate neglect;[3] 2) day-ahead air pollution forecasts delivered by text message (SMS) for eight months. In theoretical terms, we model these two treatments as shocks to inputs in an agent's forecast production function: training increases human capital, while text-message pollution forecasts increase information.[4] Broadly, three types of outcomes interest us: 1) expectations, e.g. error in forecasting air pollution; 2) adaptive behavioural responses, e.g. willingness to pay for particulate-filtering face masks; and 3) demand for inputs to expectations and adaptation, e.g. willingness to pay (WTP) for our air pollution forecast product.

Our experiment involved 999 subjects in Lahore, Pakistan. In 2019, Lahore ranked as the twelfth most polluted city in the world, with air quality roughly comparable to cities like New Delhi, India; N'Djamena, Chad; and Baghdad, Iraq [IQAir, 2020, Riaz and Hamid, 2018, Zahra-Malik, 2017]. While Lahore experiences acute pollution, its residents face a challenging information landscape in which to make useful forecasts. Some sources (public and private) provide retrospective information, but such efforts remain incomplete in space and time and information quality is uncertain.[5] The Punjab Government's Environmental Protection Department (EPD) posts past measurements, but only online in English.[6] The US consulate in Lahore recently began providing hourly pollution averages online, but these represent one point in a city with an area of more than 680 square miles. Retrospective and real-time air pollution readings are not readily available to residents—especially the majority who do not speak English—while air pollution forecasts are entirely absent.[7] This kind of information environment is common in developing cities.[8]

Average levels of human capital in Lahore may also hamper residents' ability to forecast

[3] For example, a person who forecasts the probability of rain tomorrow without considering the long-term mean probability of rain in her location exhibits base-rate neglect.

[4] In our theoretical model (Section 2), information and human capital may be complements or substitutes.

[5] Manipulation of air pollution readings has been documented in other developing-country settings [Ghanem and Zhang, 2014, Ghanem et al., 2020].

[6] According to the Punjab Government: "Data on air quality in the province is scant. Sporadic monitoring of air pollutants suggests that ambient air standards for particulate matter with size 2.5 micron ($PM_{2.5}$) ... are exceeded frequently" [Punjab Environmental Protection Department, 2017].

[7] In pilot surveys, respondents were asked to rank real-time alerts, retrospective readings, and forecasts from most to least desirable. 69 percent ranked forecasts first, and 25 percent ranked them second.

[8] Section 5.7 provides evidence for this claim.

accurately. Citywide, average educational attainment lies between 6.2 and 6.5 years [NIPS and ICF, 2019]. In our subject population, it is 9.3 years. Pakistan's nationwide educational attainment (4.8 years) is a year lower than India's, and roughly comparable to Uganda's, Ethiopia's, and Nigeria's [World Bank, 2017]. These countries' urban residents may face skill constraints similar to those of our subjects. Moreover, Lahore's residents may confront the same behavioural biases that generate forecasting errors even in highly educated populations [Kahneman and Tversky, 1973].

Using incentive-compatible elicitations, we find that treatment with SMS forecasts and training reduced error in incentivized one-day-ahead forecasts of fine particulates ($PM_{2.5}$) by 0.11 and 0.15 standard deviations respectively (6 and 8 $\mu g/m^3$).[9] This equals approximately 50 percent of the World Health Organization's corresponding maximum safe 24-hour standard.[10] These improvements in mean forecast error are driven in part by reduced underestimation of air pollution levels. At endline, the median respondent underestimated $PM_{2.5}$ levels by 40.2 $\mu g/m^3$ in the control group, compared to 35.7 $\mu g/m^3$ in the treatment groups. Additionally, while both interventions reduced mean forecast error, training reduced the variance of forecast error across subjects. Given that four to six months elapsed between the training and the forecast elicitation, these forecast improvements are notable and consistent with a durable increase in human capital.[11] More generally, the effect of training suggests that helping people use information is at least as important as delivering new information [Hanna et al., 2014].

Forecast provision increased willingness to pay for particulate-filtering masks by 6.6 Pakistani Rupees (PKR), roughly five percent of the retail price.[12] While the estimated effect of training on mask demand is positive (4 PKR), it is imprecise. These results indicate that a sparse information environment is one of the factors limiting the takeup of masks—and more broadly pollution avoidance—in developing countries. This may be particularly true when less informed people underestimate air pollution levels, as in our context. We also estimate the effect of forecast receipt on outdoor time. This is an important margin of adaptation to air pollution, as outdoor pollution exposure is frequently higher than indoor [US Centers for Disease Control and Prevention, 2022]. Forecast receipt increased outdoor

[9] Forecasts were incentivized using payments for responses within 5, 10, or 20 percent of realised particulate pollution. For more details, see Section 3.2.

[10] $PM_{2.5}$ level denotes particulates with diameter 2.5 microns or less, measured in micrograms per cubic meter ($\mu g/m^3$). The World Health Organization has set the daily standard at 15 $\mu g/m^3$ and the annual standard at 5 $\mu g/m^3$ [World Health Organization, 2021].

[11] The trainings occurred in August 2019, the endline survey in January-February 2020.

[12] N95 masks filter 95 percent of small particles. According to the mask manufacturer 3M, a genuine N95 mask retailed for 135 PKR (on average) in Lahore in November 2019, while our experiment was in progress. Endline surveys were completed prior to the outbreak of Covid-19 in Pakistan.

time by 16 percent on (infrequent) less polluted days and reduced outdoor time by 3 percent on more polluted days. That is, SMS forecasts improved the alignment of outdoor time with the level of air pollution. This pattern of responses was more pronounced for subjects who reported caring about air quality at baseline, and for children. Increased avoidance by children may have large welfare consequences because in-utero and childhood air pollution exposures affect long-run health, labour force participation, and human capital [Currie et al., 2014].

Subjects exposed to our one-day-ahead air pollution forecasts were willing to pay an average of 93 PKR to continue receiving forecasts for 90 days.[13] Viewed through the lens of our model, this willingness to pay reflects the value of additional avoidance facilitated by the forecasts. On a monthly basis, 93 PKR equals roughly 41 percent of households' total mobile phone costs. It stands in contrast to low willingness to pay for health-promoting goods like insecticide-treated nets and chlorine [Kremer et al., 2011]. Drawing on third-party demographic data, we show that our demand estimate implies substantial welfare benefits from air pollution forecasting in developing countries, even allowing for uncertain benefit transfer and substantial monitoring costs.

The timing of our training—roughly mid-way through the eight months of SMS forecasts—allows for a better understanding of the dynamics of our treatment effects, as well as of the interaction of the two interventions. Incentive-compatible forecast elicitations were conducted at the beginning and end of the one-hour training. Subjects who had been receiving our SMS forecasts for several months started the training session performing better than those who had not, and this intent-to-treat effect on forecast error at the study mid-point was similar to the effect at endline.[14] Over the training session, those who had not been receiving forecasts caught up in terms of forecast error, while those who had been receiving forecasts saw no additional improvement. These results could be interpreted as evidence of a ceiling on forecast accuracy, operating perhaps through memory or cognition. In the context of our model, they could be interpreted as evidence that information and human capital are substitutes in subjects' forecast production functions.

Our study contributes to several literatures, of which the first is on environmental beliefs. In a US setting, Bakkensen and Barrage [2022] study expectations of future flood risk. Previous work in developing countries has focused on farmers, including their beliefs over soil salinity [Patel, 2023], monsoon onset [Giné et al., 2015], and drought risk [Zappala, 2024]. Other studies have inferred precipitation beliefs from planting decisions [Kala,

[13]Willingness to pay for both forecasts and masks was elicited using a Becker-DeGroot-Marschak mechanism [Becker et al., 1964].

[14]The intent-to-treat effect of SMS forecasts on t+1 forecast error was -0.13 standard deviations at the beginning of training, versus -0.11 standard deviations at endline.

2019] or estimated responses to precipitation forecasts [Burlig et al., 2024, Patt et al., 2005, Rosenzweig and Udry, 2013, 2014]. While we also study responses to forecasts, this paper makes novel contributions on several dimensions. Perhaps most importantly, our experiment shocks beliefs not only through subjects' information sets (as prior studies have done), but also through their human capital. We elicit beliefs directly using incentive-compatible mechanisms, rather than relying on stated beliefs or inferring beliefs within a structural model.

Second, we contribute to the body of work on training interventions in developing countries. Previous research has focused on business and entrepreneurship skills [Karlan and Valdivia, 2011, McKenzie and Woodruff, 2014, Martínez A et al., 2018], or job training [Card et al., 2011, Acevedo et al., 2017]. Our paper instead considers training in general-purpose forecasting skills. Researchers have sought to improve forecasting performance in high-income settings, typically with highly educated subjects [Mellers et al., 2014, Morewedge et al., 2015, Soll et al., 2015]. Our study adapts such techniques to a low-income setting.

The third relevant literature is on adaptation to environmental threats. A large body of work naturally focuses on adaptation to climate change [Auffhammer, 2018]. Examples include Burke and Emerick [2016], Mullins and Bharadwaj [2021], and Garg et al. [2020]; for a development-focused review, see Kala et al. [2023]. Within this literature, our paper is most closely related to a set of studies on pollution avoidance behaviour [Graff Zivin and Neidell, 2013].[15] Many studies address avoidance behaviour in developed countries. Prominent examples include Neidell [2004], Graff Zivin and Neidell [2009], and Moretti and Neidell [2011].[16] We provide evidence from a low-income developing country, where both preferences and the scope for avoidance may differ (e.g. because of available technologies or jobs). This literature has not explicitly addressed expectations, and as a result one cannot tell whether studied populations are choosing avoidance consistent with the true state of the world. Our experimental design allows us to observe both expectations and avoidance behaviours for the same subjects.

Lastly our results fit within the literature on demand for environmental information.[17] Our paper contributes by eliciting the value of air pollution information directly, using an incentive-compatible mechanism. This recovers the entire demand curve, with an average elasticity of -.93.[18] In similar concurrent work, Patel [2023] experimentally estimates demand

[15] Air pollution avoidance behaviour may take the form of ex ante adaptation, e.g. purchasing a protective mask, or ex post adaptation, e.g. rescheduling a planned outdoor activity at the last minute.

[16] Graff Zivin and Neidell [2013] provides a thorough review, including a brief theoretical foundation.

[17] A related literature studies health information in developing countries. For a review, see Dupas and Miguel [2017].

[18] This estimate joins the large set roughly consistent with Samuelson [1965].

for soil salinity information in Bangladesh. In prior work, Barwick et al. [2019] recover a lower bound on the value of air pollution information in China.

The rest of the paper proceeds as follows. Section 2 presents our theoretical model and Section 3 discusses experimental design. Section 4 describes our approach to empirical analysis. Section 5 discusses estimated treatment effects and Section 6 concludes.

2 Theoretical model

In this section, we build a simple model of pollution avoidance by a forward-looking agent. Consider an individual who at the end of the day ($t = 0$) is planning activities for the next day ($t = 1$). Her payoff depends on the level of air pollution tomorrow and there are two possible states of pollution $s \in \{h, l\}$, high and low. The agent consumes only at $t = 1$. The effects of pollution can be mitigated by engaging in avoidance behaviour, which can be purchased in both periods. Examples of avoidance in our setting include protective face masks and cancellation or rescheduling of planned outdoor activities.[19] Let x and y denote the amount of avoidance purchased in periods 0 and 1 respectively. The agent's payoff is

$$E - d^s(x + y) - c(x, y),$$

where $E > d(0)$ is her initial endowment, large enough to avoid any credit constraints. The function d^s is state-dependent damage,[20] assumed to be decreasing and strictly convex in the sum of avoidance purchased. The assumption that damage is decreasing in the sum of avoidance implies avoidance actions are perfect substitutes across the two periods. This matches our setting where, for example, cancelling plans a day ahead is a perfect substitute for a cancelling them on the day.[21] We further assume that both the magnitude of damage and the marginal benefit of avoidance are increasing with the level of pollution, that is $d^h(A) \geq d^l(A)\ \forall A$ and $d^h_1(A) \leq d^l_1(A)\ \forall A$.[22] The cost of avoidance is captured through

[19]One might object that masks are a durable good. We do not model them as such because 1) masks have a limited life span, roughly 1 to 30 days in our Lahore setting, and 2) the cost of avoidance can be viewed in terms of opportunity cost, i.e. use of a mask today prevents usage later.

[20]For simplicity, we assume that there is no credit constraint and that the agent is risk neutral. While extension to risk aversion is possible, it reduces tractability without adding interesting results. We are unable to study changes in risk aversion, as those involve comparing lotteries that are significantly different. A specialized model making this point is presented in Appendix I.

[21]The damage function can be generalized to any weighted sum, e.g $A = x + \epsilon y$. Generalizing further is possible, say to a damage function of the form $d(x, y)$, if we either 1) make an interim assumption while solving, similar to Rosenzweig and Udry, 2013, or 2) make assumptions on the third derivatives of the damage function.

[22]A notation reminder is in order: we denote the partial derivative of a real valued function $f(\vec{a})$ with respect to the i-th argument as f_i.

the cost function c, which is assumed to be strictly convex and increasing in both x and y. The marginal cost of avoidance increases if the agent waits. This may be thought of as capturing increased search costs or higher price from a time-constrained search for a mask, or the increased difficulty of rescheduling outdoor activities at the last minute.

Mathematically, this requires that globally $c_1 \leq c_2$. We ensure this by assuming that $c(0,0) = c_1(0,0) = c_2(0,0) = 0$ and that for all x and y, $c_{11} \leq c_{12} \leq c_{22}$. As costs are convex in each period's avoidance, making this assumption ensures that at any (x, y), buying more x increases the marginal cost of x by less than the marginal cost of more y ($c_{11} \leq c_{21}$). Similarly, buying more y raises marginal cost of y by more than that of x ($c_{22} \geq c_{12}$). This is perhaps easier to see for the example where $c(x + \beta y)$, with $\beta > 1$. Then the marginal cost of y is always higher than that of x, and the above assumptions hold. Also note that $c(0,0) = c_1(0,0) = c_2(0,0) = 0$ is an elective normalization; the only requirement is that $c_1(0,0) \leq c_2(0,0)$.

The level of pollution is unknown at time 0 but revealed at time 1. The probability of high pollution is $P(h) = \pi$, which can also be interpreted as the agent's unbiased prior before she begins optimizing. While we assume that π is an exogenous constant, we note that it can be a function of various other factors, such as the weather, and while we employ a simple day-ahead model, the prior may itself vary day by day.

Then given π, in the process of optimizing the agent forms an internal forecast, $F \in \{H, L\}$, of tomorrow's pollution. Her forecasting performance depends on her human capital τ and her level of information ι at $t = 0$, both exogenous. Our treatments are designed to vary (increase) the level of ι and τ, through either direct provision of our forecast or through a learning exercise designed to improve forecast ability, respectively. We define the probability of a correct forecast as the agent's skill, $P(H|h, \iota, \tau) = P(L|l, \iota, \tau) = \rho(\iota, \tau)$, and assume she is equally good at predicting high and low pollution. We assume that skill is increasing in both information and human capital, but make no assumption on their interaction (i.e. whether they are substitutes or complements),[23] though our empirical results suggest that they are substitutes. Finally we assume that, given ι and τ, the forecast is weakly useful. Formally this requires $\rho(\iota, \tau) \geq \max\{\pi, 1 - \pi\}$.

2.1 Hypotheses

The results of our model, provided in detail in Appendix A, yield a set of testable hypotheses. First, our model shows that there is a positive willingness to pay for improvements in forecast

[23] In principle information and human capital could be substitutes or complements. Substitution might arise, for example, if greater forecast skill leads subjects to pay less attention to information or remember it with greater error. Complementarity might arise, for example, if greater forecast skill helps subjects weigh information more appropriately when producing a forecast.

ability (ρ). Given the assumption that our treatments (detailed in the next section) improve an agent's forecast ability, our model yields the hypothesis below.

Hypothesis 1. *Willingness to pay for services that improve the agent's forecast is non-zero.*

Turning to avoidance behaviour, our model predicts that those who anticipate higher levels of pollution tend to invest in higher levels of avoidance. Avoidance acts as insurance against damage caused by pollution. Improved forecast information and skill allow our agent to avoid pollution in a more sophisticated manner, undertaking more (costly) avoidance when pollution is high, less when it is low. We can further establish that willingness to pay for avoidance (e.g. masks) is increasing in forecast skill. We note that Lahore experienced high air pollution throughout our study. Assuming that our interventions reduce forecast errors and pre-intervention agents underestimated the level of pollution, we expect the following.[24]

Hypothesis 2. *Subjects receiving our treatments should undertake more avoidance behaviour. In particular, we expect those in all treatment arms to have higher willingness to pay for masks, compared to those in the control arm.*

Similarly, our time-use data provide us with information on avoidance as a function of the forecast sent a day ahead. Our model suggests that avoidance is increasing in the level of the air pollution forecast. One of our treatments sent day-ahead forecasts of pollution to respondents through mobile SMS. If agents incorporate this information and use it to form better forecasts, we expect the following.

Hypothesis 3. *Avoidance (e.g. reduced outdoor time) is expected to better match the state (high or low pollution) among recipients of the SMS service. In particular, subjects receiving SMS forecasts should avoid more than control subjects on high-pollution days and less on low-pollution days.*

Under the additional assumption that experience with our SMS forecast increases its perceived skill, we expect the following.

Hypothesis 4. *Willingness to pay for the forecast service will be greater for those who have experience receiving the SMS service, compared to those without.*

Finally we note that the interaction effects of the two treatments are ambiguous in sign, largely because we impose no structure on the agent's forecast function $\rho(\iota, \tau)$. There is little empirical basis for restricting ρ in our setting. Agents' behaviour in combining information

[24] At end-line, control subjects underestimated the level of pollution compared to those in our treatments. The median respondent underestimated $PM_{2.5}$ levels by 40.2 $\mu g/m^3$ in the control group compared to 35.7 $\mu g/m^3$ in the treatment groups.

and human capital to produce forecasts raises interesting research questions, but they are mostly beyond the scope of this paper. While we don't formalize this as a hypothesis, we note that among participants who received the SMS service, we expect training will decrease WTP for the SMS service if training and information are substitutes ($\frac{\partial^2 \rho}{\partial \iota \partial \tau} \leq 0$) and increase it if they are complements.

3 Experimental design

We carried out our experiment with a representative sample of 1088 households from two of the eight *Tehsils* (sub-districts) of Lahore: Walton and Model Town. We selected these *Tehsils* partly for our ability to place a high-quality air pollution monitor in a locally central location. This ensured the ground truth data for our forecasts was as accurate as possible. The two *Tehsils* are middle-class, with education levels above the average for Lahore (27 percent of households have tertiary education compared to Lahore's 18.5 percent average). The external validity of results from this sample is discussed in Section 5.7. Further details of sampling and randomization are discussed in Appendix F. Figure A1 shows the division of our sample into treatment and control groups. We find no evidence of imbalance across these groups at either baseline or end-line (Tables A1 through A4).

One adult per household was invited to participate in our study. Households were randomly assigned to either have a male or a female respondent. We prioritized household heads or spouses of households heads for selection, but if they could not be reached we randomly selected another adult of the same gender from the household. All of our outcomes, including our incentive-compatible forecast elicitations, our demand elicitations, and our outdoor time use, were asked about this single respondent, except for child time use, which was asked about the youngest physically active child in the household (i.e. that is able to walk). Both of our treatments were also applied to the same respondent.

At baseline all subjects received a pamphlet explaining fine particulate air pollution ($PM_{2.5}$). A colour-coded table described potential health effects for different pollution ranges in neutral language. The pamphlet also provided the mean and 5th and 95th percentiles of the distribution of daily average fine particulate readings.[25] Broadly the goal of the pamphlet was to put all subjects—including the control group—in a position to make grossly reasonable forecasts. In the forecasts-only and forecasts-plus-training groups, we delivered SMS air pollution forecast messages to respondents every evening over a period of eight months. In the training-only and forecasts-plus-training groups, we implemented the forecast training once for every subject. More detailed descriptions of these interventions follow.

[25] The percentiles were described in colloquial language that assumed no knowledge of probability.

3.1 Treatments

3.1.1 Day-ahead air pollution forecasts

We designed a model to forecast day-ahead ($t+1$) $PM_{2.5}$ air pollution in our study neighbourhood. Our ensemble forecast combined the following inputs, using as the ground truth for weighting our own air pollution monitor in the study neighbourhood, ensuring our forecasts were local to our study respondents.[26]

1. A model based on data from our own air pollution monitors.[27] $PM_{2.5}$ levels for $t+1$ were predicted using a seven-day moving-average (MA7) model with day-of-week fixed effects and weather forecast controls. The MA7 form was selected using a cross-validation exercise.

2. A similar MA7 model based on data from the US Consulate's air pollution monitor.

3. *MeteoBlue* and *SPRINTARS* models. These are daily third-party forecasts of fine particulate pollution based on satellite data. *MeteoBlue* is a private Swiss provider of atmospheric data. *SPRINTARS* stands for Spectral Radiation-Transport Model for Aerosol Species. This model was developed primarily by Kyushu University, Japan.

For additional detail on how these models were estimated and aggregated into an ensemble forecast, see Appendix G. Appendix Figure A2 presents the accuracy of this ensemble forecast relative to the ground truth during our study period. Importantly, our model was substantially more accurate than respondent baseline forecasts, with an average absolute error of 1.06 $\mu g/m^3$ (standard deviation = 20.95) in the three months following baseline versus 39.69 $\mu g/m^3$ (standard deviation = 46.4) for respondents. We provided our forecast treatment group (*T1 and T3*) respondents two pieces of information in each SMS message: 1) an average $PM_{2.5}$ air pollution forecast for $t+1$; and 2) the realised average $PM_{2.5}$ level for the previous day ($t-1$).[28] The latter was intended to allow subjects to assess the accuracy of our previous forecasts. SMS forecast messages were delivered to subjects around 8 PM, e.g. a forecast of Tuesday's particulate air pollution would have arrived on Monday evening.

3.1.2 Forecast Training

We implemented a one-hour training in forecasting skills based on the principles of Tetlock [2017] and Kahneman [2011]. Broadly speaking, the training aimed to reduce behavioural

[26] For more detail, see Appendix Section F.4.

[27] All respondents were within a 2.5-mile radius of our primary monitor.

[28] Some governments make similar interventions. For example, the US EPA's *EnviroFlash* service allows users to receive one-day-ahead air pollution forecasts via text message.

and psychological mistakes that decrease the accuracy of subjects' forecasts. A group of specially selected and trained enumerators conducted the trainings in Urdu in subjects' homes, and subjects received 150 PKR for their participation.[29]

The first several exercises taught subjects to combine "outside" and "inside" views when making a forecast [Kahneman and Lovallo, 1993, Lovallo et al., 2012]. The outside view is a mean outcome or base rate from a reference class of similar uncertain events. In our setting, long-run mean air pollution in Lahore would be a reasonable base rate. The inside view incorporates information particular to the event being forecast, like the probability of rain tomorrow. Subjects were taught how to choose a good reference class and warned of the tendency to give too much weight to the inside view.

Next subjects learned to distinguish fast- and slow-changing systems prior to forecasting. They were encouraged to explicitly assess the accuracy of a source before using its information as a forecast input. The following exercise taught subjects an important heuristic for forecasting time series: they were instructed to consider a history at least as long as the time horizon of the forecast task. That is, to forecast three days ahead one should consider at least three days of history.

Subjects then completed an exercise that encouraged them not to round their forecasts excessively. In the final set of exercises, subjects were asked to reflect on an earlier forecasting task and had the opportunity to change their previous forecasts. This taught subjects to slow down and to engage "System Two" in the language of Kahneman [2011].

All exercises involved the active participation of subjects and were followed by clear feedback. The training was designed to be general: the large majority of the lessons and examples did not involve air pollution. Sessions were relatively brief, with an average duration of 51 minutes.[30]

3.2 Primary outcomes

Endline surveys were conducted in person in subjects' homes ten months after baseline, and measured five primary outcomes.[31]

1. **Air pollution forecast errors.** We asked respondents to forecast average $PM_{2.5}$ levels at $t+1$ and $t+3$ in their neighbourhood.[32] We incentivized the forecasts by

[29] Urdu is one of the primary local languages spoken in Lahore. Training was in-person rather than online to reduce barriers to participation and increase take-up.

[30] The standard deviation of training duration was 15 minutes.

[31] Baseline surveys were in April-May 2019. Endline surveys were in January-February 2020, prior to the outbreak of Covid-19 in Pakistan.

[32] Absolute-value forecast errors were standardized by subtracting the mean and dividing by the control-group standard deviation at each time horizon $t+1$ and $t+3$.

offering payments for responses within 5, 10, and 20 percent of realised $PM_{2.5}$ levels. This outcome allows us to examine expectation formation—that is, do our treatments improve respondents' ability to forecast? Just before providing an air pollution forecast, subjects were asked if they wanted to view a weather forecast (at no cost). Subjects who answered yes were shown a weather forecast for the target date ($t + 1$ or $t + 3$) on a tablet computer, and then proceeded to make their incentivized air pollution forecast. Weather forecasts are potentially relevant because, for example, rain greatly reduces particulate pollution. This secondary feature of the experiment was designed to evaluate whether treatment would affect take-up and use of relevant information.

2. **Willingness to pay for a particulate-filtering face mask.** We elicited respondents' willingness to pay for a air pollution mask (N95) using a Becker-DeGroot-Marschak (BDM) mechanism [Becker et al., 1964], with the price in PKR drawn from a uniform distribution on the interval $[0, 200]$.[33] This outcome allows us to measure adaptive behavioural response—that is, do our treatments increase respondents' valuation of an avoidance good?

3. **Air pollution avoidance index.** We asked respondents to report (yes or no) whether in the past week they: (i) reduced the number of hours spent on non-work outdoor activities; (ii) reduced the number of hours worked significantly; or (iii) rescheduled activities across days in response to poor air quality. We indexed these responses into a single measure. This outcome offers an additional dimension of adaptive behavioural response—that is, do our treatments alter respondents' time allocations in ways that reduce air pollution exposure?

4. **Happiness variance.** On a five-point Likert scale, we asked respondents to report *"how variable has [their] level of happiness been from day to day over the past week."* This measures whether our treatments help subjects to better smooth subjective well-being across days.

5. **Willingness to pay (WTP) for pollution forecasts.** We elicited respondents' willingness to pay for a 90-day subscription to our $PM_{2.5}$ forecast SMS service. We used a BDM mechanism, drawing the price in Pakistani Rupees (PKR) from a uniform

[33] We used BDM auctions to elicit WTP for our SMS forecast service and a particulate filtering mask (N95). In each auction respondents received a monetary endowment equal to the maximum possible bid (200 PKR). A respondent who won—i.e., their bid price was higher than the randomly drawn price in the interval $[0, 200]$—received the forecasts (or mask), keeping the difference between their endowment and their bid. A respondent who lost kept their entire endowment. Before bidding on the forecasts or mask, subjects completed a practice BDM auction using real money and answered comprehension questions. Enumerators explained any errors in answering the comprehension questions.

distribution on the interval $[0, 200]$. This outcome allows us to measure demand for inputs to expectations.

Four of these five primary outcomes were also measured at baseline, for use as variance-reducing controls (see Section 4). WTP for pollution forecasts could not be elicited at baseline, as this would have required delivery of forecasts to winners of the BDM auction outside the group randomly assigned to receive forecasts.

4 Empirical strategy

This section explains our strategy for estimating causal effects of treatment. Meaningful deviations from the pre-analysis plan are described in Appendix H.3.

4.1 Intent to treat

We estimate effects within subject for the following primary outcomes: air pollution forecast errors, willingness to pay for a particulate-filtering mask, an index of air pollution avoidance, and self-reported happiness variance. The estimating equation is as follows.

$$Y_i = \beta_F Forecasts_i + \beta_T Training_i + \beta_{FT} Forecasts_i \times Training_i + \gamma Y_{0i} + \boldsymbol{X_i'}\boldsymbol{\delta} + \varepsilon_i \qquad (1)$$

In this equation i indexes subject and Y is the outcome.[34] $Forecasts_i$ denotes random assignment to SMS forecasts, and $Training_i$ random assignment to training. Y_0 is the baseline variable corresponding to the outcome Y. $\boldsymbol{X}$ is a vector of controls, including randomization block dummies. As pre-specified, other elements of $\boldsymbol{X}$ were chosen using post-double-selection LASSO applied separately to each primary outcome.[35]

Our pre-analysis plan anticipated power concerns under correction for multiple testing across eight primary estimates (discussed in Section 4.3). With such concerns in view, the plan pre-specified theoretically motivated one-tailed tests for some treatment-outcome combinations. For air pollution forecast errors, theory predicts that more information and better forecast training should both weakly improve forecast quality. The tests are one-tailed, against the alternatives $\beta_F < 0$, $\beta_T < 0$. The substitutability or complementarity of our two interventions is theoretically ambiguous, so the test of their interaction is two-tailed ($\beta_{FT} \neq 0$) for this and all other outcomes. We expect both treatments to improve subjects' ability to smooth utility over time, so tests in the model of self-reported happiness variance

[34] All treatment regressions include a constant term, but we omit it from most equations in this document in the interest of clarity.

[35] See Appendix H.2 for more discussion.

are one-tailed ($\beta_F < 0$, $\beta_T < 0$). Finally our model predicts that both treatments will increase avoidance when pollution is high (Hypothesis 2), so tests for mask demand and the avoidance index are against the following alternatives: $\beta_F > 0$, $\beta_T > 0$.

We estimate willingness to pay for 90 days of SMS forecasts between subjects.

$$Y_i = \alpha + \beta_F \mathit{Forecasts}_i + \beta_T \mathit{Training}_i + \beta_{FT} \mathit{Forecasts}_i \times \mathit{Training}_i + \varepsilon_i \tag{2}$$

Notation for outcomes and treatments is as in Equation 1. For willingness to pay for forecasts, our pre-specified hypothesis test takes the one-tailed form: $\alpha + \beta_F > 0$. That is, we test whether mean willingness to pay is positive among subjects in the SMS-forecast-only group.[36] This is the test that will be included in our multiple-testing correction procedure.[37]

4.2 Treatment on the treated

For the training arms ($Training_i = 1$) we observe participation in the training session ($P_{Ti} = 1$). For the forecast arms ($Forecasts_i = 1$) take-up means looking at our SMS forecast. This was not directly observable. Moreover it plausibly varied, both across individuals and within individual over time. As pre-specified, we construct a take-up measure using end-line survey responses to the question: "How many times in the last week have you seen our pollution forecast message?"[38] Denote the response of subject i as R_i.[39] Then a subject's take-up is defined as $P_{Fi} = \frac{1}{7} R_i$. This variable will range from zero to one, and can be interpreted as the fraction of forecasts taken up. While P_{Fi} is measured with error, in expectation this error has zero covariance with our random treatment assignment. Importantly, we also allow for take-up by those in our control group. In the end-line survey, we showed control respondents a picture of a forecast treatment SMS message and asked "Did you receive any LUMS air pollution text messages similar to these from someone else?"[40] If the respondent said yes, we followed up with "If yes, how frequently did this happen?" We estimate a frequency in the last week by dividing the reported (total) frequency by the number of weeks of the forecast intervention. Just 31 of 544 subjects (5.7 percent) outside the text message group reported receiving any of our pollution forecasts. Of these 31 subjects, 22 reported receiving one to

[36]Note that because randomization block dummies are not included in Equation 2, treatment effects are not identified and estimates of $\boldsymbol{\beta}$ should not be interpreted causally. The sum $\alpha + \beta_F$ is of research and policy interest even though it does not reflect causal effects of treatment.

[37]The hypotheses that willingness to pay among control subjects is positive $\alpha > 0$, that training affects willingness to pay $\beta_T \neq 0$, and that the treatments interact $\beta_{FT} \neq 0$, are interesting but secondary, as specified in our pre-analysis plan.

[38]This question was asked only of subjects assigned to the forecast treatment.

[39]Subjects who responded "not sure" were assigned $R_i = 0$.

[40]LUMS is the Lahore University of Management Sciences.

nine of our messages over the entire course of the study, and just nine reported receiving ten or more (Table A.22).

The interaction of take-up measures is simply $P_{FTi} = P_{Fi}P_{Ti}$. Effects of treatment on the treated are estimated using two-stage least squares (2SLS), with $\{Forecasts_i, Training_i, Forecasts_i \times Training_i\}$ instrumenting for $\{P_{Fi}, P_{Ti}, P_{FTi}\}$. Estimating equations appear in Appendix H.1. One- and two-tailed hypothesis tests for primary outcomes are analogous to those in our ITT regressions.

4.3 Multiple hypothesis testing

To address concerns of multiple hypothesis testing, we follow the procedures in Anderson [2008] and report Anderson sharpened q-values for our primary variables of interest. Sharpened q-values account for multiple hypothesis testing by using False Discovery Rate (FDR) adjustments as proposed by Benjamini et al. [2006]. While traditionally the process requires the researcher to specify the threshold "q-value" (i.e. the proportion of false rejections of the null in expectation), sharpened q-values allow us, like other recent papers,[41] to side-step this issue. Sharpened q-values are found by varying the false discovery rate and for each hypothesis reporting the smallest q-value for which the null is rejected. For example, a sharpened q-value of 0.1 means that a hypothesis would only be rejected if the false discovery rate is above 10%.

We adjust for potential false discoveries in the subset of alternative hypotheses related to our primary outcomes: willingness to pay for forecast information ($\alpha + \beta_F > 0$), air pollution forecast error at $t+1$ ($\beta_F < 0, \beta_T < 0$),[42] self-reported happiness variance ($\beta_T < 0$), willingness to pay for masks ($\beta_F > 0, \beta_T > 0$), and the avoidance index ($\beta_F > 0$, $\beta_T > 0$). The total count of included tests is eight. Note this is not an exhaustive list of hypotheses involving treatment effects on our primary outcomes. As pre-specified, where a test is less interesting we exclude it from the adjustment procedure.

[41] See for example Banerjee et al. [2015], Bryan et al. [2021] and Ahmad et al. [2024].

[42] Note that in the PAP we specified an index that combined forecast errors at $t+1$ and $t+3$ as a primary outcome. Following the recommendation of referees, we instead use the more intuitive measure of forecast errors at $t+1$. Details of this change and the original results table are provided in Appendix H.3.

5 Results

5.1 Primary outcomes, intent to treat

Four of our four primary hypotheses pertain to regression estimates of intent-to-treat effects, which are presented in Table 1.[43] Column headings indicate dependent variables. As described in Section 4, we pre-specified a one- or two-tailed test at the outcome-treatment level. The resulting p-values appear in square brackets. For pre-specified primary hypotheses we also report Anderson sharpened q-values in curly brackets, to account for multiple hypothesis testing as discussed in Section 4.3.

Column 1 presents estimates for $t+1$ forecasting error. This is our primary outcome in the domain of expectation formation. Provision of SMS forecasts reduced forecast error by .11 standard deviations, while training reduced it by .15 standard deviations. Figure 1 illustrates these error reductions in levels; a detailed discussion follows in Section 5.5.1.[44] Subjects in the SMS forecast group had not yet received the next day's forecast message at the time they made their own incentivized forecasts, so the reduced error is not a mechanical consequence of treatment.[45] Instead the negative treatment effect for this group is consistent with learning about the data-generating process for air pollution over the course of the experiment. The negative effect of training on forecast error is consistent with increased forecasting-relevant human capital. We also present p-values for a test of equality of the SMS and training effects in Table 1. We cannot reject that the two treatment effects on forecast error are the same, nor can we do so for our remaining primary outcomes.

The interaction effect on $t+1$ air pollution forecast error is positive (column 1 of Table 1), so the effect on the group that received both treatments was $-.11 - .15 + .13 = -.13\sigma$ (p-value of 0.014). The overall effect of receiving both treatments is consistent with net substitutability of information and human capital in the production of forecasts. A similar pattern obtains in all columns of Table 1, with estimated interaction effects taking the sign opposite that of the forecast and training effects. Our data do not speak to the sources of this substitutability. Potential explanations include crowd-out of training by recent, salient SMS

[43] Table A23 reports primary ITT results without controlling for baseline measures of the outcomes, and without the controls selected by the post-double-selection LASSO.

[44] Note the confidence intervals associated with one-tailed hypothesis tests are also one-sided. In the case of the effect of training on forecast error, as an example, the associated 95% confidence interval is $(-\infty, -0.063)$. The upper bound is calculated as $-0.15 + (1.645 * 0.053)$. That is to say, with our study's power, we can confidently rule out all positive treatment effects and very small negative effects but we cannot rule out much larger negative effects.

[45] That is, subjects were not in a position to simply parrot the prediction of our forecast model because the relevant message arrived in the evening, after all endline surveys had been completed. Furthermore, we find that only one percent of forecast-group subjects simply repeated the forecast for the previous day.

forecasts and constraints on recall or cognition. As treatment interactions were not the focus of our experimental design—none were included in our pre-specified primary outcomes—we do not discuss them further.

Column 2 reports effects on $t+3$ air pollution forecast error. Point estimates indicate that both treatments reduced error much less at $t+3$ than $t+1$, though we cannot reject a null hypothesis of equality. Intriguingly, the relative advantage of the training treatment is greater at the longer time horizon. One day ahead, training reduces error by 31 percent more than SMS forecasts do. Three days ahead, training reduces error by 99 percent more.[46] Given the large standard errors, we do not make strong claims about this pattern. It could reflect the fact that our SMS messages contained forecasts for $t+1$ but not $t+3$. Over the period they received messages, subjects might have learned lessons about forecasting one day ahead that proved unhelpful or even counterproductive when forecasting three days ahead. In contrast, the training treatment was designed to be general-purpose and produced practically meaningful reductions in error at both time horizons.

The reductions in $t+1$ forecast error from forecast provision and training are practically large. Estimating effects in concentration rather than standard deviations (Table A7, Column 1), the treatments reduced forecast error by 6.2 to 8.4 $\mu g/m^3$, or 9.5 to 13 percent of the control mean. The WHO 24-hour standard for $PM_{2.5}$ is 15 $\mu g/m^3$, so the marginal effects of forecasts and training are roughly 50 percent of the maximum healthy level.[47] The $.15\sigma$ reduction from forecast training is particularly remarkable, as our endline surveys took place four to six months after the training sessions. This suggests that our relatively brief sessions—average duration was 51 minutes—produced durable improvements in subjects' forecasting ability.[48] Table A7, Column 2, presents ITTs on mean squared forecast error, which are larger as a proportion of the control mean than ITTs for absolute errors. This is because treatment both improved the mean and lowered the variance of errors.

Comparisons to other studies in which experimenters designed treatments to reduce forecast error require care, owing to differences in setting, time horizon, and forecast scoring. Mellers et al. [2014] found that probability training improved mean standardized Brier score—a measure of forecast skill—by roughly $.1\sigma$. The improvement persisted over two years. Following the same annual training intervention over four years, Chang et al. [2016b] found a 6 to 12 percent improvement in Brier scores, again roughly similar to our estimated effects. While the participants belonged to many countries, they all had bachelor's degrees,

[46] These percentage changes use midpoints as bases.

[47] Both the United States and the European Union employ more stringent standards. Average $PM_{2.5}$ levels during endline surveys were 147 $\mu g/m^3$. As a proportion of this level, the 8.4 $\mu g/m^3$ error reduction is 5.7 percent.

[48] The standard deviation of training duration was 15 minutes.

and two thirds had graduate degrees. The probability training of Mellers et al. [2014] and Chang et al. [2016b] contained substantially more material and more complex tasks than ours. We find that a shorter, simpler training, conducted with less educated subjects, yielded a coarsely similar improvement in forecast performance for air pollution.

Column 3 of Table 1 presents effects on the variance of happiness, as reported by subjects on a five-point Likert scale.[49] Larger values correspond to higher variability. Estimated effects are small and not statistically distinguishable from zero. These coefficients potentially reflect both small or null treatment effects on this outcome and the measurement problems that attend questions of this type [Bond and Lang, 2019]. Note that the sample size in column 2 is 951, rather than 999 as in the other columns of Table 1. Here and throughout the paper, sample sizes less than 999 reflect non-response.

Column 4 reports effects on willingness to pay for N95 particulate-filtering masks.[50] The SMS forecast intervention increased WTP by 6.58 PKR.[51] The estimated effect of training is also positive at 3.95 PKR, but less precise. These positive estimates are consistent with Hypothesis 2 from the theoretical model in Section 2. That is, treated subjects may have higher WTP for masks because their forecasts of high pollution are more likely to be accurate.[52] More generally, better forecasts enable subjects to wear masks on the high-pollution days when they are most needed and conserve masks on less-polluted days. Estimated coefficients for the avoidance index are similarly positive (column 5), but are not statistically significant for either treatment. The results for mask demand are consistent with the finding of Ito and Zhang [2020] that willingness to pay for air purifiers increased in China after the US embassy in Beijing began posting air pollution readings. Our findings are also qualitatively consistent with studies of behaviours related to water pollution in developing countries. Madajewicz et al. [2007] found a large increase in the probability of switching wells when they informed households of arsenic contamination, while Jalan and Somanathan [2008] found that informing households of fecal water contamination led them to begin purifying their water.

---

[49]The question at endline was, "How variable has your level of happiness been from day to day over the past week?" At baseline, we asked "How variable has your level of happiness been over the past month?" While these questions are not identical, we use this baseline measure as a control to improve precision.

[50]Our endline survey concluded prior to the outbreak of the Covid-19 pandemic. At baseline, the maximum bid was 150 PKR. Despite this difference in censoring, we employ baseline WTP as a control corresponding to Y_{0i} in Equation 1.

[51]As explained in Appendix H.3, our pre-analysis mistakenly specified a two-tailed test for this coefficient. Table 1 reports a one-tailed p-value congruent with Hypothesis 2 of our theoretical model (Section 2). The two-tailed p-value is .06. Table A8 reports an alternative set of sharpened q-values in which tests on willingness to pay for masks and the avoidance index are two-tailed.

[52]We can rule out the possibility that the effect on mask demand stems from learning about mean pollution, as both treatment and control subjects received this information at baseline. See Section 3 for more details.

5.2 Outdoor time use and WTP for forecasts, intent to treat

In this section we evaluate additional hypotheses from our theoretical model (Section 2). The first two are related to avoidance behaviour. Column one of Table 2 presents SMS forecast effects on time spent outside on the day before the end-line survey, pooling over adults and children. We elicited outdoor time use through time diaries, asking subjects to describe their activities—and whether the activities occurred outdoors or indoors—for each hour of the day.[53] Such diaries are considered best practice in time-use research because they minimize reporting biases, including recall bias and experimenter demand [Seymour et al., 2017, Field et al., 2022, Giménez-Nadal and Molina, 2022].

We focus on the SMS treatment rather than training because our SMS forecasts varied over the course of the end-line survey and training did not, but results are robust to estimating the full suite of treatment effects (Table A18). On relatively cleaner days, with our forecast of particulate pollution below 150 micrograms per cubic meter, subjects receiving SMS messages increased outdoor time by .74 hours, or 16 percent of the control-group mean. The 150-microgram value was chosen because it was the threshold for the most polluted category of days in the pamphlet provided to all subjects (including control subjects; see Section 3). Because the end-line surveys were conducted during a high-pollution time of year (January-February), all of our SMS forecasts were in either the highest- or second-highest pollution category. On relatively more polluted days, subjects receiving SMS forecasts spent slightly less time outdoors than control subjects. That is, the sum of the "Forecasts" and "Forecasts * High pollution" coefficients is negative ($.74 - .88 = -.14$) and three percent of the control mean, but one cannot reject a hypothesized zero marginal effect. When presented with a relatively good forecast during a bad season for air pollution, SMS-treated subjects took advantage and control subjects did not. This is consistent with Hypothesis 3. When presented with a relatively bad forecast, SMS-treated subjects avoided slightly more than control subjects. Again this is consistent with Hypothesis 3, but we emphasize that the estimate is imprecise.

Column two of Table 2 presents heterogeneous treatment effects by whether subjects care about air quality.[54] At baseline 85 percent of subjects reported caring. This dimension of heterogeneity was not pre-specified and results should be interpreted cautiously. Having said

[53] Subjects completed 24-hour time diaries for both themselves and the youngest physically active child in their household. The estimating equation is $Y_i = \beta_F Forecasts_i + \beta_H High\,pollution_t + \beta_{FH} Forecasts_i High\,pollution_t + \gamma Y_{0i} + \boldsymbol{X_i'}\boldsymbol{\delta} + \varepsilon_i$. $High\,pollution_t$ is a dummy for a high air pollution forecast (fine particulate concentration above 150 $\mu g/m^3$) on the day of the subject's end-line time diary (the day before the end-line survey). As elsewhere in the paper, baseline controls in $\boldsymbol{X}$ were chosen using post-double-selection LASSO.

[54] The estimating equation adds a triple interaction with an indicator for caring about air quality; this indicator also enters in non-interacted and double-interacted control terms.

that, the indicator for caring co-varies with other attributes in ways that suggest it is not merely cheap talk. Covariances with baseline avoidance, end-line reports of viewing SMS forecasts, and end-line demand for SMS forecasts are all positive and statistically significant; the covariance with baseline demand for masks is positive but imprecisely estimated (Table A19). For subjects who reported caring, the pattern of signs in column two of Table 2 is similar to that of column one. On days with lower forecast pollution, SMS-treated subjects who care about air pollution spent three additional quarters of an hour outside, relative to the control group. On days with higher forecast pollution, they spent one quarter of an hour less outside.[55] Broadly the results in column two are consistent with Hypothesis 3.

Columns three through six of Table 2 repeat the specifications of the first two columns separately for adults and children. Broadly the patterns of signs for adults and children are similar to the pooled estimates.[56] The smaller samples reduce precision, especially for children, preventing us from making strong statements about relative magnitudes. Bearing that caveat in mind, the point estimates are consistent with greater avoidance among forecast-treated children ($.60 - 1.08 = -.48$ hours in column five) than adults ($.60 - .45 = -.15$ hours in column three). Similarly the marginal effects of SMS forecasts on high-pollution days in households that care about air quality are consistent with more avoidance among children than adults.[57]

Our experiment produced no direct evidence on whether the changes in outdoor time seen in Table 2 reduced pollution exposure.[58] We cannot exclude the possibility that subjects were making mistakes, particularly if pollution was high inside their home or workplace. But staying indoors can be an effective air pollution avoidance strategy. Levels of some pollutants, e.g. ozone, are generally much lower indoors [US Centers for Disease Control and Prevention, 2022] and indoor activities often involve less physical exertion [Laumbach, 2010]. Our time-use findings are consistent with Barwick et al. [2019], which finds that credit-card transactions outside the home decline with higher air pollution after the rollout of real-time pollution information in China.

---

[55] Among subjects who care about air pollution, the marginal effect of SMS forecast treatment is $.10 + .64 = .74$ hours on cleaner days and $.10 + .64 + 1.22 - 2.21 = -.25$ hours on dirtier days. The latter marginal effect is not statistically significant at any conventional threshold.

[56] Note that the pooled coefficient on "Forecasts" in column one is not a convex combination of the corresponding adult- and child-specific coefficients because of our pre-specified LASSO procedure for control selection. Appendix Table A20 shows that without LASSO-selected controls, the pooled coefficient is a convex combination of those for adults and children. Patterns of signs and magnitudes are qualitatively unchanged, though precision is predictably reduced.

[57] The marginal effects are $.86 + 1.07 - .10 - 2.40 = -.57$ for children and $-.41 + 1.53 + 1.17 - 2.16 = .13$ for adults.

[58] More generally, the welfare effects of the changes in Table 2 depend on exposure changes, the shape of the air pollution damage function at high levels of exposure, and subjects' valuation of time use changes.

Finally, our theoretical model delivers two predictions related to willingness to pay for continued receipt of our SMS pollution forecasts. The first is that willingness to pay will be higher among subjects who have received the (free) SMS forecast treatment. The corresponding regression estimate (Table A6, column two) is positive 5.3 PKR and large in proportional terms, consistent with Hypothesis 4, however, it is not statistically significant (two-tailed $p = .14$). The estimated interaction effect of the forecast and training interventions is negative. This negative sign implies substitutability of training (human capital) and information in subjects' forecast production functions. We caution, however, that this estimate is imprecise and the associated 95 percent confidence interval includes practically meaningful values on both sides of zero.

5.3 Demand for air pollution forecasts

Our final primary hypothesis pertains to demand for 90 additional days of our SMS air pollution forecasts. As pre-specified, our analysis focuses on subjects exposed only to the forecast treatment. Forecasts are plausibly an experience good, and these subjects' demand reflects months of interaction and learning. This informed demand constitutes the relevant estimand for a policymaker contemplating distribution of government forecasts and conducting a benefit-cost analysis. Figure 2 Panel A presents a histogram of willingness to pay (WTP) for this group. There is evidence of round-number heaping, particularly at multiples of 10 and 50. Vertical lines indicate the mean at 93.22 PKR and the median at 100 PKR. Roughly two percent of respondents in this group bid the maximum of 200 PKR and their willingness to pay is potentially censored. This implies that true mean willingness to pay is weakly greater than our reported value.[59] In a right-tailed test against a zero null hypothesis $p = .000$ (see Table A6). This is consistent with Hypothesis 1 from the model in Section 2, that willingness to pay for useful forecasts is non-zero. On a monthly basis, mean WTP of 93 PKR represents roughly 41 percent of total mobile phone costs in Lahore.[60] Considering a different benchmark, 93 PKR is approximately 20 percent of a day's earnings for an unskilled labourer. Section 5.7 aggregates WTP over Lahore and compares the resulting figure to monitoring costs. Under the assumption that our forecasts provide no direct utility (as in the theoretical model of Section 2),[61] mean WTP can be interpreted as the expected welfare gain from additional avoidance facilitated by the information. To the extent that

[59] Similarly, if some subjects expected to receive forecasts free from others after endline, that would have reduced their observed WTP, and again our estimated mean WTP is a lower bound.

[60] Table 22 of Pakistan Bureau of Statistics [2017] gives monthly per capita communications expenditure in the third quintile at 75.62 PKR. Dividing our WTP estimate by three gives a monthly WTP of 31.07 PKR. As a proportion of communications expenditure this is 41 percent.

[61] By "no direct utility" we mean that subjects do not derive satisfaction from the forecast itself, even without acting on it.

avoidance behaviour is constrained, e.g. by conditions of employment, forecast demand will be constrained as well. If expectations remain biased, forecast demand will not reflect all available welfare gains from avoidance. Figure 2 Panel B presents the same underlying WTP responses as a demand curve. The average elasticity of quantity demanded—expressed here as the share of subjects purchasing—with respect to price is -.93.

In our low-income subject population, finding an appreciably positive mean willingness to pay was by no means obvious *ex ante*. Barnwal et al. [2017] discovered low and elastic demand for arsenic testing of wells in Bihar, India. More broadly, a large body of work in development economics has revealed both low and strongly elastic demand for preventative health care [Kremer and Glennerster, 2011]. This suggests that demand for health information (or information complementary to health care) may share features with demand for other preventative measures, like insecticide-treated nets and water treatment.

The relatively high willingness to pay for air pollution forecasts may stem from several factors. First, because we delivered the forecasts by text message, subjects did not face the takeup barriers in time, distance, and inconvenience identified by studies like Thornton [2008] and Kremer et al. [2011]. Second, many previous studies have not used BDM elicitation. Finally, differences in setting may be important. Studies like Kremer et al. [2011] have examined rural populations, while ours is urban. Air pollution is a salient issue in Lahore because of its severity. While our results may not transfer to settings like Accra or Santiago—which experience substantially better air quality—they potentially shed light on cities with air pollution similar to Lahore's (for example, New Delhi and N'Djamena) and on past periods of acute fine particulate pollution in cities like Beijing. Section 5.7 discusses the external validity of our forecast demand estimates in greater depth.

5.4 Primary outcomes, effect of treatment on the treated

Table 3 reports estimated effects of treatment on the treated, instrumenting for take-up with treatment assignment as described in Section 4.2. First-stage F statistics are far above relevant critical values. Pre-specified LASSO control selection and other details are just as in Table 1. Some 98 percent of subjects assigned to training took up training, so TOT effects are not meaningfully different from their ITT counterparts.[62] Subjects receiving text messages viewed them slightly less than half the time, so TOT estimates are roughly twice as large as their ITT counterparts. As a result, the relative magnitudes of effects from the two treatments are reversed. Among perfect compliers, text messages reduced

[62] The TOT effects of training in columns 1-3 of Table 3 are slightly smaller in magnitude than the corresponding ITT effects in Table 1 because the double-selection LASSO algorithm chooses a different set of controls.

air pollution forecast error by more, and increased willingness to pay for masks by more, than did forecasting training.[63] Perfect compliers in the text message group increased their willingness to pay for masks by approximately 14 percent of the control-group mean.[64]

Of interest in Table 3 is the fact that the combined treatment effect of forecasts and training on willingness-to-pay for masks is near-zero. Adding coefficients, for those who received both treatments, the total effect is $14.9 + 3.92 - 18.4 = .42$. One possible explanation is that receiving both treatments creates a greater ability to avoid air pollution inter-temporally and thus lowers demand for other avoidance goods (i.e. wearing masks and staying indoors could be substitutes). Table A9 provides evidence consistent with this explanation. When we subset our analysis to those respondents who report an at-or-above-median preference for masks at baseline,[65] we see a stronger effect of SMS forecast take-up on WTP for masks and no correspondingly large negative coefficient on the interaction with training take-up. The combined treatment effect for this group is $20.3 + 3.23 - 13.6 = 9.93$. While such ex-post analysis should be interpreted cautiously, it suggests that respondents who particularly dislike masks may substitute to other avoidance behaviour, while those who prefer masks exhibit more consistent demand.

5.5 Mechanisms

5.5.1 Sources of reduced air pollution forecast error

Figure 1 investigates *how* our interventions reduced error in one-day-ahead forecasts. For the control group and each treated group, a separate probability density function is estimated over $t + 1$ forecast error. Unlike in most of this paper's exhibits, in Figure 1 errors are denominated in $\mu g/m^3$ rather than control-group standard deviations; no absolute value operator is applied. At endline control subjects under-predicted pollution substantially, by 40.2 $\mu g/m^3$ on average. If subjects face convex pollution-damage and abatement-cost functions, as hypothesized in our theoretical model, then such underprediction is more costly than overprediction of similar magnitude. As endline surveys took place during a high-pollution season (January-February), these prediction errors are plausibly consequential for health and well-being.[66] In contrast, the distributions for the treated groups are shifted

[63] We cannot reject a null hypothesis that the TOT effects are equal in any column.

[64] By "perfect compliers" we mean subjects who viewed 100 percent of the SMS forecasts they received.

[65] Mask preference is an average of five dummy variables: the respondent has seen people wearing N95 masks before, the respondent believes N95 masks work, the respondent does not believe N95 masks are ugly, the respondent believes the masks are comfortable (after trying one on), the respondent knows of a convenient place to purchase N95 masks. The median is three of five dummies equal to one.

[66] At baseline average $t + 1$ forecast error was positive: subjects over-predicted pollution. This may have been because baseline surveys occurred during a relatively low-pollution season (April-May). Figure A3 illustrates the distribution of baseline $t + 1$ forecast errors.

rightward, indicating reduced underprediction. Dispersion is also reduced. Tables A10 and A11 quantify these differences in means and standard deviations, respectively. Treatment increased means (reduced underprediction) by 1.3 to 6.3 $\mu g/m^3$, with the largest change in the SMS-forecast group.[67] Treatment also reduced the standard deviation of errors by 2.7 to 14.3 $\mu g/m^3$, with the largest change in the training group.[68] This is apparent in Figure 1, where the height of the distribution function at the mode is much greater for the training group than for the others. Because our measures of forecast error (as in Table 1) are built from absolute values, effects on those variables reflect both the reduced underprediction and the reduced dispersion in the underlying (non-absolute, non-standardized) forecast errors.

Appendix Section B investigates two additional possible sources of reduced air pollution forecast error specific to those who received our training intervention: information seeking and processing (Section B.1) and additional training effects (Section B.2). We find suggestive evidence that trained subjects make better use of relevant (i.e. weather) information when forecasting. Trained subjects better internalize mean reversion after experiencing an outlier day for air pollution, and they better adjust forecasts for predictable day-of-the-week pollution effects than untrained subjects, both concepts taught during the training. At the same time, trained subjects are no better than the untrained at adjusting forecasts when rain is on the horizon (untrained subjects are already good at this) and they round their forecasts no less (one of the training modules stressed not rounding forecasts).

5.5.2 Analysis from the midline training intervention

If the training intervention genuinely improved forecasting ability, that should have been apparent not only at endline, but also immediately after completion of the training. Subjects made incentivized one-day-ahead air pollution forecasts at the beginning of the training session and again at the end, yielding two observations for each of 522 subjects who completed training. Recall that subjects received training in both the training-only and forecasts-plus-training groups. This allows us to estimate simple difference-in-differences models of forecast errors at $t+1$ and $t+3$ (Table 4).[69]

The effect of SMS forecast receipt on forecast error at $t+1$ (row one, column one) is negative. At the start of the training session, subjects who had been receiving SMS forecasts made better one-day-ahead forecasts than subjects who had not been. Because

[67] These estimates are not statistically significant at conventional thresholds.

[68] The reduction in standard deviation for the training group is statistically significant ($p = .03$), but reductions for the other groups are not statistically significant at conventional thresholds.

[69] The estimating equation is $Y_{it} = \beta_1 Forecasts_i + \beta_2 Post_t + \beta_3 Forecasts_i * Post_t + \boldsymbol{X_i'}\boldsymbol{\delta} + \varepsilon_{it}$, with i indexing subject and t period (beginning or end of the training session). As elsewhere in the paper, baseline controls in $\boldsymbol{X}$ were chosen using post-double-selection LASSO.

both treatments were randomized and the forecast-only subjects had not yet been treated at the start of the training session, this estimate can sustain a causal interpretation. The negative effect is consistent with subjects learning about air pollution (or more formally, the data-generating process) through exposure to SMS forecasts. The effects of SMS forecasts on $t+3$ forecast error (column 2) are imprecise and one can reject neither a zero null hypothesis, nor a null hypothesis of equality with the estimate for $t+1$.

At the end of the one-hour training, forecast errors fell in the training-only group–the "Post training" coefficient is the marginal effect on this group. Point estimates are negative in both columns, and statistically significant in column one. This is consistent with the training functioning as intended. In the forecasts-and-training group, though, there was little change from the beginning of training to the end. Summing the coefficients in the second and third rows gives the marginal effect of the "Post" variable on this group. These sums are quite close to zero, and one cannot reject a zero null hypothesis at any conventional threshold. As the "Post" variable was not randomly assigned, speaking strictly one cannot interpret these marginal effects as causal effects of training. The scope for confounding in the course of a one-hour training was quite limited, however, and subjects had little ability to influence the timing of the training sessions.

Broadly, subjects who had been receiving SMS forecasts started the training session performing better than those who had not. But over the course of the session, the other subjects caught up in terms of forecast error. One could interpret this as evidence of a ceiling on forecast accuracy, operating perhaps through memory or cognition. Viewed through the lens of the model in Section 2, Table 4 provides corroborating evidence that information and human capital are substitutes in subjects' forecast production functions. Some of these trained subjects attrited between training and endline. Table A12 presents the same analysis for the endline sample and results are strongly similar. Table A13 presents a complementary analysis in which we examine heterogeneity in our primary intent-to-treat effects on forecast error by baseline forecast error. While there is little evidence of heterogeneity for the forecasts-only and both-treatments groups, there is suggestive evidence that our lessons were less successful for training-only subjects with lower baseline errors (p-value 0.11).

5.5.3 Mask demand

To investigate the positive treatment effect of SMS forecasts on WTP for masks, Figure A4 presents demand curves by experimental group. From these curves one can see that the increase in mean WTP for the forecast group is driven primarily by increases in takeup at higher prices (100-200 PKR). Demand elasticity in the control group is -1.6. Demand is less responsive to price in the three treated groups, with elasticities ranging from -.9 to

-1.2. Note however that the local elasticities near the retail price—135 PKR at the time of our study—are greater at roughly -2.4 (Table A16). This implies that small price changes or subsidies could produce large changes in mask takeup. At the time of our study mask wearing was uncommon in Lahore, with 74 percent of subjects reporting at baseline that they had never seen other people wearing N95 masks. If social norms around mask wearing changed in response to Covid-19, demand curves could have changed as a consequence.

5.6 Robustness

5.6.1 Experimenter demand

One might worry that some subject responses, especially non-incentivized measures of air pollution avoidance, might have been influenced by perceived experimenter demand. That is, subjects might have said they took action to avoid air pollution, when in fact they did not, if they believed we hoped to increase avoidance. This tendency could have been exacerbated if subjects thought future interactions and payouts could depend on responses. We attempted to mitigate these effects in several ways. First, all of our enumerators were trained to distance themselves from the implementation of treatments and to act as unbiased observers, with no promises of future interactions. We also ensured endline enumerators were not those that were involved in inviting subjects to treatment or providing them forecast training. Second, we phrased questions to mitigate experimenter demand effects and relied heavily on incentive-compatible elicitations for our primary analyses. Third, we included a social desirability module in our endline survey, as in Crowne and Marlowe [1960] and recent studies such as Dhar et al. [2018]. From this module, we construct an index of social desirability and report treatment effects on this variable in Table A21. Point estimates are small and not statistically significant. Marginal effects on all three experimental groups are negative, suggesting that if anything our treatment reduced the propensity to give socially desirable survey responses.

In addition, we indirectly evaluate experimenter demand effects on willingness to pay for our SMS air pollution forecasts. First we evaluate heterogeneity in willingness to pay for SMS forecasts by subjects' baseline forecast error. If willingness to pay were driven largely by experimenter demand, we would expect no difference in willingness to pay across subjects with above- and below-median errors. If instead willingness to pay was driven by the value of information to subjects, then we would expect subjects with higher baseline error to exhibit greater demand. Figure A5 displays the latter pattern: a third-party forecast is more valuable to someone who cannot forecast well alone. While baseline error could be correlated with other subject characteristics, the observed demand heterogeneity is inconsistent with pure experimenter demand.[70] Second, note that willingness to pay in the control group was still considerable at 89 PKR (Table A6, column one; cf. 93 PKR in the forecast-only treatment group).

5.6.2 Spillovers

Given the ease of relaying our forecasts, spillovers might in principle be a concern for our text message forecast treatment. Our sampling was designed to mitigate these concerns by separating subjects in space (sampled households were at least five houses apart), but some social networks might have included both treatment and control subjects nonetheless. We also asked subjects not to share pollution forecasts outside their households.

We sought to measure those spillovers we could not eliminate. At endline, subjects in the control group were asked if they received our forecasts from someone else. Just 31 of 544 subjects (5.7 percent) outside the text message group reported receiving any of our pollution forecasts. Of these 31 subjects, 22 reported receiving one to nine of our messages, and just nine reported receiving ten or more; Table A.22 reports the complete set of spillover frequencies. This evidence on spillovers does not raise substantial bias concerns. In addition, we account for measured spillovers as a form of control non-compliance in our treatment-on-the-treated estimates in Section 5.4. Because spillovers were so infrequent, accounting for them produces minimal changes in our estimates.

5.7 External validity of demand for air pollution forecasts

How far does our estimated demand curve for air pollution forecasts generalize? A complete reply to this question would require similar experiments in other settings, but theory and descriptive evidence allow for thinking about external validity in a structured way. The value of new air pollution information plausibly depends on: 1) air pollution levels; 2) the information environment (sources and modes of dissemination); and 3) scope for, and welfare gains from, avoidance. For this discussion of external validity we will hold air pollution fixed at high levels. Highly polluted cities like Lahore attract both research and policy attention because the returns to new knowledge are arguably highest there.

In Appendix C we compare Lahore, which had the world's highest levels of fine particulate pollution in 2022, to the other 24 of the 25 most polluted cities on dimensions 2) and 3).[71] In this section, we predict WTP in Lahore using commonly available demographics.

---

[70] As an important aside, this heterogeneity is also inconsistent with subjects valuing SMS forecasts for reasons other than their information content, e.g. because the messages help subjects remain attentive to pollution.

[71] According to IQAir, the 25 cities with the highest levels of PM2.5 (fine particulates) are: Lahore,

Aggregating across the city, we calculate total WTP of roughly 3.6 billion PKR, or US$12.7 million. We then project WTP for the highly polluted comparison cities. The resulting means range from 70PKR to 107PKR (Table A26). Underlying household-level estimates have large standard deviations (12.3 to 28 PKR) suggesting that demand varies substantially within cities. These estimates should be interpreted cautiously, as large benefit transfer errors are possible: Kaul et al. [2013] document mean transfer error of 39 percent.

In summary, we find that our experimental setting in Lahore is reasonably similar to most of the world's highly polluted cities in its information environment and its history. Adjusting for population differences moves estimated willingness to pay only modestly. As a result, our estimates are useful inputs to research and benefit-cost analysis of new air pollution monitoring in the urban settings where those efforts matter most. In cities for which we calculate WTP, the resulting estimate is at least one order of magnitude larger than monitoring cost, and possibly two [Hussey et al., 2022]. Even a large benefit-transfer error would be unlikely to reverse the sign of a conventional benefit-cost analysis under these circumstances. In cities with relatively richer information environments, e.g. New Delhi with its 10 monitors [Central Pollution Control Board, 2022], our estimates are less relevant to expansion of the monitoring network, but they nonetheless speak to the value of the monitoring already in place. This may be important to governments considering the continuation of monitoring in the context of competing policy initiatives.

6 Conclusion

We show that increasing information and human-capital inputs allows developing-country urbanites to form more accurate expectations over an environmental threat: air pollution. Most strikingly, our one-hour forecast training reduced forecast error for incentivized predictions made up to six months later. This is consistent with the training building human capital that works against common prediction biases. Exercises of this type could be a useful complement to education and job training in the developing world. While our training was relatively expensive to administer, other work has demonstrated successful de-biasing from videos and video games, which scale much more cheaply [Morewedge et al., 2015]. The constituent lessons and exercises from our training could be delivered via such low-cost channels. More generally, our training results argue that assisting people in using information they already have is at least as important as delivering novel information [Hanna et al., 2014].

Hotan, Bhiwadi, Delhi (NCT), Peshawar, Dharbanga, Asopur, N'Djamena, New Delhi, Patna, Ghaziabad, Dharuhera, Baghdad, Chapra, Muzaffarnagar, Faisalabad, Greater Noida, Bahadurgarh, Faridabad, Muzaffarpur, Noida, Jind, Karagandy, Charkhi Dadri, and Rohtak [IQAir, 2023].

Exposure to information—pollution forecasts—also increased willingness to pay for protective masks. This suggests that in areas where mask-wearing is not yet commonplace, information provision could be an important spur to mask adoption and other adaptive behaviours. Our findings that mean WTP for masks is roughly 70 percent of the retail price and demand is locally elastic suggest that modest subsidies could produce large changes in takeup, with concomitant health benefits.

In addition, we present evidence of meaningful willingness to pay for air pollution forecasts among developing-country urbanites. This argues that the scarcity of environmental information in many developing countries does not stem from a lack of demand. While capital and operating costs for reference-quality air pollution monitors are considerable—the equipment for a single site typically costs more than US$20,000 [Hussey et al., 2022]—the level of demand we estimate indicates that the welfare gain from investments in air pollution monitoring and forecasting are likely an order of magnitude greater than the costs. This is plausibly true not only in Lahore, but also in other developing-country settings with high pollution, low information, and comparable or higher incomes.

Many developing cities combine high, variable air pollution with relatively sparse information and low stocks of human capital. Residents face considerable risk, not only from environmental threats, but also in domains from family to employment. While our experiment was not designed to measure the broad welfare effects of providing forecasts or training agents to produce more accurate expectations, they are plausibly considerable, and warrant future research.

References

Paloma Acevedo, Guillermo Cruces, Paul Gertler, and Sebastian Martinez. Living up to expectations: How job training made women better off and men worse off. Technical report, National Bureau of Economic Research, 2017.

Achyuta Adhvaryu, Namrata Kala, and Anant Nyshadham. Management and shocks to worker productivity. *Journal of Political Economy*, 130(1):1–47, 2022.

Sandra Aguilar-Gomez, Holt Dwyer, Joshua S Graff Zivin, and Matthew J Neidell. This is air: The "non-health" effects of air pollution. Technical report, National Bureau of Economic Research, 2022.

Husnain Ahmad, Sheheryar Banuri, and Farasat Bokhari. Discrimination in healthcare: A field experiment with Pakistan's transgender community. *Labour Economics*, 87:102490, 4 2024. doi: 10.1016/j.labeco.2023.102490.

Anna Alberini, Maureen Cropper, Tsu-Tan Fu, Alan Krupnick, Jin-Tan Liu, Daigee Shaw, and Winston Harrington. Valuing health effects of air pollution in developing countries: the case of taiwan. *Journal of Environmental Economics and Management*, 34(2):107–126, 1997.

Michael L Anderson. 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, 2008.

Maximilian Auffhammer. Quantifying economic damages from climate change. *Journal of Economic Perspectives*, 32(4):33–52, 2018.

Abubakr Ayesh. Burned agricultural biomass, air pollution and crime. *Journal of Environmental Economics and Management*, 122:102887, 2023. doi: 10.1016/j.jeem.2023.102887.

Laura A Bakkensen and Lint Barrage. Going underwater? flood risk belief heterogeneity and coastal home price dynamics. *The Review of Financial Studies*, 35(8):3666–3709, 2022.

Abhijit Banerjee, Esther Duflo, Nathanael Goldberg, Dean Karlan, Robert Osei, William Parienté, Jeremy Shapiro, Bram Thuysbaert, and Christopher Udry. A multifaceted program causes lasting progress for the very poor: Evidence from six countries. *Science*, 348 (6236):1260799, 5 2015. doi: 10.1126/science.1260799.

Prabhat Barnwal, Alexander van Geen, Jan von der Goltz, and Chander Kumar Singh. Demand for environmental quality information and household response: Evidence from well-water arsenic testing. *Journal of Environmental Economics and Management*, 86: 160–192, 2017.

Alan I Barreca, Matthew Neidell, and Nicholas J Sanders. Long-run pollution exposure and mortality: Evidence from the acid rain program. *Journal of Public Economics*, 200, 2021.

Geoffrey Barrows, Teevrat Garg, and Akshaya Jha. The health costs of coal-fired power plants in India. Technical report, IZA DP No. 12838, 2019.

Panle Jia Barwick, Shanjun Li, Liguo Lin, and Eric Zou. From fog to smog: The value of pollution information. Technical report, National Bureau of Economic Research, 2019.

Gordon M Becker, Morris H DeGroot, and Jacob Marschak. Measuring utility by a single-response sequential method. *Behavioral Science*, 9(3):226–232, 1964.

Michel Beine, Frederic Docquier, and Hillel Rapoport. Brain drain and human capital formation in developing countries: winners and losers. *The Economic Journal*, 118(528):631–652, 2008.

Yoav Benjamini, Abba M Krieger, and Daniel Yekutieli. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93(3):491–507, 2006.

Kelly C Bishop, Jonathan D Ketcham, and Nicolai V Kuminoff. Hazed and confused: the effect of air pollution on dementia. *Review of Economic Studies*, page 2188–2214, 2022.

Timothy N Bond and Kevin Lang. The sad truth about happiness scales. *Journal of Political Economy*, 127(4):1629–1640, 2019.

Gharad Bryan, James Choi, and Dean Karlan. Randomizing religion: The impact of protestant evangelism on economic outcomes. *The Quarterly Journal of Economics*, 136:293–380, 2021. doi: 10.3386/w24278.

Marshall Burke and Kyle Emerick. Adaptation to climate change: Evidence from us agriculture. *American Economic Journal: Economic Policy*, 8(3):106–140, 2016.

Fiona Burlig, Amir Jina, Erin M Kelley, Gregory Lane, and Harshil Sahai. The value of forecasts: Experimental evidence from india. *NBER Working Paper No. 32173*, 2024.

David Card, Pablo Ibarrarán, Ferdinando Regalia, David Rosas-Shady, and Yuri Soares. The labor market impacts of youth training in the dominican republic. *Journal of Labor Economics*, 29(2):267–300, 2011.

Central Pollution Control Board. Manual ambient air quality monitoring stations in the country (as on 15.09.2022). Technical report, Ministry of Environment, Forest, and Climate Change, 2022.

Tom Chang, Joshua Graff Zivin, Tal Gross, and Matthew Neidell. Particulate pollution and the productivity of pear packers. *American Economic Journal: Economic Policy*, 8(3): 141–69, 2016a.

Tom Chang, Joshua Graff Zivin, Tal Gross, and Matthew Neidell. The effect of pollution on worker productivity: Evidence from call center workers in china. *American Economic Journal: Applied Economics*, 11(1):151–72, 2019.

Welton Chang, Eva Chen, Barbara Mellers, and Philip Tetlock. Developing expert political judgment: The impact of training and practice on judgmental accuracy in geopolitical forecasting tournaments. *Judgment and Decision Making*, 11(5):509, 2016b.

Maureen L Cropper, Nathalie B Simon, Anna Alberini, Seema Arora, and PK Sharma. The health benefits of air pollution control in delhi. *American Journal of Agricultural Economics*, 79(5):1625–1629, 1997.

Douglas P Crowne and David Marlowe. A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, 24(4):349, 1960.

Janet Currie, Joshua Graff Zivin, Jamie Mullins, and Matthew Neidell. What do we know about short-and long-term effects of early-life exposure to pollution? *Annu. Rev. Resour. Econ.*, 6(1):217–247, 2014.

Diva Dhar, Tarun Jain, and Seema Jayachandran. Reshaping adolescents' gender attitudes: Evidence from a school-based experiment in india. Technical report, National Bureau of Economic Research, 2018.

Pascaline Dupas and Edward Miguel. Impacts and determinants of health levels in low-income countries. In *Handbook of Economic Field Experiments*, volume 2, pages 3–93. Elsevier, 2017.

Erica M Field, Rohini Pande, Natalia Rigol, Simone G Schaner, Elena M Stacy, and Charity M Troyer Moore. Understanding rural households' time use in a developing setting: Validating a low-cost time use module. Technical report, National Bureau of Economic Research, 2022.

Teevrat Garg, Maulik Jagnani, and Vis Taraz. Temperature and human capital in india. *Journal of the Association of Environmental and Resource Economists*, 7(6):1113–1150, 2020.

Teevrat Garg, Maulik Jagnani, and H Pullabhotla. Rural roads, farm labor exits, and crop fires. *American Economic Journal: Economic Policy, Forthcoming*, pages 1–81, 2023.

Dalia Ghanem and Junjie Zhang. Effortless perfection: Do chinese cities manipulate air pollution data? *Journal of Environmental Economics and Management*, 68(2):203–225, 2014.

Dalia Ghanem, Shu Shen, and Junjie Zhang. A censored maximum likelihood approach to quantifying manipulation in china's air pollution data. *Journal of the Association of Environmental and Resource Economists*, 7(5):965–1003, 2020.

Michael Gilraine and Angela Zheng. Air pollution and student performance in the us. Technical report, National Bureau of Economic Research, 2022.

José Ignacio Giménez-Nadal and José Alberto Molina. Time use surveys. *Handbook of Labor, Human Resources and Population Economics*, pages 1–18, 2022.

Xavier Giné, Robert Townsend, and James Vickery. Forecasting when it matters: evidence from semi-arid india. *Manuscript, FRB of NY*, 2015.

Joshua Graff Zivin and Matthew Neidell. Days of haze: Environmental information disclosure and intertemporal avoidance behavior. *Journal of Environmental Economics and Management*, 58(2):119–128, 2009.

Joshua Graff Zivin and Matthew Neidell. Environment, health, and human capital. *Journal of Economic Literature*, 51(3):689–730, 2013.

Joshua Graff Zivin and Matthew Neidell. Air pollution's hidden impacts. *Science*, 359(6371): 39–40, 2018.

Rema Hanna, Sendhil Mullainathan, and Joshua Schwartzstein. Learning through noticing: Theory and evidence from a field experiment. *The Quarterly Journal of Economics*, 129 (3):1311–1353, 2014.

Jiaxiu He, Haoming Liu, and Alberto Salvo. Severe air pollution and labor productivity: Evidence from industrial towns in china. *American Economic Journal: Applied Economics*, 11(1):173–201, 2019.

Travis Hussey, Amy Lemelin, and Matthew Lulofs. Maine dep low-cost pm sensor comparison. *Technical Report, Maine Department of Environmental Protection*, 2022.

IQAir. 2019 world air quality report. Technical report, IQAir Foundation, 2020.

IQAir. 2022 world air quality report. Technical report, IQAir Foundation, 2023.

Koichiro Ito and Shuang Zhang. Willingness to pay for clean air: Evidence from air purifier markets in china. *Journal of Political Economy*, 128(5):1627–1672, 2020.

Hanan G Jacoby and Emmanuel Skoufias. Risk, financial markets, and human capital in a developing country. *The Review of Economic Studies*, 64(3):311–335, 1997.

Jyotsna Jalan and Eswaran Somanathan. The importance of being informed: Experimental evidence on demand for environmental quality. *Journal of Development Economics*, 87(1): 14–28, 2008.

Marc Jeuland, Subhrendu K Pattanayak, and Randall Bluffstone. The economics of household air pollution. *Annu. Rev. Resour. Econ.*, 7(1):81–108, 2015.

Daniel Kahneman. *Thinking, fast and slow.* Macmillan, 2011.

Daniel Kahneman and Dan Lovallo. Timid choices and bold forecasts: A cognitive perspective on risk taking. *Management Science*, 39(1):17–31, 1993.

Daniel Kahneman and Amos Tversky. On the psychology of prediction. *Psychological Review*, 80(4):237, 1973.

Namrata Kala. Learning, adaptation, and climate uncertainty: Evidence from indian agriculture. *MIT Center for Energy and Environmental Policy Research Working Paper*, 23, 2019.

Namrata Kala, Clare Balboni, and Shweta Bhogale. Climate adaptation. *VoxDevLit*, 7:3, 2023.

Dean Karlan and Martin Valdivia. Teaching entrepreneurship: Impact of business training on microfinance clients and institutions. *Review of Economics and Statistics*, 93(2):510–527, 2011.

Sapna Kaul, Kevin J Boyle, Nicolai V Kuminoff, Christopher F Parmeter, and Jaren C Pope. What can we learn from benefit transfer errors? evidence from 20 years of research on convergent validity. *Journal of Environmental Economics and Management*, 66(1):90–104, 2013.

Christopher R Knittel, Douglas L Miller, and Nicholas J Sanders. Caution, drivers! children present: Traffic, pollution, and infant health. *Review of Economics and Statistics*, 98(2): 350–366, 2016.

Michael Kremer and Rachel Glennerster. Improving health in developing countries: evidence from randomized evaluations. In *Handbook of Health Economics*, volume 2, pages 201–315. Elsevier, 2011.

Michael Kremer, Jessica Leino, Edward Miguel, and Alix Peterson Zwane. Spring cleaning: Rural water impacts, valuation, and property rights institutions. *The Quarterly Journal of Economics*, 126(1):145–205, 2011.

Robert J Laumbach. Outdoor air pollutants and patient health. *American Family Physician*, 81(2):175, 2010.

Dan Lovallo, Carmina Clarke, and Colin Camerer. Robust analogizing and the outside view: two empirical tests of case-based decision making. *Strategic Management Journal*, 33(5): 496–512, 2012.

Malgosia Madajewicz, Alexander Pfaff, Alexander Van Geen, Joseph Graziano, Iftikhar Hussein, Hasina Momotaj, Roksana Sylvi, and Habibul Ahsan. Can information alone change behavior? response to arsenic contamination of groundwater in bangladesh. *Journal of Development Economics*, 84(2):731–754, 2007.

Claudia Martínez A, Esteban Puentes, and Jaime Ruiz-Tagle. The effects of micro-entrepreneurship programs on labor market performance: experimental evidence from chile. *American Economic Journal: Applied Economics*, 10(2):101–124, 2018.

David McKenzie and Christopher Woodruff. What are we learning from business training and entrepreneurship evaluations around the developing world? *The World Bank Research Observer*, 29(1):48–82, 2014.

Barbara Mellers, Lyle Ungar, Jonathan Baron, Jaime Ramos, Burcu Gurcay, Katrina Fincher, Sydney E Scott, Don Moore, Pavel Atanasov, Samuel A Swift, et al. Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*, 25(5):1106–1115, 2014.

Enrico Moretti and Matthew Neidell. Pollution, health, and avoidance behavior evidence from the ports of los angeles. *Journal of Human Resources*, 46(1):154–175, 2011.

Carey K Morewedge, Haewon Yoon, Irene Scopelliti, Carl W Symborski, James H Korris, and Karim S Kassam. Debiasing decisions: Improved decision making with a single training intervention. *Policy Insights from the Behavioral and Brain Sciences*, 2(1):129–140, 2015.

Jamie T Mullins and Prashant Bharadwaj. Weather, climate, and migration in the united states. Technical report, National Bureau of Economic Research, 2021.

Matthew Neidell. Air pollution, health, and socio-economic status: the effect of outdoor air quality on childhood asthma. *Journal of Health Economics*, 23(6):1209–1236, 2004.

Matthew Neidell. Air pollution and worker productivity. *IZA World of Labor*, pages 1–11, 2017.

NIPS and ICF. Pakistan demographic and health survey 2017-18. Technical report, National Institute of Population Studies (NIPS) Pakistan and ICF, Islamabad, Pakistan, and Rockville, Maryland, USA, 2019.

Pakistan Bureau of Statistics. Household integrated economic survey 2015-16. Technical report, Government of Pakistan, 2017. URL `http://www.pbs.gov.pk/sites/default/files//pslm/publications/hies15-16/TABLE_22.pdf`.

Dev Patel. Learning About a Warming World: Attention and Adaptation in Agriculture. *SSRN Electronic Journal*, 2023.

Anthony Patt, Pablo Suarez, and Chiedza Gwata. Effects of seasonal climate forecasts and participatory workshops among subsistence farmers in zimbabwe. *Proceedings of the National Academy of Sciences*, 102(35):12623–12628, 2005.

Claudia Persico. Can pollution cause poverty? the effects of pollution on educational, health and economic outcomes. Technical report, National Bureau of Economic Research, 2022.

Punjab Environmental Protection Department. Policy on controlling smog 2017, Oct 2017.

Ramsha Riaz and Khizar Hamid. Existing smog in lahore, pakistan: An alarming public health concern. *Cureus*, 10(1), 2018. doi: 10.7759/cureus.2111.

Mark Rosenzweig and Christopher R. Udry. Forecasting profitability. *NBER Working Paper*, 2013.

Mark R Rosenzweig and Christopher Udry. Rainfall forecasts, weather, and wages over the agricultural production cycle. *American Economic Review*, 104(5):278–83, 2014.

Paul Samuelson. Using full duality to show that simultaneously additive direct and indirect utilities implies unitary price elasticity of demand. *Econometrica*, 33(4):781–796, 1965.

Greg Seymour, Hazel Jean Malapit, and Agnes R Quisumbing. Measuring time use in development settings. *World Bank Policy Research Working Paper*, pages 1–40, 2017.

Jack B Soll, Katherine L Milkman, and John W Payne. A user's guide to debiasing. *The Wiley Blackwell Handbook of Judgment and Decision Making*, 2:924–951, 2015.

Joseph E Stiglitz. The contributions of the economics of information to twentieth century economics. *The Quarterly Journal of Economics*, 115(4):1441–1478, 2000.

Philip E Tetlock. *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton University Press, 2017.

Rebecca L Thornton. The demand for, and impact of, learning hiv status. *American Economic Review*, 98(5):1829–63, 2008.

US Centers for Disease Control and Prevention. Ozone and your health, 2022. URL https://www.cdc.gov/air/ozone.html.

World Bank. Education statistics (edstats): All indicators. Technical report / data overview, World Bank Group, Education Global Practice, 2017.

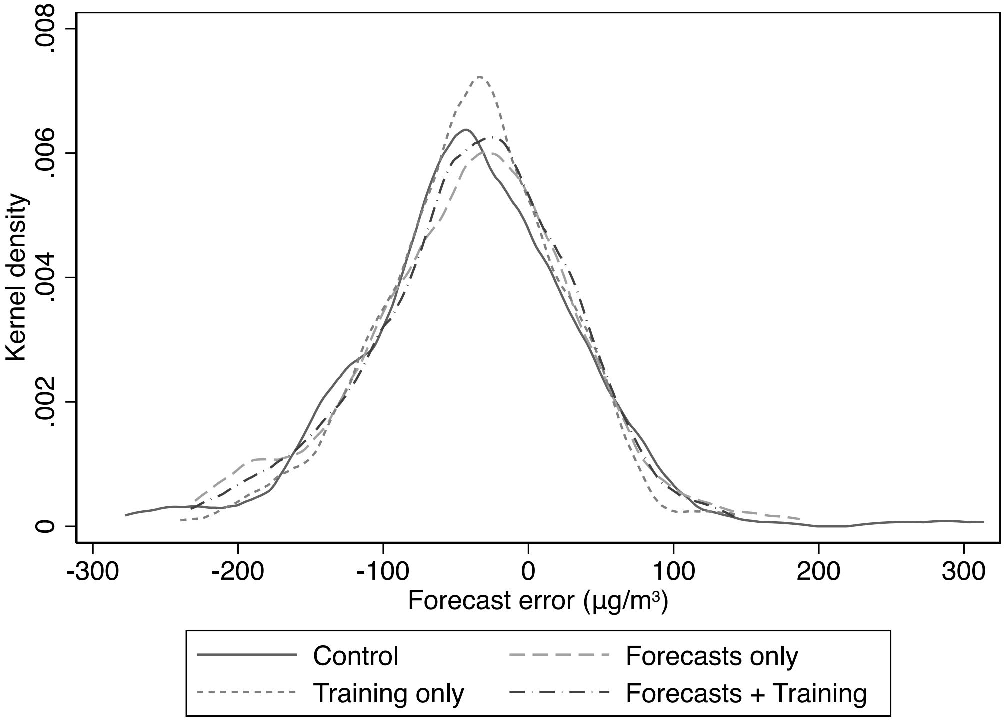
World Health Organization. Who air quality guidelines for particulate matter, ozone, nitrogen dioxide, sulfur dioxide, and carbon monoxide. Technical report, Geneva: World Health Organization, 2021.

Mehreen Zahra-Malik. In lahore, pakistan, smog has become a 'fifth season'. *The New York Times*, page A11, Nov 2017. URL https://www.nytimes.com/2017/11/10/world/asia/lahore-smog-pakistan.html.

Guglielmo Zappala. Adapting to climate change accounting for individual beliefs. *Journal of Development Economics*, 2024.

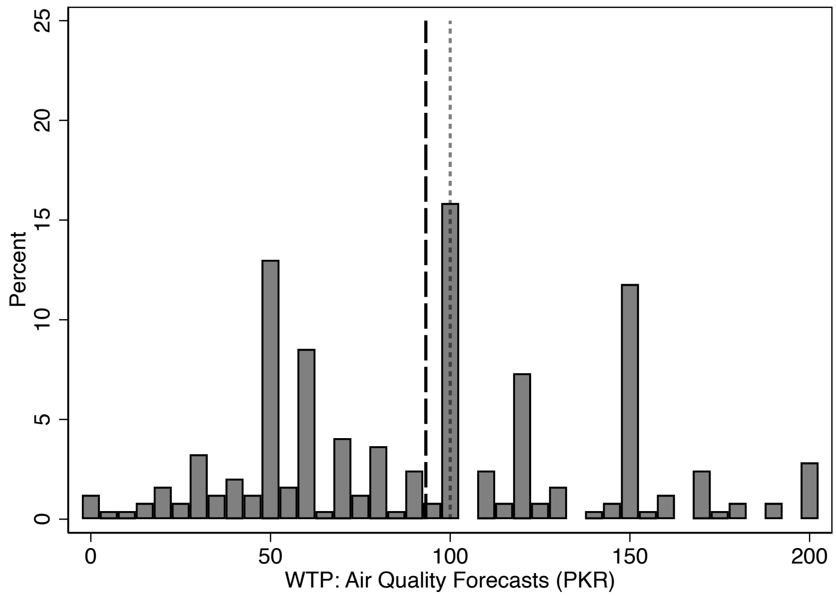
7 Figures

Figure 1: Air pollution forecast errors, t+1

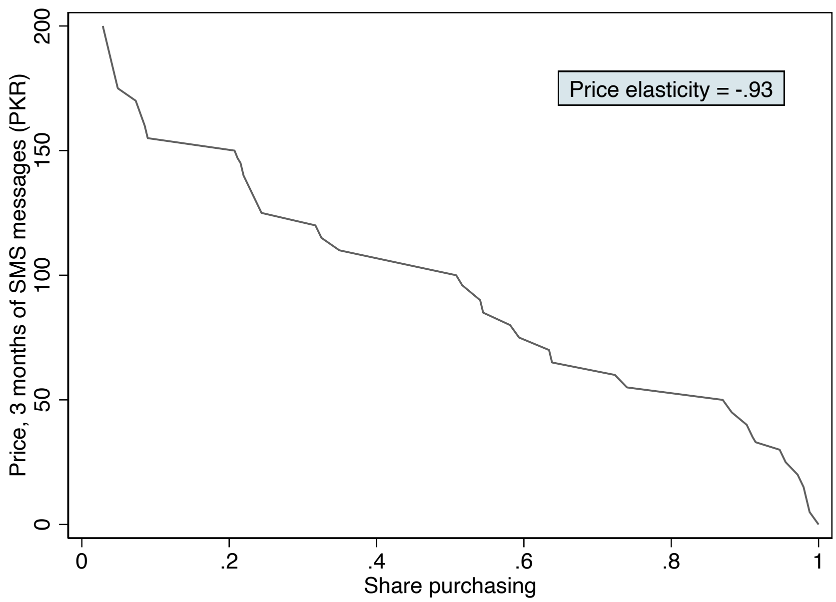


Note: Both treatments reduced mean under-prediction in air pollution forecasts. In addition, both treatments reduced the variance of forecast errors across subjects. Errors are the difference between subjects' incentive-compatible one-day-ahead (t+1) air pollution forecast and realised pollution on the day after the endline survey. That is, a negative error represents an underprediction of pollution. Units are $\mu g/m^3$, rather than control-group standard deviations as in most exhibits in this paper. Densities were estimated under Stata-default kernel and bandwidth.

Figure 2: Willingness to pay (WTP) for air pollution forecasts, forecast-only group



(a) Panel A: Histogram of willingness to pay



(b) Panel B: Demand curve for air pollution forecasts

Note: End-line willingness to pay (WTP) for air pollution information, specifically 90 additional days of our SMS air pollution forecasts (Section 3.1), was practically large. The vertical long-dashed line in Panel A marks the mean at 93.22 PKR, while the vertical short-dashed line marks the median at 100 PKR. For a formal hypothesis test of the mean against a zero null, see Table A6. In addition, Panel A illustrates the full distribution of WTP across subjects. Panel B expresses quantity demanded as the share of subjects purchasing; that is, the share with WTP greater than or equal to a given price. WTP was elicited at endline using a Becker-DeGroot-Marschak mechanism [Becker et al., 1964] with a maximum bid of 200 Pakistani Rupees (PKR). Both panels reflect the forecast-only treatment group (246 subjects), as explained in Section 5.1, because forecasts are plausibly an experience good.

8 Tables

Table 1: Primary outcomes, intent to treat

	Forecast error (t+1)	Forecast error (t+3)	Happiness variance	WTP: Masks	Avoidance index
Forecasts	-0.11	-0.022	0.052	6.58	0.046
	(0.058)	(0.056)	(0.070)	(3.53)	(0.059)
	[0.02]	[0.35]		[0.03]	[0.22]
	{0.05}			{0.05}	{0.17}
Training	-0.15	-0.065	0.078	3.95	0.019
	(0.053)	(0.057)	(0.071)	(3.54)	(0.059)
	[0.00]	[0.12]	[0.86]	[0.13]	[0.37]
	{0.01}		{0.42}	{0.12}	{0.27}
Forecasts + Training	0.13	0.070	-0.11	-7.58	-0.022
	(0.075)	(0.081)	(0.099)	(5.02)	(0.083)
Observations	999	999	951	999	999
Control mean	0.000	0.000	0.017	104.1	-0.0019
F = T p-value	0.44	0.44	0.71	0.47	0.65

Note: Both treatments reduced air pollution forecast error at t+1, and receipt of SMS forecasts increased willingness to pay for masks. Coefficients are intent-to-treat effects, with the dependent variable indicated in the column heading. Units are standard deviations for forecast errors, the variance of happiness, and the avoidance index. Units are Pakistani Rupees (PKR) for willingness to pay for masks. All columns include randomization block indicators. A pre-specified LASSO procedure was used to select additional controls separately for each outcome. Heteroskedasticity-robust standard errors are in parentheses. A pre-specified left-, right-, or two-tailed test was conducted for each estimate of interest: air pollution forecast error (t+1 and t+3) ($\beta_F < 0, \beta_T < 0$), self-reported happiness variance ($\beta_T < 0$), willingness to pay for masks ($\beta_F > 0, \beta_T > 0$), and the avoidance index ($\beta_F > 0$, $\beta_T > 0$). The resulting p-values appear in square brackets. For pre-specified variables of interest, Anderson [2008] sharpened q-values are reported in curly brackets. WTP for forecasts (not included in this table; see Fig. 2) was also a pre-specified primary hypothesis. It was accounted for in the calculation of q-values and the q-value for Forecast WTP is [0.001]. The F = T p-value corresponds to a t-test for equality of the coefficients on Forecasts and Training.

Table 2: Outdoor time, effect of receiving forecasts

	Outdoor hours					
Forecasts	0.74	0.10	0.60	-0.41	0.60	0.86
	(0.29)	(0.56)	(0.33)	(0.61)	(0.59)	(1.66)
Forecasts * High pollution	-0.88	1.22	-0.45	1.53	-1.08	1.07
	(0.36)	(0.99)	(0.41)	(0.97)	(0.74)	(2.88)
Forecasts * Cares about air quality		0.64		1.17		-0.10
		(0.65)		(0.69)		(1.86)
Forecasts * High pollution * Cares		-2.21		-2.16		-2.40
		(1.05)		(1.04)		(3.05)
Observations	1442	1442	980	980	462	462
Control mean	4.74	4.74	4.18	4.18	5.96	5.96
Adult and/or child time?	Both	Both	Adult	Adult	Child	Child
F = F * HP p-val	0.0094	0.42	0.14	0.18	0.19	0.96
F * C = F * HP * C p-val		0.064		0.036		0.61

Note: Subjects treated with SMS forecasts better matched their outdoor time to air pollution levels, increasing it on relatively cleaner days and decreasing it on relatively more polluted days. These effects were stronger among subjects who reported caring about air quality at baseline. The initial estimating equation (odd columns) is $Y_i = \beta_F Forecasts_i + \beta_H High\,pollution_t + \beta_{FH} Forecasts_i High\,pollution_t + \gamma Y_{0i} + \boldsymbol{X_i'\delta} + \varepsilon_i$. The dependent variable is outdoor time in hours, elicited as part of a 24-hour time diary. $High\,pollution_t$ is a dummy for a high air pollution forecast (fine particulate concentration above 150 $\mu g/m^3$) on the day of the subject's endline time diary (the day before the endline survey). Even columns add triple interactions with a baseline indicator for caring about air quality; this indicator also enters in non-interacted and double-interacted control terms. All columns include randomization block indicators. A pre-specified LASSO procedure was used to select additional controls, which also included weather controls, week, and day-of-the-week controls for the day of the time diary. Standard errors are clustered at the household level. The F = F * HP p-value corresponds to a t-test for equality of the coefficients on Forecasts and Forecasts * High pollution. The F * C = F * HP * C p-value corresponds to a t-test for equality of the coefficients on Forecasts * Cares about air quality and Forecasts * High pollution * Cares.

Table 3: Primary outcomes, effect of treatment on the treated

	Forecast error (t+1)	Forecast error (t+3)	Happiness variance	WTP: Masks	Avoidance index
% Forecasts seen	-0.27	-0.028	0.060	14.9	0.093
	(0.13)	(0.13)	(0.16)	(8.04)	(0.14)
	[0.02]	[0.41]		[0.03]	[0.25]
Attended training	-0.15	-0.042	0.059	3.92	0.026
	(0.056)	(0.060)	(0.074)	(3.79)	(0.062)
	[0.00]	[0.24]	[0.79]	[0.15]	[0.34]
% For. x Att. train.	0.30	0.14	-0.22	-18.4	-0.056
	(0.18)	(0.20)	(0.24)	(12.2)	(0.20)
Observations	999	999	951	999	999
Control mean	-0.00	-0.00	0.017	104.1	-0.0019
1st stage F-stat	173.6	173.6	168.2	174.0	171.5
F = T p-value	0.30	0.90	1.00	0.12	0.57

Note: Takeup of training (.98) was higher than takeup (viewing) of SMS forecasts (.43). As a result estimated TOTs for SMS forecasts are larger, relative to the ITTs, while estimated TOTs for training are similar to the ITTs. Coefficients are effects of treatment on the treated, with the dependent variable indicated in the column heading. Units are standard deviations for the forecast errors, the variance of happiness, and the avoidance index. Units are Pakistani Rupees (PKR) for willingness to pay for masks. All columns include randomization block indicators. A pre-specified LASSO procedure was used to select additional controls separately for each outcome. Heteroskedasticity-robust standard errors are in parentheses. A pre-specified left-, right-, or two-tailed test was conducted for each estimate of interest: air pollution forecast error (t+1 and t+3) ($\beta_F < 0, \beta_T < 0$), self-reported happiness variance ($\beta_T < 0$), willingness to pay for masks ($\beta_F > 0, \beta_T > 0$), and the avoidance index ($\beta_F > 0$, $\beta_T > 0$). The resulting p-values appear in square brackets. The F = T p-value corresponds to a t-test for equality of the coefficients on % Forecast seen and Attended training.

Table 4: Forecast errors, beginning and end of training

	Forecast error (t + 1)	Forecast error (t + 3)
Forecasts	-0.13	0.027
	(0.069)	(0.074)
Post training	-0.14	-0.051
	(0.054)	(0.046)
Forecasts * Post	0.17	0.024
	(0.064)	(0.065)
Observations	1044	1044
Control mean	-0.20	-0.30

Note: Both the training-only group and the forecasts-plus-training group were offered training. At the start of the training session, the forecasts-plus-training group produced smaller air pollution forecast errors ("Forecasts" coefficients). But by the end of the session, the training-only group caught up ("Post training" coefficients). The forecasts-plus-training group showed little change relative to the start of the session (summing coefficients on "Post" and "Forecasts * Post"). The sample is comprised of two observations for each of 522 subjects. The estimating equation is $Y_{it} = \beta_1 Forecasts_i + \beta_2 Post_t + \beta_3 Forecasts_i * Post_t + \boldsymbol{X_i'}\boldsymbol{\delta} + \varepsilon_{it}$, with i indexing subject and t period (beginning or end of the training session). Units are standard deviations in all columns. All columns include randomization block indicators. A pre-specified LASSO procedure was used to select additional controls separately for each outcome. Heteroskedasticity-robust standard errors are in parentheses.