Forecasts: Consumption, Production, and Behavioral Responses*

Husnain F. Ahmad, Matthew Gibson, Fatiq Nadeem,

Sanval Nasim, Arman Rezaee

July 2023

Abstract

Scarce information and human capital may make it difficult for residents of developing countries to produce accurate forecasts, limiting responses to uncertain future events like air pollution. We study two randomized interventions in Lahore, Pakistan: 1) provision of air pollution forecasts; 2) general training in forecasting. 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 increased the responsiveness of outdoor time to pollution. Forecast recipients were willing to pay 60 percent of the cost of mobile internet for continued access.

Keywords: forecasts, training, pollution avoidance, environmental information

JEL: Q56, Q53, D84, D90

*Ahmad: Department of Economics, Sewanee: The University of the South. Gibson: Department of Economics, Williams College, and IZA; mg17@williams.edu. 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 Chris Blattman, 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); Williams College and the University of the South. We would like to thank 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).

1 Introduction

Economic theory predicts that poor forecasts reduce welfare. An agent who relies on a biased forecast of high prices tomorrow may consume too little today. An agent who does not anticipate high temperatures may fail to reschedule physically intense outdoor work. Even well-informed experts commonly make forecasting mistakes (Tetlock, 2017). Residents of developing countries may face a substantially more difficult information environment, with relevant third-party forecasts often unavailable or of poor quality (Rosenzweig and Udry, 2014a,b). Moreover in developing countries, underlying behavioral biases may interact with information scarcity and lower levels of human capital (Stiglitz, 2000; Hanushek, 2013; Hanna, Mullainathan, and Schwartzstein, 2014).[1] The resulting errors are consequential, as people in developing countries face considerable risk in domains from health (Blakely et al., 2005) to employment and income (Jalan and Ravallion, 1999).

Air pollution provides a suitable domain to study decision problems involving forecasts, 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, Miller, and Sanders, 2016; Arceo, Hanna, and Oliva, 2016; Barreca, Neidell, and Sanders, 2021; Gong et al., 2022), labor productivity (Chang et al., 2016a; Neidell, 2017; Chang et al., 2019; Adhvaryu, Kala, and Nyshadham, 2022) and labor supply (Hanna and Oliva, 2015).[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), rendering 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 developing-world urbanites solve forecasting problems in the presence of limited information and human capital. We concern ourselves with the following broad questions. Do residents of developing cities exhibit positive demand for forecast products? Can they form useful forecasts, and can their forecasting ability be improved? How does consuming forecasts influence behavior, especially avoidance of environmental harm? The answers to these questions shed light on human decision making, and also form important inputs to benefit-cost analyses of policies

[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, Pattanayak, and Bluffstone (2015); He, Liu, and Salvo (2019); Bishop, Ketcham, and Kuminoff (2022). For reviews, see Graff Zivin and Neidell (2018) and Aguilar-Gomez et al. (2022).

concerning air pollution monitoring and abatement.

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

Our experiment involved 999 subjects in Lahore, Pakistan. In 2019, Lahore ranked as the twelfth most polluted city in the world, with air roughly comparable to that of Delhi and Dhaka (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 accurate forecasts (i.e., to form unbiased expectations). 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.

Average levels of human capital in Lahore may also hamper residents' ability to forecast 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 behavioral biases that generate forecasting errors even in highly educated populations

[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. Improved forecasting performance could arise from changes in both the first and higher moments of the forecast distribution.

[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, Shen, and Zhang, 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).

(Kahneman and Tversky, 1973).

Using incentive-compatible elicitations, we find that subjects exposed to our one-day-ahead air pollution forecasts were willing to pay an average of 93 Pakistani Rupees (PKR) to continue receiving forecasts for 90 days.[7] On a monthly basis, this equals roughly 60 percent of the cost of 4G mobile internet access. It stands in contrast to low willingness to pay for health-promoting goods like insecticide-treated nets and chlorine (Kremer et al., 2011). Both forecast provision and training reduced error in incentivized forecasts of fine particulates ($PM_{2.5}$) by roughly one-tenth of a standard deviation, or 5 $\mu g/m^3$.[8] This equals approximately 20 percent of the World Health Organization's corresponding maximum safe 24-hour standard.[9] Given that four to six months elapsed between the training and the forecast elicitation, the error reduction is notable and consistent with a durable increase in human capital.[10] Forecast provision increased willingness to pay for particulate-filtering masks by 6.6 PKR, roughly five percent of the retail price.[11] While the estimated effect of training on mask demand is positive (4 PKR), it is imprecise.[12]

We also estimate the effect of forecast receipt on outdoor time. This is an important margin of response to air pollution, as outdoor pollution exposure is frequently higher than indoor (US Centers for Disease Control and Prevention, 2022).[13] Forecast receipt increased outdoor time by 16 percent on relatively 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.[14] This pattern of responses was more pronounced for subjects who reported caring about air quality at baseline, and for children.

Our research design allows for investigation of the mechanisms driving reductions in air pollution forecast errors. Error reductions were greater at the one-day horizon than at the three-day horizon, reflecting improvements in both accuracy and precision. The SMS forecast treatment resulted in increased seeking of air pollution information from other

[7] Willingness to pay for both forecasts and masks was elicited using a Becker-DeGroot-Marschak mechanism (Becker, DeGroot, and Marschak, 1964).

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

[9] Here "fine particulates" denotes $PM_{2.5}$ air pollution: the concentration of particulates with diameter 2.5 microns or less, measured in micrograms per cubic meter ($\mu g/m^3$). For $PM_{2.5}$, the World Health Organization has set the daily standard at 25 $\mu g/m^3$ and the annual standard at 10 $\mu g/m^3$ (World Health Organization, 2006).

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

[11] 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.

[12] This estimate is not statistically significant at conventional thresholds.

[13] Relative indoor and outdoor air pollution exposure depends on the pollutant, combustion within the dwelling, and the intensity of physical exertion.

[14] For a formalization of this claim, please see the theoretical model of Section 2.

sources, consistent with complementarity and potentially an error-reducing mechanism in this group. Forecast error reductions were largest for trained subjects who choose to view a weather forecast before making their air pollution forecasts. This is potentially consistent with improved information processing by trained subjects.

Our project contributes to several literatures, of which the first is on forecasting in developing countries. Previous work has focused on farmers, including their yield expectations (Maertens, Michelson, and Nourani, 2021) and responses to precipitation forecasts (Rosenzweig and Udry, 2014a,b; Kala, 2017). While we also study responses to forecasts, this paper makes novel contributions on several dimensions. We elicit beliefs directly using incentive-compatible mechanisms, rather than inferring them within a structural model. Our experiment examines a different but consequential type of uncertainty–air pollution–and different responses–e.g. time spent outside. Studying these beliefs and behaviors is increasingly important, as rural citizens in the developing world continue moving to cities (Henderson, 2002).

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; Valdivia, 2015; Brudevold-Newman et al., 2017), or job training (Card et al., 2011; Acevedo et al., 2017). Our paper instead considers training in general-purpose forecast 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, Milkman, and Payne, 2015). So far as we are aware, ours is the first study to adapt such techniques to the constraints of a developing city.

The third relevant literature is on pollution avoidance behavior (Graff Zivin and Neidell, 2013).[15] A substantial empirical literature addresses avoidance behavior 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). Previous work on avoidance largely relies on natural experiments for identification (Neidell, 2009, 2010). For example, while we can often observe an avoidance behavior, such as a canceled trip to a movie theater (He, Luo, and Zhang, 2022) or a mask purchase (Zhang and Mu, 2018; Wang and Zhang, 2021), agents' air pollution expectations go unobserved. Our experimental design allows us to observe both expectations and avoidance behaviors for the same subjects, including total outdoor time rather than a proxy. Our finding that average willingness to pay for masks is roughly 70 percent of the retail price

[15] Some work prefers the term "averting behavior"; we view the two as synonymous.

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

offers a potential explanation for low take-up in some developing cities with high air pollution.

Lastly our results add to the literature on demand for environmental information.[17] To the best of our knowledge, ours is the first paper in this literature to study demand for forecasts.[18] Barwick et al. (2019) estimate effects of real-time air pollution information in China on a variety of outcomes, including shopping and mortality. Combined with an income-adjusted value of a statistical life (VSL) from the United States, these estimates allow recovery of a lower bound on the value of air pollution information.[19] Our study differs in eliciting the value of air pollution information directly, using an incentive-compatible mechanism. This avoids the need to specify channels through which information affects utility, and recovers the entire demand curve. Another related study is Barnwal et al. (2017), which randomized prices for arsenic testing of drinking-water wells in Bihar, India. In contrast, our experiment elicited willingness to pay for information from subjects who had experience with our forecast product. The resulting estimate is policy-relevant, and it pertains to a near-universal exposure (airborne fine particulates).[20] Our results on acquisition and processing of complementary environmental information (e.g. weather forecasts) are also novel.

The rest of the paper proceeds as follows. Section 2 presents our theoretical model and Section 3 discusses the design of our experiment. Section 4 describes our approach to empirical analysis. Section 5 discusses estimated treatment effects and mechanisms, 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 $s \in \{h, l\}$, high and low. The agent consumes only at $t = 1$. Pollution effects can be mitigated by engaging in avoidance behavior, which can be purchased in both periods. Examples of avoidance in our setting include protective face masks and cancellation

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

[18] 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.

[19] This VSL forms the basis of estimated benefits through both reduced mortality and reduced morbidity (Barwick et al., 2019).

[20] In Barnwal et al. (2017) and many other studies eliciting demand for static environmental and/or health information, demand can only be elicited once, before agents have any experience with the information. Since air pollution varies at high frequency, we can first expose agents to the information and then elicit willingness to pay.

or rescheduling of planned outdoor activities.[21] Let x and y denote the amount of avoidance purchased in periods 0 and 1 respectively, so the agent's payoff is

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

where $E > d(0)$ is her initial endowment that is large enough to avoid any credit constraints. and d^s is the state-dependent damage function,[22] 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, a mask purchased yesterday is a perfect substitute for a mask purchased today.[23] 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 $d^h_1(A) \leq d^l_1(A)\ \forall A$.[24] The cost of avoidance is captured through the cost function c, which is assumed to be strictly convex and increasing in both x and y. The marginal cost of avoidance rises if the agent waits to purchase. 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

[21] 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.

[22] We assume 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 G.

[23] 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, 2014a, or 2) make assumptions on the third derivatives of the damage function.

[24] 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.

pollution is $P(h) = \pi$, which can also be interpreted as the agent's unbiased prior before she begins optimizing. 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 information set ι at $t = 0$, both exogenous. We define the probability of a correct forecast as the 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). 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

Given the assumption that our treatments increase an agent's forecast skill, we are able to establish intuitive results: willingness to engage in (or pay for) avoidance is positive and increasing for those who receive our treatments. 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 mitigation strategies (e.g. masks) is increasing in forecast skill. Derivations of our results are presented in Appendix A for the sake of concision, but we state the hypotheses derived from our model below.

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

Turning to avoidance behavior, note that Lahore experienced high air pollution throughout our study. If subjects forecast high air pollution, we expect the following.

Hypothesis 2. *Subjects receiving our treatments should undertake more avoidance behavior. 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.

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 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. While agents' behavior in combining information and human capital to produce forecasts raises interesting research questions, they are mostly beyond the scope of this paper. Among participants who received the SMS service, we expect training will increase WTP for the SMS service if training and information are complements ($\frac{\partial^2 \rho}{\partial \iota \partial \tau} \geq 0$) and decrease it if they are substitutes.

3 Experimental design

Details of sampling and randomization are discussed in Appendix D. Figure 1 shows the division of our sample into treatment and control groups. We find no evidence of imbalance across these groups at either baseline or endline (Tables A1 through A4).

At baseline all subjects received a pamphlet explaining fine particulate air pollution ($PM_{2.5}$). A color-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 Treatment Groups 1 and 3 (*T1* and *T3*), we delivered SMS air pollution forecast messages to respondents every evening over a period of eight months. In Treatment Groups 2 and 3 (*T2* and *T3*), we implemented the forecast training once for every subject. More detailed descriptions of these interventions follow.

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. Our ensemble forecast combined the following inputs.[26]

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

[26] For more detail, see Section D.4.

1. A model based on data from our own air pollution monitors. $PM_{2.5}$ levels for *t+1* were predicted using a seven-day moving-average (MA7) model with day of the 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 E. We provided our treatment-group (*T1*) respondents two pieces of information in each SMS message: 1) an average $PM_{2.5}$ air pollution forecast for *t+1*; and 2) the realized average $PM_{2.5}$ level for the previous day (*t-1*). 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.

Effects of this treatment may reflect both the two included pieces of information and the manner in which they were communicated. They are not "pure" information effects. This is trivially true of any information intervention, however; the message cannot be separated completely from the medium.

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 behavioral 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.[27]

The first set of training exercises covered the concept of calibration. In pilot sessions, most subjects made large errors and demonstrated overconfidence, consistent with evidence from high-income countries (Mellers et al., 2014). The calibration exercises were intended to show subjects that they had room for improvement and open their minds to subsequent lessons.

[27] Urdu is one of the primary local languages spoken in Lahore.

The next set of exercises taught subjects to combine "outside" and "inside" views when making a forecast (Kahneman and Lovallo, 1993; Lovallo, Clarke, and Camerer, 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.

In the following 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). Subjects then completed an exercise that encouraged them not to round their forecasts excessively.

The next 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. The final exercise reminded subjects that people tend to allow their emotions and preferences to influence their forecasts. For example, a person who plans to spend the day outside tomorrow may underrate the chance of rain.

All exercises involved the active participation of subjects and were followed by clear feedback. The training was designed to be general: none of the exercises involved air pollution, nor was any air pollution information provided. Sessions were relatively brief, with an average duration of 51 minutes.[28]

3.2 Primary outcomes

Endline surveys were conducted in person in subjects' homes ten months after baseline, and measured five primary outcomes.[29]

1. **Willingness to pay (WTP) for pollution forecasts.** We elicited respondents' willingness to pay for a 90-day subscription to our $\text{PM}_{2.5}$ forecast SMS service. We used a Becker-DeGroot-Marschak (BDM) mechanism (Becker, DeGroot, and Marschak, 1964), drawing the price in Pakistani Rupees (PKR) from a uniform distribution on the interval $[0, 200]$.[30] This outcome allows us to measure forecast consumption—that is, do our respondents value forecast information?

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

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

[30] Before bidding on masks or forecasts, subjects completed a practice BDM auction using real money and answered comprehension questions. Enumerators explained any errors in answering these questions.

2. **Air pollution forecast error index.** We asked respondents to forecast Lahore's average $PM_{2.5}$ levels at $t + 1$ and $t + 3$ and calculated an index of the two forecast errors.[31] We incentivized the forecasts by offering payments for responses within 5, 10, and 20 percent of realized $PM_{2.5}$ levels. This outcome allows us to examine forecast production—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 takeup and use of relevant information.

3. **Willingness to pay for particulate-filtering face masks.** We elicited respondents' willingness to pay for air pollution masks using a BDM mechanism, with the price in PKR drawn from a uniform distribution on the interval $[0, 200]$. This outcome allows us to measure behavioral response—that is, do our treatments increase respondents' valuation of an avoidance good?

4. **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 behavioral response—that is, do our treatments alter respondents' time allocations in ways that reduce air pollution exposure?

5. **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."*[32] This measures whether our treatments help subjects to better smooth subjective well-being across days.

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.

[31] 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). We then averaged to form the index.

[32] Happiness variance and the air pollution avoidance index are self-reported measures.

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 F.3.

4.1 Intent to treat

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

$$Y_i = \alpha + \beta_F \mathit{Forecasts}_i + \beta_T \mathit{Training}_i + \beta_{FT} \mathit{Forecasts}_i \mathit{Training}_i + \varepsilon_i \tag{1}$$

In this equation i indexes subject and Y is the outcome. $\mathit{Forecasts}_i$ denotes random assignment to SMS forecasts, and $\mathit{Training}_i$ random assignment to training. 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 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.[33] This is the test that will be included in our multiple-testing correction procedure.[34]

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

$$Y_i = \beta_F \mathit{Forecasts}_i + \beta_T \mathit{Training}_i + \beta_{FT} \mathit{Forecasts}_i \mathit{Training}_i + \gamma Y_{0i} + \boldsymbol{X_i'}\boldsymbol{\delta} + \varepsilon_i \tag{2}$$

Notation for outcomes and treatments is as in Equation 1.[35] 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.[36]

Again as pre-specified, hypothesis testing on estimates of $\{\beta_F, \beta_T, \beta_{FT}\}$ varies by outcome. For the air pollution forecast error index, theory predicts that more information

[33] Note that because randomization block dummies are not included in Equation 1, 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.

[34] 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.

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

[36] See Appendix F.2 for more discussion.

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 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: $\beta_F > 0$, $\beta_T > 0$.

4.2 Treatment on the treated

For the training arm ($Training_i = 1$) we observe participation in the training session ($P_{Ti} = 1$). For the forecast arm ($Forecasts_i = 1$) takeup 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 takeup measure using endline survey responses to the question: "How many times in the last week have you seen our pollution forecast message?"[37] Denote the response of subject i as R_i.[38] Then a subject's takeup 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 takeup by those in our control group.[39] In the endline 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 nine of our messages over the entire course of the study, and just nine reported receiving ten or more (Table A16).

The interaction of takeup measures is simply $P_{FTi} = P_{Fi}P_{Ti}$. Effects of treatment on the treated are estimated using two-stage least squares (2SLS), with $\{\mathit{Forecasts}_i, \mathit{Training}_i, \mathit{Forecasts}_i\,\mathit{Training}_i\}$ instrumenting for $\{P_{Fi}, P_{Ti}, P_{FTi}\}$. Estimating equations appear in Appendix F.1. One- and two-tailed hypothesis tests for primary

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

[38] Subjects who responded "not sure" are assigned $R_i = 0$.

[39] Such non-compliance was not possible with the training treatment as we had absolute control over who participated.

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

outcomes are analogous to those in our ITT regressions.

4.3 p value adjustments

To address the problem of multiple hypothesis testing, we follow the procedures in Benjamini, Krieger, and Yekutieli (2006) to control the false discovery rate for a pre-specified subset of alternative hypotheses related to our primary outcomes: willingness to pay for forecast information ($\alpha + \beta_F > 0$), air pollution forecast error index ($\beta_F < 0, \beta_T < 0$), 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 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.

5 Results

5.1 Primary outcomes, intent to treat

We begin by examining 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. In a right-tailed test against a zero null hypothesis $p = .000$ and we reject at the one percent level of significance (see Table A5). 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 60 percent of the cost of 4G mobile internet access.[41] Considering a different benchmark, 93 PKR is approximately 20 percent of a day's earnings for an unskilled laborer.[42] Under the assumption that our

[41] Alternatively, one can use total mobile phone costs as a reference point. 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.

[42] Given mean WTP of 93 PKR, one might ask why there is not more provision of air pollution forecasts by for-profit firms. Even setting aside the sundry market failures that may be present in a developing-country

forecasts provide no direct utility (as in the theoretical model of Section 2),[43] mean WTP can be interpreted as the expected welfare gain from additional avoidance facilitated by the information. 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.[44]

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). Thornton (2008) found that even at a zero price, only 34 percent of subjects pick up HIV test results. Small incentives doubled takeup. This suggests that demand for health information (or alternatively, 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: in 2019, the city ranked as the twelfth-most polluted in the world based on $PM_{2.5}$ (IQAir, 2020). 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 5.6 discusses the external validity of our forecast demand estimates in greater depth and presents demand estimates for other high-pollution cities.

Our other primary hypotheses pertain to regression estimates of intent-to-treat effects, which are presented in Table 1. Column headings indicate dependent variables and shaded cells denote pre-specified primary hypotheses. Column 1 presents estimates for an index of air pollution forecast errors, aggregating errors at one- and three-day horizons ($t + 1$ and $t + 3$). This is our primary outcome in the domain of forecast production. Provision of SMS

setting like Lahore, such a firm could not prevent customers from passing forecasts to others. Information spillovers in the context of our experiment are addressed in Section 5.5.2.

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

[44] This estimate joins the large set roughly consistent with Samuelson (1965).

forecasts reduced forecast error by .074 standard deviations, while training reduced forecast error by .11 standard deviations. As discussed 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. The SMS effect on forecast error is statistically significant at the ten percent level ($p = .056$), while the training effect is statistically significant at the one percent level.[45] 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.[46] 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.

The interaction effect on air pollution forecast error is positive (column 1 of Table 1), so the effect on the group that received both treatments was $-.074 + -.11 + .11 = -.074\sigma$. While the estimated interaction effect is only marginally statistically significant, it 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 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.

The reductions in forecast error from forecast provision and training are practically large. Estimating effects in concentration rather than standard deviations (Table A6), both treatments reduced forecast error by approximately 5 $\mu g/m^3$, or 8 percent of the control mean. The WHO 24-hour standard for $PM_{2.5}$ is 25 $\mu g/m^3$, so the marginal effects of forecasts and training are roughly 20 percent of the maximum healthy level.[47] The $.11\sigma$ 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]

[45] Correction of these p-values for multiple testing is discussed later in this section.

[46] 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. It is of course possible that subjects in the SMS group based their forecasts in part on recently received messages. Recall that our forecast messages contained predictions for t+1 but not t+3.

[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 5 $\mu g/m^3$ error reduction is 3.4 percent.

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

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, 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. Thus we find a striking result—a shorter, simpler training, conducted with less educated subjects, yielded a coarsely similar improvement in forecast performance for air pollution.

Column 2 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 3 reports effects on willingness to pay for N95 particulate-filtering masks.[50] The SMS forecast intervention increased WTP by 6.58 PKR and this estimate is statistically significant at the five percent level.[51] The estimated effect of training is positive at 3.95 PKR, but not statistically significant. 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. 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, but are not statistically significant for either treatment. Together the results for mask demand and avoidance are qualitatively consistent with studies of

[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, we had capped the maximum bid at 150 PKR. Despite this difference in censoring, we employ baseline WTP as a control corresponding to Y_{0i} in Equation 2.

[51] As explained in Appendix F.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 A7 reports an alternative set of multiple-testing-corrected p-values in which tests on willingness to pay for masks and the avoidance index are two-tailed.

behaviors 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.

As discussed in Section 4.3, we adjust p values corresponding to primary hypotheses for multiple testing using the procedure of Benjamini, Krieger, and Yekutieli (2006), which controls the false discovery rate. Note that this procedure can yield corrected p values that are larger or smaller than uncorrected values. Table 2 presents the corrected probabilities. In the test of mean willingness to pay for forecasts (column 1; see also Figure 2) against a zero null, the estimate remains significant at the one percent level. For treatment-driven reductions in forecast error (column 2), $p = .09$ for forecasts and $p = .03$ for training and we reject a zero null hypothesis at the ten and five percent levels, respectively. Similarly, for the effect of SMS forecasts on willingness to pay for masks, $p = .07$ and we reject a zero null hypothesis at the ten percent level. We fail to reject the null for the effect of training on willingness to pay for masks at the ten percent level, but we note that the adjusted p value is not far above the threshold ($p = .13$).

5.2 Primary outcomes, effect of treatment on the treated

Table 3 reports estimated effects of treatment on the treated, instrumenting for takeup with treatment assignment as described in Section 4.2.[52] 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.[53] 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 air pollution forecast error by more, and increased willingness to pay for masks by more, than did forecasting training.[54] To put the point another way, the apparent advantage of training in ITT estimates arises largely from higher takeup rates, rather than larger local average treatment effects on compliers. Perfect compliers in the text message group increased their

[52] In the text message forecast condition, we define endogenous takeup as the average share of forecasts viewed, ranging from zero to one, including for the control group. In the training condition, takeup is a dummy for participation in training.

[53] 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.

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

willingness to pay for masks by approximately 14 percent of the control-group mean.[55]

5.3 Mechanisms, primary outcomes

5.3.1 Sources of reduced air pollution forecast error

Column one of Table 1 demonstrates that both SMS forecasts and training reduced an index of air pollution forecast errors at t+1 and t+3. Which element of the index drives the estimate? Table 4 separately reports treatment effects on standardized forecast errors at the two time horizons. Point estimates indicate that both treatments reduce error much more at t+1 than t+3, 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 95 percent more.[56] 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.[57] 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.

Figure 3 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 3 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 39.6 $\mu g/m^3$ 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.[58] In contrast, the distributions for the treated groups are shifted rightward, indicating reduced under-prediction. Dispersion is also reduced. Tables A8 and A9 quantify these differences in means and standard deviations, respectively. Treatment

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

[56] These percentage changes use midpoints as bases.

[57] The SMS effect at t+1 is not mechanical. At the time they made their incentivized endline predictions, subjects in the SMS group had not yet received our forecast for the next day.

[58] 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 A1 illustrates the distribution of baseline $t+1$ forecast errors.

increased means (reduced underprediction) by 2.1 to 6.4 $\mu g/m^3$, with the largest change in the SMS-forecast group.[59] 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.[60] This is apparent in Figure 3, where the height of the distribution function at the mode is much greater for the training group than for the others. Because our indexed measure of forecast error (as in Table 1 and Table 4) is built from absolute values, effects on this variable reflect both the reduced underprediction and the reduced dispersion in the underlying (non-absolute, non-standardized) forecast errors.

5.3.2 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 $t+3$, and an index of errors at both horizons (Table 5).[61]

The effect of SMS forecast receipt on forecast error at $t+1$ (row one, column one) is negative and statistically significant at the ten percent level. 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 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 and the error index (columns two and three) 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 all three columns, and statistically significant at conventional thresholds in columns one and three. This is consistent with the training functioning as intended. In the forecasts-and-

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

[60] 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.

[61] The estimating equation is $Y_{it} = \beta_1 Forecasts_i + \beta_2 Post_t + \beta_3 Forecasts_i * Post_t + \boldsymbol{X}_{\boldsymbol{i}}'\boldsymbol{\delta} + \varepsilon_{it}$, 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.

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 5 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 A10 presents the same analysis for the endline sample and results are strongly similar.

5.3.3 Information seeking and processing

Our endline survey asked a number of questions about subjects' information diets, especially pertaining to weather and air quality. Columns one and two of Table 6 present ITT effects on counts of sources consulted in the past week for a given category. Subjects receiving SMS forecasts increased the number of air pollution information sources they consulted by .23, or 15 percent of the control group mean. The estimate is statistically significant at the five percent level. Our SMS forecasts were deliberately excluded from the question, so the effect is not mechanical. The positive estimate is consistent with complementarity of our SMS pollution forecasts and other air pollution information, e.g. social media posts.

Columns three and four of Table 6 evaluate the role of weather forecasts in production of subjects' own incentivized air pollution forecasts. Recall that before making an incentivized forecast (at both baseline and endline), subjects were offered the opportunity to view a weather forecast. In column three estimates are small and not statistically distinguishable from zero; neither SMS forecasts nor training changed subjects' takeup of weather forecasts. It is possible these null results arise from a ceiling effect, as 92 percent of control subjects took up the weather forecast. Column four interacts treatments with weather forecast takeup to estimate heterogeneous effects on air pollution forecast error. Among trained subjects who did not take up the weather forecast, air pollution forecast error actually increased by .24 standard deviations. Weather forecast takeup is endogenous and the result must be interpreted cautiously. It is worth noting, however, that the training emphasized the importance of carefully combining an outside view (base rate) and an inside view (situation-specific information like a weather forecast). A trained subject who did not take up the weather forecast plausibly missed this important lesson, perhaps because it was not understood or the subject did not take the training seriously. It is unsurprising that such a subject might perform worse in forecasting air pollution. Among trained subjects who did take up the weather forecast, the marginal effect of training was $.24 - .37 = -.13$. These are the subjects who drive the reductions in air pollution forecast error estimated in our primary results (Table 1). The interaction with weather forecast takeup is larger for training (-.37 standard deviations) than for SMS forecasts (-.10 standard deviations). This pattern is consistent with trained subjects making better use of relevant information, but we cannot reject a null hypothesis that the coefficients on these two interactions are equal at any conventional threshold.

5.3.4 Mask demand

To investigate the positive treatment effect of SMS forecasts on WTP for masks, Figure A2 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 A11). 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.

5.4 Secondary outcomes, intent to treat

In this section we evaluate the remaining hypotheses from our theoretical model (Section 2). The first two are related to avoidance behavior. Column one of Table 7 presents SMS forecast effects on time spent outside on the day before the endline survey, pooling over adults and children.[62] We focus on the SMS treatment rather than training because our SMS forecasts

[62] 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 endline time diary (the day before the endline survey). As elsewhere in the paper, baseline controls in $\boldsymbol{X}$ were chosen using post-double-selection LASSO.

varied over the course of the endline survey and training did not, but results are robust to estimating the full suite of treatment effects (Table A12). 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. This estimate is statistically significant at the five percent level ($p = .011$). 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 endline 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 7 presents heterogeneous treatment effects by whether subjects care about air quality.[63] At baseline 85 percent of subjects reported caring. This dimension of heterogeneity was not pre-specified and results should be interpreted cautiously. Having said that, the indicator for caring co-varies with other attributes in ways that suggest it is not merely cheap talk. Covariances with baseline avoidance, endline reports of viewing SMS forecasts, and endline demand for SMS forecasts are all positive and statistically significant; the covariance with baseline demand for masks is positive but imprecisely estimated (Table A13). For subjects who reported caring, the pattern of signs in column two of Table 7 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.[64] Broadly the results in column two are consistent with Hypothesis 3.

Columns three through six of Table 7 repeat the specifications of the first two columns separately for adults and children. Broadly the patterns of signs for adults and children

[63] 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.

[64] Among subjects who care about air pollution, the marginal effect of SMS forecast treatment is $.068 + .69 = .76$ hours on cleaner days and $.068 + .69 + 1.26 - 2.27 = -.25$ hours on dirtier days. The latter marginal effect is not statistically significant at any conventional threshold.

are similar to the pooled estimates.[65] 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.[66]

Our experiment produced no direct evidence on whether the changes in outdoor time seen in Table 7 reduced pollution exposure. 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.

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 A5, column two) is positive 5.3 PKR and large in proportional terms, consistent with Hypothesis 4. This estimate is not statistically significant at conventional thresholds (two-tailed $p = .14$). The estimated interaction effect of the forecast and training interventions is negative. Under Hypothesis 5, 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.

[65] 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 A14 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.

[66] The marginal effects are $.86 + 1.07 - .10 - 2.40 = -.57$ for children and $-.41 + 1.54 + 1.17 - 2.17 = .13$ for adults.

5.5 Robustness

5.5.1 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, Jain, and Jayachandran (2018). From this module, we construct an index of social desirability and report treatment effects on this variable in Table A15. 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. No measure of social desirability is complete and we cannot rule out this type of bias with certainty, but there is no evidence of it in Table A15.

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 A3 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. 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. Second, note that willingness to pay in the control group was still considerable at 89PKR (Table A5, column one; cf. 93PKR in the forecast-only treatment group). Subjects in the control group received no treatment involving forecasts, nor did they have strong reason to believe forecast information was a focus of the study. Control-group surveys asked about several kinds of air pollution information and included two other BDM elicitations (for a low-cost practice item and a N95 mask). Given the experimental environment experienced by control subjects, it is unlikely that the willingness to pay for SMS forecasts that they exhibited in an incentive-compatible mechanism stemmed largely from experimenter demand effects.

5.5.2 Spillovers

Given the ease of relaying our forecasts, spillovers might in principle be a concern for our text message forecast treatment. The sampling was designed to mitigate these concerns by separating subjects in space, 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 A16 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 5.2. Because spillovers were so infrequent, accounting for them produces minimal changes in our estimates.

5.5.3 Other robustness

Table A17 reports primary ITT results without controlling for baseline measures of the outcomes, and without the controls selected by the post-double-selection LASSO. As expected precision is somewhat worse than in Table 1. Point estimates for pre-specified primary effects are strongly similar across the two tables.

5.6 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 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. Let us consider how Lahore, which had the world's highest levels of fine particulate pollution in 2022, compares to the other 24 of the 25 most polluted cities on dimensions 2) and 3).[67]

As discussed in Section 1, both the US State Department and the Punjab Environmental Protection Department (EPD) operate reference-quality monitors in Lahore. Only retrospective measurements are given, online and in English, which many residents do not speak. Monitoring is incomplete in both space and time, and residents may not trust these information sources. Some private individuals post readings from low-cost air pollution sensors online, but again coverage is spotty and data quality varies widely (Williams et al., 2014). This information environment is common in the world's most polluted cities. The US State Department also operates monitors in New Delhi, Peshawar, and N'Djamena. Of the 18 top-25 cities in India, 11 are covered by government monitors.[68] In summary, three of 24 comparison cities resemble Lahore in that a US State Department monitor is present, and at least 11 of 24 resemble Lahore in that monitors operated by the Government of India are present.[69]

Comparison of the scope for avoidance in Lahore and other highly polluted cities is more difficult. Avoidance may depend on the availability of N95 masks and air purifiers. It may also depend on high-frequency pollution variation in time, and fine-scale pollution variation in space. The industry-occupation composition of the labor market and its institutions plausibly matter as well. A complete accounting is beyond the scope of this paper, but our data allow us to learn something nonetheless. To begin, note that 20 of 24 comparison cities are in South Asia, with 2 in Pakistan and 18 in India. Owing partly to a common colonial history, these cities share some cultural and economic characteristics with Lahore, which may limit the importance of unobserved confounders.

To account for population differences, we regress willingness to pay for air pollution forecasts on a set of demographic variables available in both our data and the 2011 Indian Census for Indian cities, and the latest round of the World Bank Multiple Indicator Cluster Survey (MICS) for others.[70] These variables include home ownership, number of household

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

[68] Asopur, Dharuhera, Chapra, Bahadurgarh, Jind, Charkhi Dadri, and Rohtak appear not to contain monitors (Central Pollution Control Board, 2022).

[69] To the best of our knowledge, the government of Pakistan does not currently operate monitors in Peshawar and Faisalabad. We have been unable to determine whether there are government monitors in Hotan, N'Djamena, Baghdad, and Karagandy.

[70] The 2011 Indian Census was downloaded in April 2023 from

members, number of rooms in the house, and whether households own a car, a computer, a motorcycle, a television, and/or an air conditioner. Importantly these variables capture wealth and income differences, and they are components of standard poverty measures collected across many surveys.[71] By combining in-sample coefficient estimates (column 1 in Tables A18 and A19) with means from these other surveys, we can estimate mean willingness to pay for 23 of the 24 comparison cities.[72] The resulting means range from 70PKR to 105PKR (Table A20). This variation is relatively limited not because our regression lacks predictive power ($R^2 = .13$), but because numerically important differences, relative to Lahore, often offset each other in calculating estimated demand. Baseline air pollution forecast error does not predict willingness to pay after conditioning on these covariates, which implies they are reasonable proxies for informedness about air pollution (column 2 in Tables A18 and A19).[73] This suggests that differences in consumption of available information will not introduce large bias in our estimates of willingness to pay in other cities.

Our out-of-sample willingness to pay estimates should nonetheless be interpreted cautiously. The empirical model used to create them is surely incomplete, and large benefit transfer errors are possible (Kaul et al., 2013). Even large benefit transfer errors, however, may be irrelevant from a policy perspective. That is because our estimates of willingness to pay dwarf the costs of monitoring. A reference-quality monitor costs approximately US$22,000 to US$24,000 (Hussey, Lemelin, and Lulofs, 2022). The total costs of monitoring are greater because, for example, such monitors require small shelters.[74] Even allowing for these additional costs, monitoring benefits exceed costs by roughly two orders of magnitude. Taking the average WTP for Lahore from Table A20, the implied annual aggregate WTP is roughly 3.6 billion PKR, or US$12.7 million.[75]

In summary, 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

https://censusindia.gov.in/census.website/data/census-tables. MICS data was downloaded in April 2023 from https://mics.unicef.org. In the case of MICS, the latest available survey was always used, with specific years varying by city.

[71] See, e.g., List (2020), for a discussion of the role of comparisons across individual-specific factors such as these variables when considering generalizability.

[72] We are unable to find comparable demographic data for Hotan, China.

[73] Figure A3 demonstrates that baseline air pollution forecast error predicts willingness to pay for forecast when one does not condition on these demographic predictors.

[74] Additional costs include those associated with sample collection and analysis.

[75] We extrapolate from mean WTP for 90 days of forecasts among the 11mn residents of Lahore: $(80.3PKR) * (365/90) * 11119985 = 3.621e + 09PKR$, or US$12.7mn as of this writing. If monitoring costs are assumed to be on the close order of 100,000, then WTP is roughly 100 times larger.

monitoring in the urban settings where those efforts matter most. 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 make more accurate forecasts. 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 suggest that assisting people in using information they already have is at least as important as delivering new information (Hanna, Mullainathan, and Schwartzstein, 2014).

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 pollution avoidance. 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.

Masks are a private response to environmental information. Somanathan (2010) has surmised that in developing countries, environmental information may also increase demand for environmental quality and lead to public action. If so, the long-run responses to air pollution forecasts may be greater in scope and magnitude than those we study.

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, Lemelin, and Lulofs, 2022)—the level of demand we estimate indicates that the welfare gain from investments in air pollution monitoring and forecasting may be considerable. 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 the health effects of air pollution, 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 forecasts, they are plausibly considerable, and warrant future research.

References

Acevedo, P., G. Cruces, P. Gertler, and S. Martinez. 2017. “Living up to expectations: How job training made women better off and men worse off.” Working paper, National Bureau of Economic Research.

Adhvaryu, A., N. Kala, and A. Nyshadham. 2022. “Management and shocks to worker productivity.” *Journal of Political Economy* 130:1–47.

Aguilar-Gomez, S., H. Dwyer, J.S. Graff Zivin, and M.J. Neidell. 2022. “This is Air: The “Non-Health” Effects of Air Pollution.” Working paper, National Bureau of Economic Research.

Ahrens, A., C. Hansen, and M. Schaffer. 2018. “pdslasso and ivlasso: Progams for post-selection and post-regularization OLS or IV estimation and inference.” Statistical Software Components S458459, Boston College Department of Economics, Revised 24 Jan 2019.

Alberini, A., M. Cropper, T.T. Fu, A. Krupnick, J.T. Liu, D. Shaw, and W. Harrington. 1997. “Valuing health effects of air pollution in developing countries: the case of Taiwan.” *Journal of Environmental Economics and Management* 34(2):107–126.

Arceo, E., R. Hanna, and P. Oliva. 2016. “Does the effect of pollution on infant mortality differ between developing and developed countries? Evidence from Mexico City.” *The Economic Journal* 126:257–280.

Athey, S., and G.W. Imbens. 2017. “The Econometrics of Randomized Experiments.” In *Handbook of Economic Field Experiments*. Elsevier, vol. 1, pp. 73–140.

Barnwal, P., A. van Geen, J. von der Goltz, and C.K. Singh. 2017. “Demand for environmental quality information and household response: Evidence from well-water arsenic testing.” *Journal of Environmental Economics and Management* 86:160–192.

Barreca, A.I., M. Neidell, and N.J. Sanders. 2021. “Long-run pollution exposure and mortality: Evidence from the Acid Rain Program.” *Journal of Public Economics* 200.

Barwick, P.J., S. Li, L. Lin, and E. Zou. 2019. “From fog to smog: The value of pollution information.” Working paper, National Bureau of Economic Research.

Becker, G.M., M.H. DeGroot, and J. Marschak. 1964. “Measuring utility by a single-response sequential method.” *Behavioral Science* 9:226–232.

Benjamini, Y., A.M. Krieger, and D. Yekutieli. 2006. “Adaptive linear step-up procedures that control the false discovery rate.” *Biometrika* 93:491–507.

Bishop, K.C., J.D. Ketcham, and N.V. Kuminoff. 2022. “Hazed and confused: the effect of air pollution on dementia.” *Review of Economic Studies*, pp. .

Blakely, T., S. Hales, C. Kieft, N. Wilson, and A. Woodward. 2005. “The global distribution of risk factors by poverty level.” *Bulletin of the World Health Organization* 83:118–126.

Bond, T.N., and K. Lang. 2019. “The sad truth about happiness scales.” *Journal of Political Economy* 127:1629–1640.

Brown, M.B., and A.B. Forsythe. 1974. “Robust tests for the equality of variances.” *Journal of the American Statistical Association* 69:364–367.

Brudevold-Newman, A.P., M. Honorati, P. Jakiela, and O.W. Ozier. 2017. “A firm of one’s own: experimental evidence on credit constraints and occupational choice.” Working paper No. 7977, World Bank Policy Research.

Card, D., P. Ibarrarán, F. Regalia, D. Rosas-Shady, and Y. Soares. 2011. “The labor market impacts of youth training in the Dominican Republic.” *Journal of Labor Economics* 29:267–300.

Central Pollution Control Board. 2022. “Manual Ambient Air Quality Monitoring Stations in the Country (as on 15.09.2022).” Working paper, Ministry of Environment, Forest, and Climate Change.

Chang, T., J. Graff Zivin, T. Gross, and M. Neidell. 2019. “The Effect of Pollution on Worker Productivity: Evidence from Call Center Workers in China.” *American Economic Journal: Applied Economics* 11:151–72.

—. 2016a. “Particulate pollution and the productivity of pear packers.” *American Economic Journal: Economic Policy* 8:141–69.

Chang, W., E. Chen, B. Mellers, and P. Tetlock. 2016b. “Developing expert political judgment: The impact of training and practice on judgmental accuracy in geopolitical forecasting tournaments.” *Judgment and Decision Making* 11:509.

Cropper, M.L., N.B. Simon, A. Alberini, S. Arora, and P. Sharma. 1997. “The health benefits of air pollution control in Delhi.” *American Journal of Agricultural Economics* 79:1625–1629.

Crowne, D.P., and D. Marlowe. 1960. “A new scale of social desirability independent of psychopathology.” *Journal of Consulting Psychology* 24:349.

Dhar, D., T. Jain, and S. Jayachandran. 2018. “Reshaping Adolescents’ Gender Attitudes: Evidence from a School-Based Experiment in India.” Working paper, National Bureau of Economic Research.

Dupas, P., and E. Miguel. 2017. “Impacts and determinants of health levels in low-income countries.” In *Handbook of Economic Field Experiments*. Elsevier, vol. 2, pp. 3–93.

Gerber, A.S., and D.P. Green. 2012. *Field experiments: Design, analysis, and interpretation*. WW Norton.

Ghanem, D., S. Shen, and J. Zhang. 2020. “A censored maximum likelihood approach to quantifying manipulation in China’s air pollution data.” *Journal of the Association of Environmental and Resource Economists* 7:965–1003.

Ghanem, D., and J. Zhang. 2014. "Effortless Perfection: Do Chinese cities manipulate air pollution data?" *Journal of Environmental Economics and Management* 68(2):203–225.

Gong, Y., S. Li, N.J. Sanders, and G. Shi. 2022. "The mortality impact of fine particulate matter in China: Evidence from trade shocks." *Journal of Environmental Economics and Management*, pp. .

Graff Zivin, J., and M. Neidell. 2018. "Air pollution's hidden impacts." *Science* 359:39–40.

—. 2009. "Days of haze: Environmental information disclosure and intertemporal avoidance behavior." *Journal of Environmental Economics and Management* 58(2):119–128.

—. 2013. "Environment, health, and human capital." *Journal of Economic Literature* 51:689–730.

Hanna, R., S. Mullainathan, and J. Schwartzstein. 2014. "Learning through noticing: Theory and evidence from a field experiment." *The Quarterly Journal of Economics* 129:1311–1353.

Hanna, R., and P. Oliva. 2015. "The effect of pollution on labor supply: Evidence from a natural experiment in Mexico City." *Journal of Public Economics* 122:68–79.

Hanushek, E.A. 2013. "Economic growth in developing countries: The role of human capital." *Economics of Education Review* 37:204–212.

He, J., H. Liu, and A. Salvo. 2019. "Severe air pollution and labor productivity: Evidence from industrial towns in China." *American Economic Journal: Applied Economics* 11:173–201.

He, X., Z. Luo, and J. Zhang. 2022. "The impact of air pollution on movie theater admissions." *Journal of Environmental Economics and Management* 112:102626.

Henderson, V. 2002. "Urbanization in developing countries." *The World Bank Research Observer* 17:89–112.

Hussey, T., A. Lemelin, and M. Lulofs. 2022. "Maine DEP Low-Cost PM Sensor Comparison.", pp. .

IQAir. 2020. "2019 World Air Quality Report." Working paper.

—. 2023. "2022 World Air Quality Report." Working paper.

Jalan, J., and M. Ravallion. 1999. "Are the poor less well insured? Evidence on vulnerability to income risk in rural China." *Journal of Development Economics* 58:61–81.

Jalan, J., and E. Somanathan. 2008. "The importance of being informed: Experimental evidence on demand for environmental quality." *Journal of Development Economics* 87:14–28.

Jeuland, M., S.K. Pattanayak, and R. Bluffstone. 2015. “The economics of household air pollution.” *Annu. Rev. Resour. Econ.* 7:81–108.

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

Kahneman, D., and D. Lovallo. 1993. “Timid choices and bold forecasts: A cognitive perspective on risk taking.” *Management Science* 39:17–31.

Kahneman, D., and A. Tversky. 1973. “On the psychology of prediction.” *Psychological Review* 80:237.

Kala, N. 2017. “Learning, Adaptation, and Climate Uncertainty: Evidence from Indian Agriculture.” *MIT Center for Energy and Environmental Policy Research Working Paper* 23.

Karlan, D., and M. Valdivia. 2011. “Teaching entrepreneurship: Impact of business training on microfinance clients and institutions.” *Review of Economics and Statistics* 93:510–527.

Kaul, S., K.J. Boyle, N.V. Kuminoff, C.F. Parmeter, and J.C. Pope. 2013. “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.

Knittel, C.R., D.L. Miller, and N.J. Sanders. 2016. “Caution, drivers! Children present: Traffic, pollution, and infant health.” *Review of Economics and Statistics* 98:350–366.

Kremer, M., and R. Glennerster. 2011. “Improving health in developing countries: evidence from randomized evaluations.” In *Handbook of Health Economics*. Elsevier, vol. 2, pp. 201–315.

Kremer, M., J. Leino, E. Miguel, and A.P. Zwane. 2011. “Spring cleaning: Rural water impacts, valuation, and property rights institutions.” *The Quarterly Journal of Economics* 126:145–205.

Laumbach, R.J. 2010. “Outdoor air pollutants and patient health.” *American Family Physician* 81:175.

List, J.A. 2020. “Non est disputandum de generalizability? a glimpse into the external validity trial.” Working paper, National Bureau of Economic Research.

Lovallo, D., C. Clarke, and C. Camerer. 2012. “Robust analogizing and the outside view: two empirical tests of case-based decision making.” *Strategic Management Journal* 33:496–512.

Ludwig, J., S. Mullainathan, and J. Spiess. 2019. “Augmenting Pre-Analysis Plans with Machine Learning.” In *AEA Papers and Proceedings*. vol. 109, pp. 71–76.

Madajewicz, M., A. Pfaff, A. Van Geen, J. Graziano, I. Hussein, H. Momotaj, R. Sylvi, and H. Ahsan. 2007. “Can information alone change behavior? Response to arsenic contamination of groundwater in Bangladesh.” *Journal of Development Economics* 84:731–754.

Maertens, A., H. Michelson, and V. Nourani. 2021. “How do farmers learn from extension services? Evidence from Malawi.” *American Journal of Agricultural Economics* 103:569–595.

Mauboussin, M.J., and D. Callahan. 2015. “Sharpening Your Forecasting Skills - Foresight Is a Measurable Skill That You Can Cultivate.” Working paper, Credit Suisse.

McKenzie, D., and C. Woodruff. 2014. “What are we learning from business training and entrepreneurship evaluations around the developing world?” *The World Bank Research Observer* 29:48–82.

Mellers, B., L. Ungar, J. Baron, J. Ramos, B. Gurcay, K. Fincher, S.E. Scott, D. Moore, P. Atanasov, S.A. Swift, et al. 2014. “Psychological strategies for winning a geopolitical forecasting tournament.” *Psychological Science* 25:1106–1115.

Moore, R.T. 2012. “Multivariate continuous blocking to improve political science experiments.” *Political Analysis* 20:460–479.

Moretti, E., and M. Neidell. 2011. “Pollution, health, and avoidance behavior evidence from the ports of Los Angeles.” *Journal of Human Resources* 46:154–175.

Morewedge, C.K., H. Yoon, I. Scopelliti, C.W. Symborski, J.H. Korris, and K.S. Kassam. 2015. “Debiasing decisions: Improved decision making with a single training intervention.” *Policy Insights from the Behavioral and Brain Sciences* 2:129–140.

Neidell, M. 2017. “Air pollution and worker productivity.” *IZA World of Labor*, pp. .

—. 2004. “Air pollution, health, and socio-economic status: the effect of outdoor air quality on childhood asthma.” *Journal of Health Economics* 23:1209–1236.

—. 2010. “Air quality warnings and outdoor activities: evidence from Southern California using a regression discontinuity design.” *Journal of Epidemiology & Community Health* 64:921–926.

—. 2009. “Information, avoidance behavior, and health the effect of ozone on asthma hospitalizations.” *Journal of Human Resources* 44:450–478.

NIPS and ICF. 2019. “Pakistan Demographic and Health Survey 2017-18.” Working paper, National Institute of Population Studies (NIPS) Pakistan and ICF, Islamabad, Pakistan, and Rockville, Maryland, USA.

Pakistan Bureau of Statistics. 2017. “Household Integrated Economic Survey.” Working paper.

Punjab Environmental Protection Department. 2017. “Policy on Controlling Smog 2017.”

Riaz, R., and K. Hamid. 2018. “Existing Smog in Lahore, Pakistan: An Alarming Public Health Concern.” *Cureus* 10.

Rosenzweig, M., and C.R. Udry. 2014a. “Forecasting profitability.” Working paper.

Rosenzweig, M.R., and C. Udry. 2014b. "Rainfall forecasts, weather, and wages over the agricultural production cycle." *American Economic Review* 104:278–83.

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

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

Somanathan, E. 2010. "Effects of information on environmental quality in developing countries." *Review of Environmental Economics and Policy* 4:275–292.

Stiglitz, J.E. 2000. "The contributions of the economics of information to twentieth century economics." *The Quarterly Journal of Economics* 115:1441–1478.

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

Thornton, R.L. 2008. "The demand for, and impact of, learning HIV status." *American Economic Review* 98:1829–63.

US Centers for Disease Control and Prevention. 2022. "Ozone and Your Health."

Valdivia, M. 2015. "Business training plus for female entrepreneurship? Short and medium-term experimental evidence from Peru." *Journal of Development Economics* 113:33–51.

Wager, S., W. Du, J. Taylor, and R.J. Tibshirani. 2016. "High-dimensional regression adjustments in randomized experiments." *Proceedings of the National Academy of Sciences* 113:12673–12678.

Wang, Z., and J. Zhang. 2021. "The Value of Information Disclosure: Evidence from Mask Consumption in China." Working paper.

Williams, R., R. Long, M. Beaver, A. Kaufman, F. Zeiger, M. Heimbinder, B.R. Acharya, B.A. Grinwald, K.A. Kupcho, S.E. Robinson, O. Zaouak, B. Aubert, M. Hannigan, R. Piedrahita, N. Masson, B. Moran, M. Rook, P. Heppner, C. Cogar, N. Nikzad, and W.G. Griswold. 2014. "Sensor Evaluation Report." Working paper No. EPA 600/R-14/143, US Environmental Protection Agency.

World Bank. 2017. "Education Statistics." Working paper.

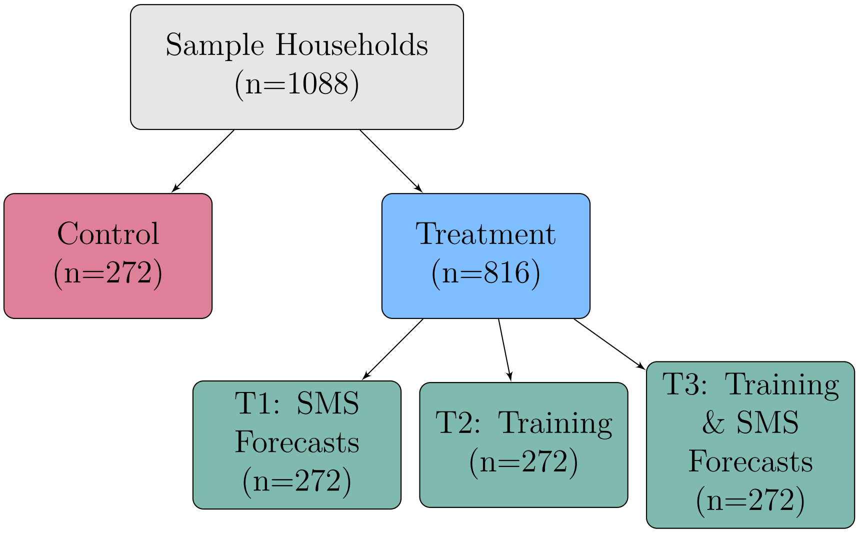
World Health Organization. 2006. "WHO Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide: global update 2005: summary of risk assessment." Working paper, Geneva: World Health Organization.

Zahra-Malik, M. 2017. "In Lahore, Pakistan, Smog Has Become a 'Fifth Season'." *The New York Times*, Nov, pp. A11.

Zhang, J., and Q. Mu. 2018. "Air pollution and defensive expenditures: Evidence from particulate-filtering facemasks." *Journal of Environmental Economics and Management* 92:517–536.

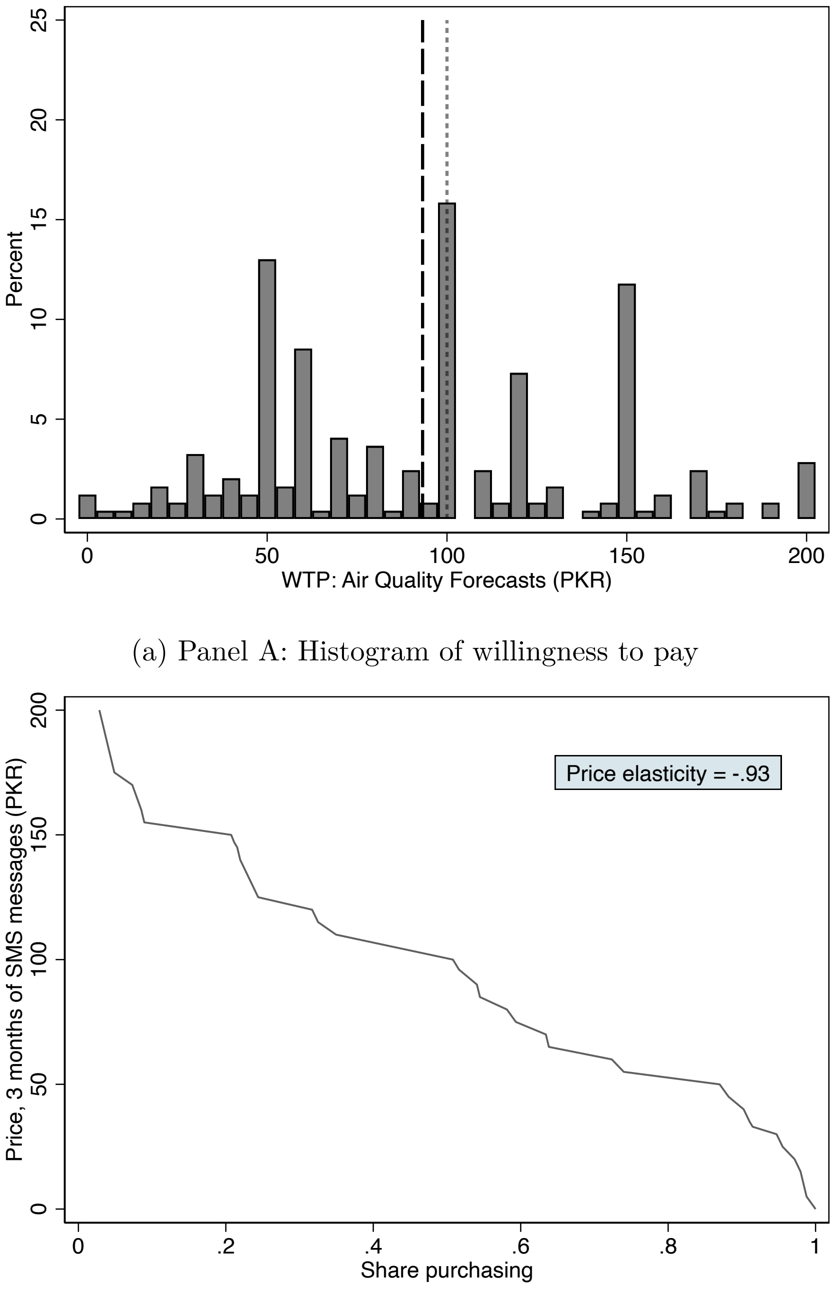
7 Figures

Figure 1: Experimental Groups



Note: Subjects were allocated over experimental groups as illustrated above. "SMS Forecasts" are the one-day-ahead air pollution forecasts delivered by text message (SMS) and described in Section 3.1. "Training" is the session focused on general-purpose forecasting skills, also described in Section 3.1. The average duration of training was roughly one hour.

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

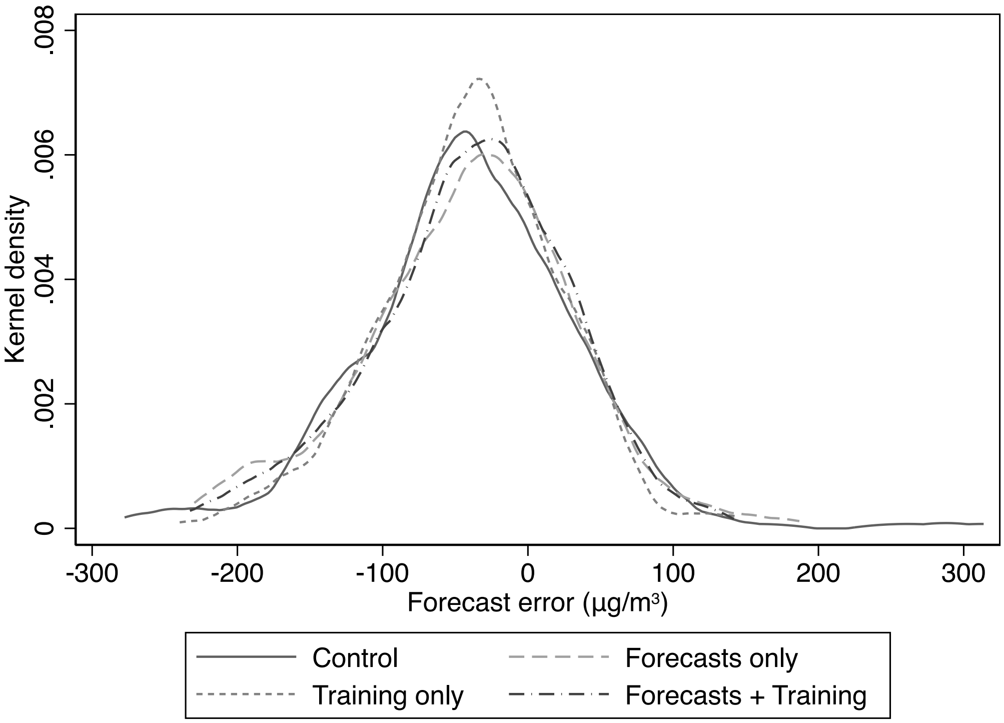


(a) Panel A: Histogram of willingness to pay

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

Note: Endline 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 A5. 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, DeGroot, and Marschak, 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.

Figure 3: Mechanisms: 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 realized 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.

8 Tables

Table 1: Primary outcomes, intent to treat

	Forecast error index	Happiness variance	WTP: Masks	Avoidance index
Forecasts	-0.074	0.052	6.58	0.046
	(0.047)	(0.070)	(3.53)	(0.059)
	[0.056]	[0.77]	[0.03]	[0.22]
Training	-0.11	0.078	3.95	0.019
	(0.047)	(0.071)	(3.54)	(0.059)
	[0.01]	[0.86]	[0.13]	[0.37]
Forecasts + Training	0.11	-0.11	-7.58	-0.022
	(0.066)	(0.099)	(5.02)	(0.083)
	[0.097]	[0.13]	[0.13]	[0.79]
Observations	999	951	999	999
Control mean	-0.000	0.017	104.1	-0.0019

Note: Both treatments reduced air pollution forecast error, 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 the forecast error index, the variance of happiness, and the avoidance index. Units are Pakistani Rupees (PKR) for willingness to pay for masks. Shaded cells denote pre-specified estimates of interest. 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 index ($\beta_F < 0, \beta_T < 0$), 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.

Table 2: MHT-adjusted p-values, primary outcomes (ITT)

	WTP: Forecasts	Forecast error index	Happiness variance	WTP: Masks	Avoidance index
Forecasts	-	0.09	-	0.07	0.17
Training	-	0.03	0.42	0.13	0.27
Forecasts + Training	-	-	-	-	-
Mean, forecast-only group	0.001	-	-	-	-

Note: Estimates remain significant at conventional thresholds after correction for multiple hypothesis testing. The p value in column 1 corresponds to the test of mean willingness to pay for forecasts in the forecast-only treatment group. This test is illustrated in Figure 2 and formalized at the bottom of column 1, Table A5. The p values in columns 2 through 5 correspond to the pre-specified estimates in Table 1 (shaded cells). The MHT correction is performed using the procedure of Benjamini, Krieger, and Yekutieli (2006), which controls the false discovery rate. As discussed in Section 4.3, the resulting corrected p-values can be larger or smaller than their uncorrected analogs.

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

	Forecast error index	Happiness variance	WTP: Masks	Avoidance index
% Forecasts seen	-0.16	0.063	14.8	0.092
	(0.11)	(0.16)	(8.09)	(0.14)
	[0.07]	[0.65]	[0.03]	[0.25]
Attended training	-0.097	0.061	3.92	0.027
	(0.049)	(0.074)	(3.79)	(0.062)
	[0.02]	[0.8]	[0.15]	[0.33]
% Forecasts seen	0.24	-0.22	-18.2	-0.061
× Attended training	(0.16)	(0.24)	(12.2)	(0.20)
	[0.13]	[0.18]	[0.14]	[0.76]
Observations	999	951	999	999
Control mean	-0.00	0.017	104.1	-0.0019
1st stage F-stat	173.6	168.2	174	171.5

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 error index, the variance of happiness, and the avoidance index. Units are Pakistani Rupees (PKR) for willingness to pay for masks. Shaded cells denote pre-specified estimates of interest. 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 index ($\beta_F < 0, \beta_T < 0$), 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.

Table 4: Mechanisms: Forecast errors, by time horizon

	Forecast error (t + 1)	Forecast error (t + 3)
Forecasts	-0.11	-0.023
	(0.058)	(0.056)
Training	-0.15	-0.065
	(0.053)	(0.057)
Forecasts + Training	0.13	0.070
	(0.075)	(0.082)
Observations	999	999
Control mean	0.00	0.00

Note: Reductions in the air pollution forecast error index (Table 1, column 1) resulted primarily from improved one-day-ahead (t+1) forecasts. Estimates correspond to Equation 2, with the dependent variable indicated in the column heading. 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.

Table 5: Mechanisms: Forecast errors, beginning and end of training

	Forecast error (t + 1)	Forecast error (t + 3)	Forecast error idx
Forecasts	-0.13	0.027	-0.045
	(0.069)	(0.074)	(0.060)
Post training	-0.14	-0.051	-0.098
	(0.054)	(0.046)	(0.044)
Forecasts * Post	0.17	0.024	0.097
	(0.064)	(0.066)	(0.053)
Observations	1044	1044	1044
Control mean	-0.20	-0.30	-0.25

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 $Y_{it} = \beta_1 \mathit{Forecasts}_i + \beta_2 \mathit{Post}_t + \beta_3 \mathit{Forecasts}_i * \mathit{Post}_t + \boldsymbol{X_i'\delta} + \varepsilon_{it}$, 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.

Table 6: Mechanisms: Information seeking

	Weather seeking	Air quality info seeking	Weather forecast take up	Forecast error index
Forecasts	0.18	0.23	0.0084	0.024
	(0.14)	(0.12)	(0.022)	(0.15)
Training	-0.11	-0.015	-0.021	0.24
	(0.14)	(0.12)	(0.024)	(0.13)
Forecasts + Training	0.17	0.019	-0.012	-0.13
	(0.20)	(0.17)	(0.034)	(0.18)
Forecasts * Took up weather				-0.10
				(0.16)
Training * Took up weather				-0.37
				(0.14)
F + T * Took up weather				0.26
				(0.20)
Observations	981	978	999	999
Control mean	2.73	1.53	0.92	-0.000

Note: SMS forecasts increased seeking of pollution information (column 2). Training reduced forecast error much more for subjects who viewed a weather forecast (column 4, summing coefficients on "Training" and "Training * Took up weather") than for those who did not (the coefficient on "Training"). The interaction of weather forecasts and treatment is more negative for training ("Training * Took up weather") than for SMS forecasts ("Forecasts * Took up weather"), potentially consistent with better information processing among trained subjects. Coefficients are intent-to-treat effects, with the dependent variable indicated in the column heading. In columns one and two dependent variables are counts of information sources. In column three the dependent variable is an indicator for taking up a free weather forecast before making incentivized air pollution forecasts. In column four the dependent variable is a standardized air pollution forecast index, as in column one of Table 1. 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.

Table 7: Outdoor time, effect of receiving forecasts

	Outdoor hours					
Forecasts	0.74	0.068	0.60	-0.41	0.60	0.86
	(0.29)	(0.56)	(0.33)	(0.61)	(0.62)	(1.79)
Forecasts * High pollution	-0.88	1.26	-0.45	1.54	-1.08	1.07
	(0.36)	(0.99)	(0.41)	(0.97)	(0.77)	(3.10)
Forecasts * Cares about air quality		0.69		1.17		-0.10
		(0.64)		(0.69)		(2.00)
Forecasts * High pollution * Cares		-2.27		-2.17		-2.40
		(1.06)		(1.04)		(3.28)
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

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 $Y_i = \beta_F Forecasts_i + \beta_H High\,pollution_t + \beta_{FH} Forecasts_i High\,pollution_t + \gamma Y_{0i} + \boldsymbol{X_i'}\boldsymbol{\delta} + \varepsilon_i$. 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. Columns 1-2 present standard errors that are clustered at the household level. Columns 3-6 present heteroskedasticity-robust standard errors.

For online publication

A Results from the theoretical model

To solve our model, we begin solving backward and consider the problem at time $t = 1$ (the second period). Note that in the appendix, we abuse notation and drop functional arguments for notational simplicity and readability.

A.1 Avoidance purchased after pollution is realized (2nd period)

The state of the world s is known, as is the previously purchased level of avoidance x. The agent's problem is given by

$$u^s(x) = \max_y \{E - d^s(x + y) - c(x, y)\}. \tag{3}$$

Under our assumptions, a unique state-dependent level of avoidance exists, though the two-stage nature of the problem precludes parsimonious assumptions that would ensure it is non-zero. We focus on the cases that yield interior solutions. Then the state-dependent optimal level of avoidance in period 1, $y^s(x)$, is implicitly defined by the first-order condition

$$-d_1^s(x + y^s) - c_2(y^s) = 0. \tag{4}$$

By the implicit function theorem we know that $y_1^s = -\dfrac{-d_{11}^s - c_{12}}{-d_{11}^s - c_{22}} \in [-1, 0]$, as d^s and c are convex in all variables and $0 < c_{11} \leq c_{12} \leq c_{22}$. The results are intuitive given that avoidance actions in the two periods are substitutes, and the marginal cost of avoidance increases in the second period. Finally note that if $a < b$, then $a + y^s(a) \leq b + y^s(b)$. That is, if the agent invests less in the first period, she does not fully make up for it in the second period.

A.2 Avoidance purchased before pollution is realized (1st period)

We now turn our attention to the full ex-ante problem. Given forecasting skill ρ the agent maximizes

$$V(\rho, \pi) = \max_{x^H, x^L} \{\pi[\rho u^h(x^H) + (1 - \rho)u^h(x^L)] + (1 - \pi)[\rho u^l(x^L) + (1 - \rho)u^l(x^H)]\}. \tag{5}$$

Interpreting the above, notice that first the state of the world is determined (with probability π), and then the agent makes a forecast (with skill ρ). The agent's forecast may incorporate external signals. Based on her forecast, the agent chooses her level of avoidance at period 0.

Once the state is realized, she purchases extra avoidance as needed and experiences utility based on the state.

We can transform this bivariate maximization problem into two simpler forecast-dependent problems using Bayes' rule. The value function can alternatively be expressed as

$$V(\rho,\pi) = \varphi\{\max_{x^H}[q^H u^h(x^H) + (1-q^H)u^l(x^H)]\} + (1-\varphi)\{\max_{x^L}[q^L u^l(x^L) + (1-q^L)u^h(x^L)]\},$$

where $\varphi = P(H) = 1-\rho-\pi+2\pi\rho$, $q^H = P(h|H) = \frac{\pi\rho}{\varphi}$ and $q^L = P(l|L) = \frac{\rho(1-\pi)}{1-\varphi}$. This transformation allows us to instead solve an interim problem at time 0 that is a function of the agent's forecast.[76] The result is similar to the rainfall forecasting problem presented in Rosenzweig and Udry (2014b), though with one important distinction. Unlike Rosenzweig and Udry (2014b) we model skill as $P(F|s) = \rho$ (suppressing exogenous variables), while Rosenzweig and Udry (2014b) model it as $P(s|F) = q$, with the quality measure assumed equal for both signals. As can be seen from our formulations of q^F, this assumption is meaningfully restrictive.

We can now solve the agent's problem based on her forecast. Consider the case when she forecasts a high level of pollution. Again, this could be based largely on an external signal. Then her optimization problem is

$$\max_x \left\{ \begin{matrix} q^H[E - d^h(x+y^h(x)) - c(x,y^h(x))] \\ +(1-q^H)[E - d^l(x+y^l(x)) - c(x,y^l(x))] \end{matrix} \right\}. \tag{6}$$

Before continuing, we note that the best-case scenario for the agent is low pollution. Given that the marginal cost of air pollution rises in the second period, it then follows directly that the agent will always pre-purchase at least the optimal level of avoidance for low pollution, $x^l = \text{argmax}_x\, E - d^l(x) - c(x)$. Furthermore in the state with low pollution, the agent will not purchase additional avoidance tomorrow, i.e. $y^l(x^l+x) = 0$ for all $x \geq 0$.

[76] The transformation is a direct application of Bayes' rule and simple algebraic manipulation. We reproduce some of the main steps below

$$\begin{aligned} &\max_{x^H,x^L}\{P(h)P(H|h)u^h(x^H) + P(h)P(L|h)u^h(x^L) + P(l)P(L|l)u^l(x^L) + P(l)P(H|L)u^l(x^H)\} \\ &= \max_{x^H,x^L}\{P(H)P(h|H)u^h(x^H) + P(L)P(h|L)u^h(x^L) + P(L)P(l|L)u^l(x^L) + P(H)P(l|H)u^l(x^H)\} \\ &= \max_{x^H,x^L}\{P(H)[P(h|H)u^h(x^H) + P(l|H)u^l(x^H)] + P(L)[P(h|L)u^h(x^L) + P(l|L)u^l(x^L)]\}. \end{aligned}$$

Then we can re-write equation 6 as

$$\max_{x} \left\{ \begin{array}{c} q^H[E - d^h(x^l + x + y^h(x)) - c(x^l + x, y^h(x))] \\ +(1 - q^H)[E - d^l(x^l + x) - c(x^l + x, 0))] \end{array} \right\}.$$

The first-order condition yields[77]

$$q^H[-d_1^h.(1 + y_1^h) - c_2 y_1^h - c_1] + (1 - q^H)[-d_1^l - c_1] = 0.$$

Rearranging and substituting in the first-order condition for the period 1 problem (equation 4) yields

$$q^H[-d_1^h] + (1 - q^H)[-d_1^l] - c_1 = 0. \tag{7}$$

We do not need to check the second-order condition as this is a simple case of partial minimization. However as we use it later, the second derivative is $-q^H d_{11}^h.(1 + y_1^h) - (1 - q^H)d_{11}^l - c_{11} < 0$, as the damage function and costs are strictly convex, and $y_1^h \in [-1, 0]$.

Equation 7 implicitly defines $x^H(q^H)$, the optimal level of investment given a forecast of high pollution. We can now ascertain the effect of forecast skill on the level of avoidance purchased in advance. By the implicit function theorem, $x_1^H = -\dfrac{-d_1^h + d_1^l}{SOC} \geq 0$, where SOC is the (negative) second order condition and $d_1^h(A) \leq d_1^l(A)\ \forall A$. Symmetric arguments imply that $x_1^L \leq 0$.

Finally, we wish to compare levels of investment based on the signal the agent receives. Under our assumptions on the agent's forecast skill ($\rho \geq \max\{\pi, 1 - \pi\}$), we know that $q^H, q^L \geq \frac{1}{2}$. Then as a first step in our comparison of $x^H(q^H)$ and $x^L(q^L)$, we investigate the artificial case where $q^L = q^H = q$. Let us consider the first-order conditions for both forecasts. For H, we need $q[-d_1^h] + (1-q)[-d_1^l] = c_1$, while for L we require $(1-q)[-d_1^h] + q[-d_1^l] = c_1$. Recall that c is increasing and convex, and that $d_1^h(A) \leq d_1^l(A)$ (equivalently, $-d_1^h(A) \geq -d_1^l(A)$). Then in the case for each forecast, we need the q-weighted average of the slopes of the damage functions to equal the slope of the period 0 cost function. For the high forecast more weight is on the steeper damage function, while the reverse is true for a forecast of low pollution. Coupled with the convexity of the cost function, this implies that $x^H(q) > x^L(q)\ \forall q$. The result is both intuitive and consistent with Rosenzweig and Udry (2014a): a forecast of high pollution, given the same q, should result in higher investment compared to a forecast of low pollution.

[77] Once again, we focus on interior solutions, though it is possible to assume Inada conditions here to ensure interiority.

While intuitive, the result is incomplete, as q^H need not equal q^L. In fact, depending on the value of π, either could be higher.[78] Recall that $x_1^H \geq 0$ and $x_1^L \leq 0$, and consider first the case when $q^H \geq q^L$. Then we have that $x^L(q^L) \leq x^H(q^L) \leq x^H(q^H)$. Similarly, when $q^L \geq q^H$, we have that $x^H(q^H) \geq x^L(q^H) \geq x^L(q^L)$. We have seen, then, that $x^H(q^H) \geq x^L(q^L)$ for all possible cases.[79]

A.3 Willingness to pay for improvements in forecast services

We now turn our attention to willingness to pay for our forecast service, represented within the model as an increase in the agent's forecast quality. Recall that the value function, $V(\rho, \pi)$ is defined by equation 5. Then application of the envelope theorem yields

$$V_1 = \pi[u^h(x^H(q^H)) - u^h(x^L(q^L))] + (1-\pi)[u^l(x^L(q^L)) - u^l(x^H(q^H))].$$

To sign this expression, we need to sign $u^h(x^H) - u^h(x^L)$ and $u^l(x^L) - u^l(x^H)$. Consider the expression $u_1^s = -d_1^s(1+y_1^s) - c_1 - c_2 y_1^s = -d_1^s - c_1$. Taking the second derivative yields $u_{11}^s = -d_{11}^s(1+y_1^s) - c_{11} < 0$, so u^s is concave and attains unique maxima (one per state).

Of interest, however, are not the maxima (as the agent cannot predict the state of the world perfectly), but rather $u_1^h(x^H)$ and $u_1^l(x^L)$. Note that x^H is implicitly defined by $q^H[-d_1^h]+(1-q^H)[-d_1^l] = c_1$ and similarly x^L is implicitly defined by $q^L[-d_1^l]+(1-q^L)[-d_1^h] = c_1$. So given that d^h is steeper than d^l (and both have negative slope), then at x^H, $u_1^h \geq 0$ and at x^L, $u_1^l \leq 0$.[80] This coupled with the fact that $x^H(q^H) \geq x^L(q^L)$, and that both u^s are concave, implies that $V_1 \geq 0$. Hence we know that as the quality of the forecast improves, the individual's utility increases. This implies that willingness to pay for a useful third-party (external) forecast is positive, and increasing in quality.

Before we move to the final step and model the effects of training and our SMS forecast service, we note that the previous results for x^H and x^L provide some useful insights. Compared to a world where the state of air pollution is known ($q^H = q^L = 1$), in a world with imperfect information, the agent under-invests when her forecast is high ($x^H(q^H) \leq \operatorname{argmax}_x u^h(x)$) and over-invests when it is low ($x^L(q^L) \geq \operatorname{argmax}_x u^l(x)$).[81] The result is intuitive, as when the agent forecasts high pollution, she under-invests to benefit

[78] In particular, if $\pi \geq 0.5$, $q^H \geq q^L$, while the reverse is true otherwise.

[79] This is driven by the fact that $\rho \geq \max\{\pi, 1-\pi\}$, which implies $q^L, q^F \geq 0.5$.

[80] To see this more clearly, focus on $u_1^h|_{x^H} = -d_1^h - c_1$. At x^H, $q^H[-d_1^h] + (1-q^H)[-d_1^l] = c_1$, and so c_1 is equal to the weighted average of slopes of d^h and d^l. Then if follows that $d^h$is steeper than c at x^H and so u_1^h is positive. Symmetric arguments apply to u_1^l at x^L.

[81] Another way of seeing this result is by noting that "perfect" forecasts would imply that $q^H = q^L = 1$. Given that $x_1^H \geq 0$ and $x_1^L \leq 0$, imperfect signals yield under- and over-investment, for high and low forecasts respectively.

from the non-zero probability of a low pollution state, and vice versa. This result adds to the set of potential explanations for low mask take-up in developing-country settings with low information and variable pollution.

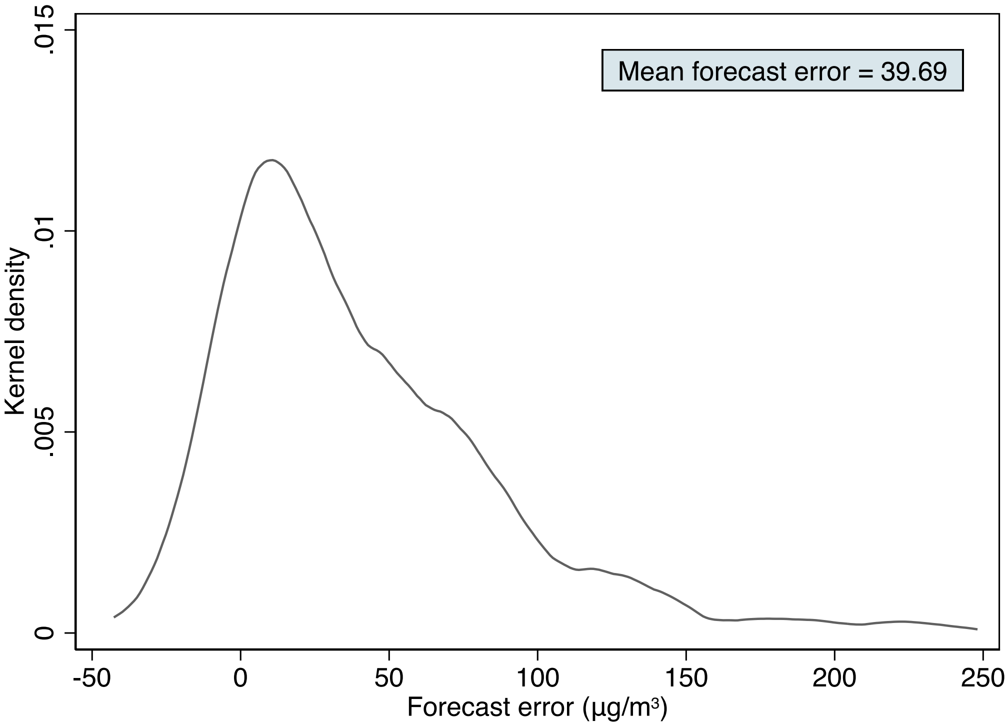
A.4 Effects of information and training

As a final step, we now model the effects of our experimental interventions: SMS forecasts and forecasting training. Recall that ρ is a function of information ι and human capital τ. Within the model, we think of our SMS forecast as an increase in the agent's information. In the extreme case an agent may simply adopt our forecast as her own.[82] Similarly, our training is designed explicitly to increase human capital in the dimension of forecast ability. Both our experimental treatments, then, should improve agents' forecast skill.

[82] Here and throughout the paper, we remain largely agnostic on questions of belief updating.

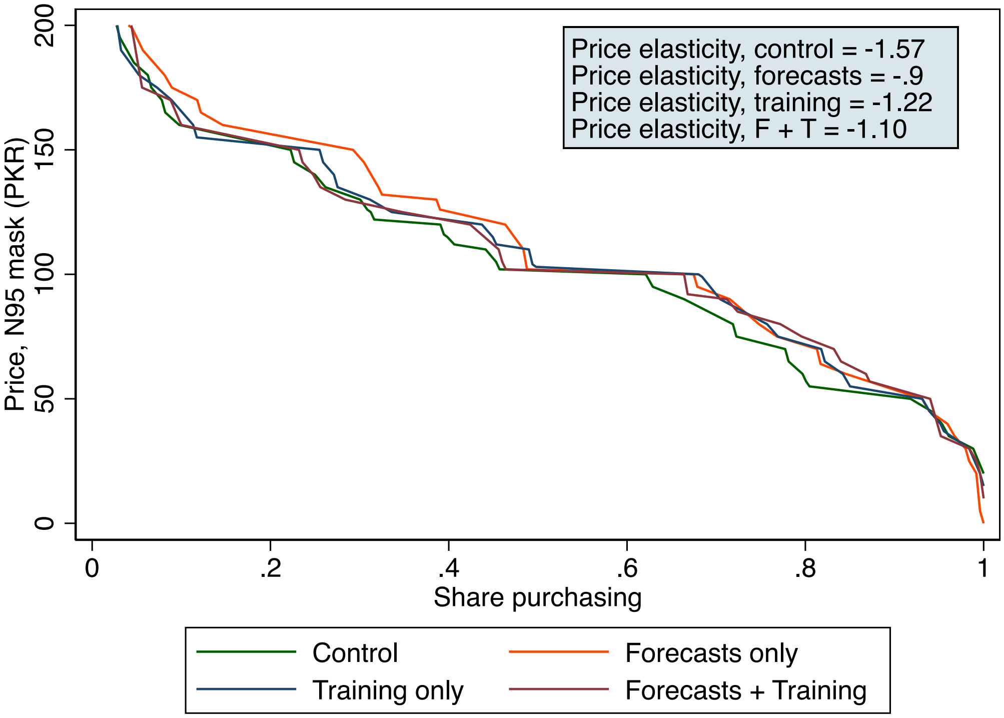
B Additional figures

Figure A1: Air pollution forecast errors (t+1), baseline



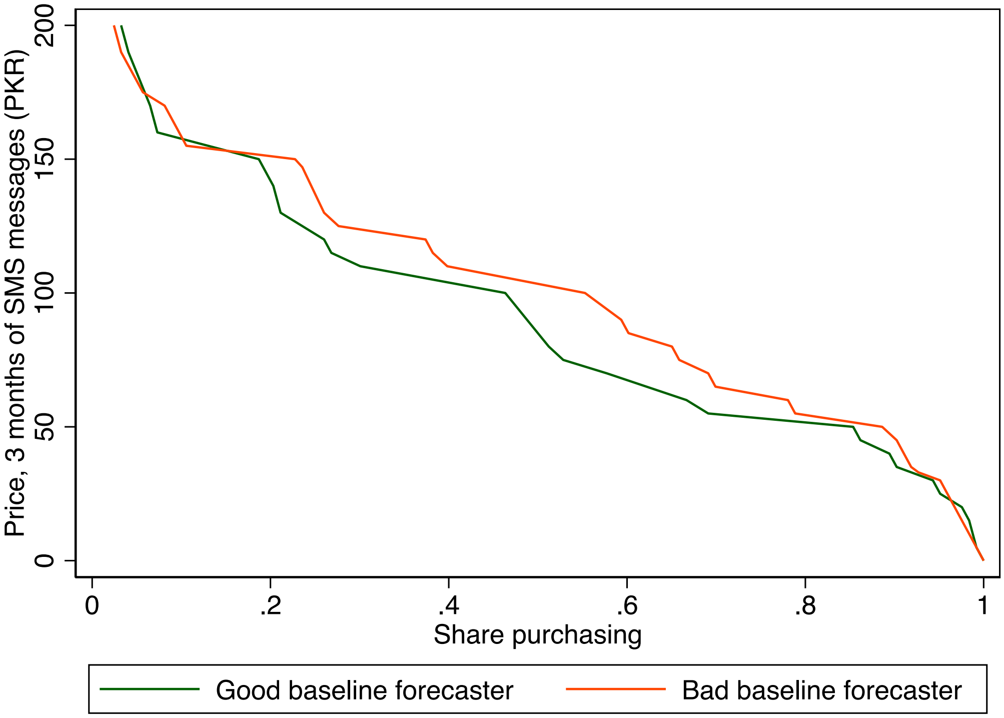
Note: Forecast errors are the differences between subjects' incentive-compatible air pollution forecasts and realized pollution on the day after the baseline survey. That is, a negative error represents an underprediction of pollution. Units are $\mu g/m^3$, rather than control-group standard deviations as in most exhibits in this paper. Density was estimated under Stata-default kernel and bandwidth. The sample includes all baseline respondents.

Figure A2: Demand curves for N95 masks, endline



Note: Willingness to pay (WTP) was elicited at endline using a Becker-DeGroot-Marschak mechanism (Becker, DeGroot, and Marschak, 1964), in which all subjects bid on an N95 mask with a retail price of 135 PKR. The maximum bid was 200 Pakistani Rupees (PKR). Quantity demanded is expressed as the share of subjects purchasing; that is, the share with WTP greater than or equal to a given price. Local elasticities near the retail price appear in Table A11.

Figure A3: Demand curves for air pollution forecasts, by baseline forecast error



Note: Willingness to pay (WTP) was elicited at endline using a Becker-DeGroot-Marschak mechanism (Becker, DeGroot, and Marschak, 1964), in which all subjects bid on 90 additional days of our SMS air pollution forecasts. The maximum bid was 200 Pakistani Rupees (PKR). The figure reflects the forecast-only treatment group (246 subjects), as explained in Section 5.1. Quantity demanded is expressed as the share of subjects purchasing; that is, the share with WTP greater than or equal to a given price. "Good baseline forecaster" denotes subjects with baseline air pollution forecast error (t+1) above the group median, while "Bad baseline forecaster" denotes subjects with error below the median.

C Additional tables

Table A1: Treatment-control balance, full baseline sample

	Control	Forecast	Training	Forecasts + Training	P-value
Age of respondent (years)	31.643	30.555	30.559	31.776	0.346
	(0.663)	(0.608)	(0.633)	(0.647)	
Respondent female (=1)	0.515	0.504	0.496	0.482	0.888
	(0.030)	(0.030)	(0.030)	(0.030)	
# of household members	5.493	5.603	5.676	5.864	0.482
	(0.161)	(0.145)	(0.143)	(0.182)	
# of elderly in household	0.404	0.397	0.408	0.441	0.886
	(0.041)	(0.042)	(0.042)	(0.042)	
# of children in household	1.680	1.952	1.746	1.941	0.173
	(0.099)	(0.112)	(0.103)	(0.116)	
A household member has a respiratory disease	1.857	1.846	1.824	1.853	0.729
	(0.021)	(0.022)	(0.023)	(0.022)	
# of employed household members	1.728	1.691	1.846	1.820	0.185
	(0.060)	(0.054)	(0.061)	(0.062)	
Cares about air quality (likert)	3.588	3.647	3.632	3.705	0.577
	(0.063)	(0.059)	(0.058)	(0.057)	
Aware of the air quality in Lahore (likert)	3.226	3.279	3.255	3.350	0.562
	(0.070)	(0.062)	(0.064)	(0.063)	
Aware of the air quality in Walton (likert)	2.570	2.513	2.543	2.625	0.837
	(0.096)	(0.089)	(0.087)	(0.092)	
# of times/week checks the weather	1.823	1.918	1.910	2.056	0.381
	(0.093)	(0.094)	(0.101)	(0.098)	
Believes n95 masks work (=1)	0.939	0.916	0.922	0.933	0.764
	(0.016)	(0.018)	(0.018)	(0.016)	
Household owns a car (=1)	1.955	1.944	1.944	1.952	0.917
	(0.013)	(0.014)	(0.014)	(0.013)	
# of mobile phones household owns	2.632	2.592	2.794	2.823	0.294
	(0.077)	(0.081)	(0.163)	(0.110)	
Observations	272	272	272	272	
F statistic	1.7	1.6	1.5	1.1	

Note: Means and heteroskedasticity-robust standard errors reported. P-values are from joint F tests of treatment orthogonality with respect to listed observables. F statistics are from joint tests of regression coefficients on observables.

Table A2: Attrition rates by experimental condition

	Control	Forecast	Training	Forecasts + Training	P-value
Attrited from endline dummy	0.059	0.096	0.092	0.081	0.333
	(0.014)	(0.018)	(0.018)	(0.017)	
Observations	272	272	272	272	

Note: Means and heteroskedasticity-robust standard errors reported. P-value is from a joint F test of treatment orthogonality with respect to an endline attrition indicator.

Table A3: Treatment-control balance, non-attritors

	Control	Forecast	Training	Forecasts + Training	P-value
Age of respondent (years)	31.746	30.671	30.308	31.580	0.346
	(0.685)	(0.649)	(0.665)	(0.660)	
Respondent female (=1)	0.500	0.472	0.482	0.484	0.936
	(0.031)	(0.032)	(0.032)	(0.032)	
# of household members	5.527	5.606	5.563	5.892	0.490
	(0.169)	(0.145)	(0.142)	(0.194)	
# of elderly in household	0.410	0.382	0.405	0.436	0.856
	(0.043)	(0.043)	(0.044)	(0.044)	
# of children in household	1.656	1.919	1.741	1.924	0.188
	(0.101)	(0.097)	(0.109)	(0.120)	
A household member has a respiratory disease	1.852	1.846	1.826	1.856	0.818
	(0.022)	(0.023)	(0.024)	(0.022)	
# of employed household members	1.734	1.687	1.838	1.824	0.213
	(0.063)	(0.054)	(0.063)	(0.065)	
Cares about air quality (likert)	3.613	3.695	3.623	3.711	0.568
	(0.064)	(0.060)	(0.061)	(0.059)	
Aware of the air quality in Lahore (likert)	3.227	3.321	3.282	3.349	0.589
	(0.072)	(0.063)	(0.066)	(0.065)	
Aware of the air quality in Walton (likert)	2.565	2.558	2.569	2.592	0.993
	(0.098)	(0.093)	(0.093)	(0.094)	
# of times/week checks the weather	1.824	1.955	1.975	2.012	0.537
	(0.095)	(0.100)	(0.106)	(0.101)	
Believes n95 masks work (=1)	0.940	0.912	0.919	0.926	0.709
	(0.016)	(0.019)	(0.019)	(0.018)	
Household owns a car (=1)	1.952	1.938	1.942	1.952	0.884
	(0.014)	(0.015)	(0.015)	(0.014)	
# of mobile phones household owns	2.645	2.565	2.838	2.815	0.214
	(0.080)	(0.074)	(0.178)	(0.117)	
Observations	256	246	247	250	
F statistic	1.4	1.4	1.4	1.3	

Note: Means and heteroskedasticity-robust standard errors reported. P-values are from joint F tests of treatment orthogonality with respect to listed observables. F statistics are from joint tests of regression coefficients on observables.

Table A4: Balance, non-attritors, primary outcomes at baseline

	Control	Forecast	Training	Forecasts + Training	P-value
Forecast error index (baseline)	-0.002	0.017	0.108	0.014	0.577
	(0.060)	(0.062)	(0.065)	(0.062)	
WTP: Masks (baseline)	90.000	89.110	89.615	89.640	0.994
	(2.241)	(2.209)	(2.335)	(2.270)	
Avoidance index (baseline)	0.005	-0.013	0.093	-0.047	0.249
	(0.050)	(0.050)	(0.052)	(0.049)	
Happiness variance (baseline)	2.803	2.746	2.703	2.611	0.158
	(0.062)	(0.064)	(0.068)	(0.061)	
Observations	256	246	247	250	
F statistic	0.16	1.89	1.29	1.70	

Note: Means and heteroskedasticity-robust standard errors reported. P-values are from joint F tests of treatment orthogonality with respect to listed observables. F statistics correspond to regressions of group dummies on all baseline measures of primary outcomes. Willingness to pay for our forecast messages was not elicited at baseline by design.

Table A5: Willingness to pay for SMS air pollution forecasts

	WTP: Forecast	WTP: Forecast
Forecasts	4.46	5.34
	(4.22)	(3.60)
Training	-1.22	2.42
	(4.08)	(3.55)
Forecasts + Training	-3.08	-5.15
	(5.72)	(4.95)
Constant	88.8	109.0
	(2.96)	(16.0)
Observations	999	999
Forecasts group mean	93.22	
	(3.00)	
	[0.00]	

Note: Column 1 reports a regression of forecast WTP (in PKR) on a constant term and the three treatment dummies. This allows a test of the mean in the forecast-only group in a regression context. As pre-specified, we conduct a right-tailed test of the sum of the Constant and the Forecasts coefficient against a zero null and report the result at the bottom of column 1, with the resulting p-value in square brackets. Note that because block dummies are not included in column 1, treatment effects are not identified and estimates should not be interpreted causally. Column 2 reports estimates corresponding to Equation 2, with forecast WTP as the outcome. Randomization block dummies are included. A pre-specified LASSO procedure was used to select additional controls. Heteroskedasticity-robust standard errors are in parentheses.

Table A6: Effects on absolute forecast error in $\mu g/m^3$

	Forecast error
Forecasts	-4.12
	(2.64)
	[0.06]
Training	-6.26
	(2.63)
	[0.01]
Forecasts + Training	6.11
	(3.71)
	[0.1]
Observations	999
Control mean	64.6

Note: Specification is as in column 1 of Table 1, but with average absolute error (t+1 and t+3) denominated in $\mu g/m^3$, rather than control-group standard deviations. As in column 1 of Table 1, tests of the Forecasts and Training estimates are left-tailed ($\beta_F < 0, \beta_T < 0$). The resulting p-values appear in square brackets.

Table A7: Alt. MHT-adjusted ITT p-values, 2-tailed tests for mask WTP and avoidance

	WTP: Forecasts	Forecast error index	Happiness variance	WTP: Masks	Avoidance index
Forecasts	-	0.10	-	0.10	0.42
Training	-	0.03	0.74	0.27	0.74
Forecasts + Training	-	-	-	-	-
Mean, forecast-only group	0.001	-	-	-	-

Note: Alternative tests for mask WTP and the avoidance index are two-tailed, rather than right-tailed as in Table 2. The p value in column 1 corresponds to the test of mean willingness to pay for forecasts in the forecast-only treatment group. This test is illustrated in Figure 2 and formalized at the bottom of column 1, Table A5. The p values in columns 2 and 3 correspond to the pre-specified estimates in Table 1 (shaded cells). The MHT correction is performed using the procedure of Benjamini, Krieger, and Yekutieli (2006), which controls the false discovery rate. As discussed in Section 4.3, the resulting corrected p-values can be larger or smaller than their uncorrected analogs.

Table A8: Effects on t+1 non-absolute, non-standardized forecast error in $\mu g/m^3$

	Forecast error (t + 1), $\mu g/m^3$
Forecasts	6.41
	(4.06)
Training	2.14
	(3.96)
Forecasts + Training	-3.49
	(5.52)
Observations	999
Control mean	-39.6

Note: Specification is as in column 1 of Table 1. Forecast errors are the differences between subjects' incentive-compatible air pollution forecasts and realized 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. Heteroskedasticity-robust standard errors are in parentheses.

Table A9: Standard deviation of t+1 air pollution forecast errors, by group

	Standard deviation	P-value
Control	75.24	
Forecasts only	72.51	0.99
Training only	60.90	0.03
Forecasts + Training	67.49	0.42

Note: Standard deviations are computed from non-absolute, non-standardized forecast error in $\mu g/m^3$ (see also note below Table t.forecast-error-alt-outcomes). P-values correspond to the Brown and Forsythe (1974) median-based robust test statistic for the equality of variances between control group respondents and respondents in each of the three treatment groups.

Table A10: Air pollution forecast errors, beginning and end of training, endline sample

	Forecast error (t + 1)	Forecast error (t + 3)	Forecast error idx
Post training	-0.15	-0.060	-0.11
	(0.058)	(0.049)	(0.046)
Forecasts	-0.12	0.018	-0.052
	(0.070)	(0.077)	(0.063)
Forecasts * Post	0.16	0.021	0.091
	(0.067)	(0.069)	(0.056)
Observations	968	968	968
Control mean	-0.19	-0.28	-0.24

Note: Unlike Table 5, this table excludes subjects who attrited between training and endline. The sample is comprised of 2 observations for each of the 484 trained subjects who completed the endline survey. 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 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.

Table A11: Price elasticity of air pollution masks near market price

Bin width	20	30	40
Price elasticity, control	-2.13	-2.40	-2.32
Price elasticity, forecasts only	-2.44	-2.27	-2.01
Price elasticity, training only	-2.04	-2.58	-2.33
Price elasticity, F + T	-2.98	-3.08	-2.65

Note: Willingness to pay (WTP) was elicited at endline using a Becker-DeGroot-Marschak mechanism (Becker, DeGroot, and Marschak, 1964), in which all subjects bid on an N95 mask with a retail price of 135 PKR. The maximum bid was 200 Pakistani Rupees (PKR). Elasticities are estimated from a log-log regression in which price interacts with a set of 3 bin indicators. Bin indicators also enter in non-interacted form. "Bin width" refers to the bandwidth of the central bin containing the market price of 135 PKR. That is, a bin width of 20 implies a price elasticity estimated over the range from 125 to 145 PKR. Complete demand curves and average elasticities appear in Figure A2.

Table A12: Outdoor time, effect of receiving forecasts, including all treatments

	Outdoor hours	
Forecasts	0.83	-0.25
	(0.41)	(0.71)
Forecasts * High pollution	-0.96	1.62
	(0.50)	(1.27)
Forecasts * Cares about air quality		0.85
		(0.87)
Forecasts * High pollution * Cares		-2.46
		(1.39)
Training	-0.69	-0.14
	(0.45)	(0.87)
Training * High pollution	0.59	0.51
	(0.56)	(1.20)
Training * Cares about air quality		-1.15
		(1.03)
Training * High pollution * Cares		0.58
		(1.39)
Forecasts + Training	-0.15	0.65
	(0.60)	(1.18)
(Forecasts + Training) * High pollution	0.10	-0.82
	(0.74)	(1.85)
(Forecasts + Training) * Cares about air quality		-0.38
		(1.37)
(Forecasts + Training) * High pollution * Cares		0.51
		(2.02)
Observations	1442	1442
Control mean	4.74	4.74
Adult and/or child time?	Both	Both

Note: Column 1 corresponds to a variant of the equation employed in Table 7, with the addition of the training treatment and the interaction of the two treatments. Column 2 adds triple interactions with a baseline indicator for caring about air quality; this indicator also enters in non-interacted and double-interacted control terms. 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). All columns include randomization block indicators. A pre-specified LASSO procedure was used to select additional controls. Standard errors in parentheses are clustered at the household level.

Table A13: Covariances with caring about air quality

	WTP: masks (Baseline)	Avoidance index (Baseline)	SMS views per week (Endline)	WTP: forecast (Endline)
Cares about air quality	2.67	0.44	0.52	10.2
	(2.91)	(0.051)	(0.28)	(4.06)
Observations	1088	1088	496	999
Control mean	86.9	-0.37	2.58	80.9

Note: Estimates correspond to regressions of the dependent variable indicated in the column heading on a constant and a dummy variable for caring about air quality at baseline. Units are in Pakistani rupees (PKR) in columns 1 and 4. Units are standard deviations in column 2. Units are an average count of views per week in column 3. Heteroskedasticity-robust standard errors are in parentheses.

Table A14: Outdoor time, effect of receiving forecasts, without LASSO procedure

	Outdoor hours					
Forecasts	0.57	0.11	0.61	-0.42	0.46	1.85
	(0.38)	(0.76)	(0.43)	(0.83)	(0.69)	(1.78)
Forecasts * High pollution	-0.74	1.02	-0.63	1.60	-0.99	-0.92
	(0.47)	(1.19)	(0.53)	(1.19)	(0.87)	(2.85)
Forecasts * Cares about air quality		0.60		1.36		-1.86
		(0.87)		(0.94)		(2.03)
Forecasts * High pollution * Cares		-2.03		-2.68**		0.18
		(1.27)		(1.29)		(3.06)
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

Note: The initial estimating equation (odd columns) is $Y_i = \beta_F \mathit{Forecasts}_i + \beta_H \mathit{High\,pollution}_t + \beta_{FH} \mathit{Forecasts}_i \mathit{High\,pollution}_t + \gamma Y_{0i} + \boldsymbol{X_i'}\boldsymbol{\delta} + \varepsilon_i$. The dependent variable is outdoor time in hours, elicited as part of a 24-hour time diary. $\mathit{High\,pollution}_t$ 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. No LASSO-selected controls are included. All columns include randomization block indicators. Columns 1-2 present standard errors that are clustered at the household level. Columns 3-6 present heteroskedasticity-robust standard errors.

Table A15: Effects on a social desirability index

	Social desirability index
Forecasts	-0.139
	(0.126)
Training	-0.00642
	(0.123)
Forecasts + Training	0.0388
	(0.179)
Observations	998
Control mean	0.00

Note: Estimates correspond to Equation 2, with the dependent variable indicated in the column heading. All columns include randomization block indicators. A pre-specified LASSO procedure was used to select additional controls. Heteroskedasticity-robust standard errors are in parentheses.

Table A16: Spillover frequencies, non-SMS groups

Num. spillover messages	Num. HH
1-9	22
10-24	6
25-49	2
50+	1

Note: Responses were collected at endline from subjects outside the SMS forecast message group: the pure control group and the training-only group. Subjects were shown an image of one of our messages and asked if they had received any such messages.

Table A17: Primary results, no baseline outcome control & no LASSO-selected controls

	Forecast error index	Happiness variance	WTP: Masks	Avoidance index
Forecasts	-0.050	-0.0066	6.79	-0.029
	(0.066)	(0.083)	(3.80)	(0.071)
	[0.23]	[0.47]	[0.04]	[0.65]
Training	-0.13	-0.0081	3.92	-0.078
	(0.065)	(0.082)	(3.82)	(0.070)
	[0.03]	[0.46]	[0.15]	[0.87]
Forecasts + Training	0.078	-0.030	-7.58	0.085
	(0.092)	(0.12)	(5.41)	(0.10)
	[0.40]	[0.40]	[0.16]	[0.40]
Observations	999	995	999	999
Control mean	0.00	0.00	104.1	0.00

Note: This is a variant of Table 1 that omits controls for baseline outcomes and LASSO-selected controls. All columns include randomization block indicators. 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 index ($\beta_F < 0, \beta_T < 0$), 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.

Table A18: Predicted Willingness to Pay using Variables Available in the 2011 Indian Census

	WTP: Forecasts	
# HH members	-4.86	-4.77
	(1.27)	(1.28)
# of rooms in the house	12.5	12.4
	(2.35)	(2.37)
HH owns a car	-28.8	-28.8
	(7.70)	(7.73)
HH owns a computer	-8.24	-8.90
	(6.81)	(7.02)
HH owns a motorcycle	-17.3	-17.7
	(13.0)	(13.0)
HH owns a television	24.8	23.7
	(13.2)	(13.6)
Constant	78.8	78.3
	(19.4)	(19.6)
Baseline forecast error (t+1)		0.046
		(0.069)
Observations	242	242
R-squared	0.12	0.12
F-stat	7.84	6.80

Note: These variables were selected from a slightly larger set of overlapping demographics through the use of a penalized LASSO regression. Sample is limited to households who received the forecast treatment only.

Table A19: Predicted Willingness to Pay using Variables Available in the MICS

	WTP: Forecasts	
HHs owns their house	2.71	2.08
	(7.32)	(7.50)
# HH members	-4.79	-4.72
	(1.26)	(1.27)
# of rooms in the house	11.4	11.4
	(2.55)	(2.56)
HHs owns a car	-31.1	-31.1
	(8.31)	(8.32)
HH owns a computer	-11.8	-12.3
	(7.22)	(7.36)
HH owns a motorcycle	-17.4	-17.6
	(13.4)	(13.5)
HH owns a television	25.5	24.4
	(13.1)	(13.6)
HH owns an air conditioner	11.6	11.5
	(9.84)	(9.88)
Constant	78.7	78.4
	(19.2)	(19.5)
Baseline forecast error (t+1)		0.040
		(0.070)
Observations	242	242
R-squared	0.13	0.13
F-stat	5.72	5.18

Note: These variables were selected from a slightly larger set of overlapping demographics through the use of a penalized LASSO regression. Sample is limited to households who received the forecast treatment only.

Table A20: Estimated Average Willingness to Pay in 25 Most Polluted Cities

Pollution Rank	City	Estimated Average WTP in PKR	Estimated Total WTP in USD
1	Lahore, Pakistan	80.3	12,674,713
2	Hotan, China	N/A	N/A
3	Bhiwadi, India	86.9	875,370
4	Delhi (NCT), India	91.7	21,851,697
5	Peshawar, Pakistan	70.4	1,968,422
6	Darbhanga, India	86.1	361,802
7	Asopur, India	79.0	6,395
8	N'Djamena, Chad	104.8	2,368,224
9	New Delhi, India	89.0	1,482,997
10	Patna, India	94.7	2,264,055
11	Ghaziabad, India	89.0	2,082,740
12	Dharuhera, India	95.0	40,918
13	Baghdad, Iraq	94.13	5,132,407
14	Chapra, India	89.9	258,217
15	Muzaffarnagar, India	90.3	503,434
16	Faisalabad, Pakistan	79.5	3,622,530
17	Greater Noida, India	90.9	131,678
18	Bahadurgarh, India	95.4	231,244
19	Faridabad, India	89.7	1,800,427
20	Muzaffarpur, India	90.9	457,353
21	Noida, India	86.9	786,073
22	Jind, India	96.0	228,372
23	Karagandy, Khazakstan	92.0	655,368
24	Charkhi Dadri, India	96.5	688,000
25	Rohtak, India	95.4	506,848

Note: Average Willingness to Pay is estimated following the procedure explained in Section 5.6. Data for Indian cities comes from the 2011 Indian Census. Data for other cities come from the latest round of the World Bank's Multiple Indicator Cluster Survey (MICS). Total Willingness to Pay is simply the average for a city multiplied by 365/90 to convert to a yearly WTP, then multiplied by the estimated population of each city according to UNdata (data.un.org), then multiplied by the current exchange rate between PKR and USD.

D Data, sampling, and randomization details

D.1 Sampling and subjects

Located in the province of Punjab, Lahore is Pakistan's second largest city by population. The Pakistan Bureau of Statistics divides Lahore's population of 11.1 million into 8 Tehsils (subdistricts). We use data from the 2011 Multiple Indicator Cluster Survey (MICS) to compare Walton (one of our selected Tehsils) to the rest of Lahore on key indicators.[83] On average, residents of Walton are slightly more educated and wealthier than residents of Lahore as a whole. For example, 27 percent of household heads have some tertiary education, compared to 18.5 percent overall in Lahore. Households in Walton are also slightly more likely to include older relatives. Using data from our pilot surveys and insights from previous surveys in Lahore, we selected two Tehsils for our survey: Walton and Model Town.[84]

To collect data on outcomes and covariates we surveyed subjects in the Walton area of Lahore at multiple points in time. Survey enumerators collected all the primary data on electronic tablets using SurveyCTO's Open Data Kit (ODK) server.

We used 7 charges for the study. Between 140 and 180 households per charge were surveyed, giving a total of 1088 respondents in 7 charges. This was accomplished by using a GIS-based system to construct 190 meter by 190 meter grid cells within each charge and selecting up to 19 survey points within each charge. The grid buffer ensured that our survey points were at least 190 meters from each other. We then drew 128 random GPS points across the entire sampling frame of 7 charges.

To select households within each charge, a pin was dropped at a random point. A pair of enumerators proceeded to the pin and selected the nearest household to the left for the first survey. The enumerators then selected nine other households using the *left hand rule*: every fifth household on the left, proceeding in a clockwise spiral fashion. Each enumerator pair surveyed 5 male and 5 female subjects at each survey point, for a total of 10 respondents. This ensured the gender distribution in the sample would match the population. Households were excluded from the sample if the dwelling was locked/empty, all members of the household were below 18 or above 60 years of age, members were not willing to subscribe to our SMS service, or the household refused to participate in the study. In any of these situations, the enumerator skipped the dwelling, recorded the reason for refusal, and selected the next closest neighbor for the survey. For each household, respondent gender was chosen using a

[83] The MICS data does not distinguish Model Town (our other selected Tehsil) from other Tehsils.

[84] To draw our sample of 1088 households within these Tehsils, we included 6 out of the 11 charges (sub-subdistricts) of Walton and 1 charge of Model Town. The excluded charges included restricted military and high-income areas, where low response rates were expected. The sampling frame for this experiment encompassed 7 charges, 41 circles and 231 census blocks.

pre-generated random list.

Within the household, all members were listed according to their status. A random number generator programmed in the survey tablet was then used to select a household member using a three step process. First, the set of household members was restricted to the eligible population;[85] Second, a random number was generated for each member. Members who were either household heads or spouses of household heads were pre-selected by allocating them a probability of 1, while all other members were assigned equal probability of being randomly selected. Third, the random numbers were used to select the *n*th household member. The enumerator then asked to speak with the *n*th listed eligible individual to conduct the baseline survey, conditional on oral consent.

D.2 Baseline survey (core modules)

The following modules were included in our baseline survey:

1. Information and trust;
2. Willingness to pay for particulate-filtering masks;
3. Air pollution forecast elicitation;
4. Air pollution-related attitudes and behavior;
5. Time use of the respondent and the youngest physically active child;
6. Risk aversion elicitation;
7. Political preference elicitation;
8. Demographics.

D.3 Survey frequency

Data were collected at two stages of the experiment.

1. **In-person surveys:** In-person baseline and endline surveys of all respondents were conducted.
2. **Treatment survey:** For each individual in the forecast training treatment groups (groups T2 and T3 in Figure 1), we conducted an in-person training session, which allowed us to collect additional survey data.

[85] The eligibility criteria were: (i) ages of 18-60 years; (ii) willingness to receive our SMS forecast messages and our forecast training; and (iii) presence in the dwelling at the time of the survey.

D.4 Air pollution data

1. **AQMesh and Dusttrak II:** We used two industrial-grade monitors: (1) the AQMesh; and (2) the Dusttrak II.[86] We installed the AQMesh on the roof of a house in central Walton. It transmitted air pollution readings via GSM continuously and data were accessed through an API. The Dusttrak II is a handheld device that a research assistant used to manually take readings in Walton 2-3 times a day, following a written protocol.

2. **AirNow International:** U.S. EPA's AirNow program is a repository of real-time air quality data and forecasts for the United States. AirNow International is a global version of the U.S.-based air quality data management and display system. It provides hourly data on $PM_{2.5}$ levels. We regularly scraped this data from the AirNow website.[87]

3. **MeteoBlue:** MeteoBlue uses nonhydrostatic mesoscale and multi-scale weather models, which we operated at resolutions between 40 km. For air quality data, MeteoBlue makes use of forecast data from the European Commission and the ECMWF (European Centre for Medium-Range Weather Forecasts).[88] MeteoBlue uses this third-party data to source its predictions and issues them from an atmospheric model with a 40 km resolution. We updated these predictions everyday at UTC 06:00, 10:00, 12:00 and 18:00 to include them in our secondary data.

4. **SPRINTARS:** Spectral Radiation-Transport Model for Aerosol Species (SPRINTARS) is a numerical model which estimates the effect of aerosols on the climatic system and its contribution to global air quality.[89] The Climate Change Science Section at the Research Institute for Applied Mechanics, Kyushu University in Fukuoka, Japan primarily developed the model. SPRINTARS uses aerosols from both natural and anthropogenic sources to estimate categories for SPM, PM 10 and $PM_{2.5}$. We used the forecasts generated from this model in our secondary data on air quality forecasts.

[86] The AQMesh is a small-sensor air quality monitoring system for measuring outdoor and indoor air quality. Details on the product can be found here: https://www.aqmesh.com/product/. The Dusttrak-II is a battery-powered handheld aerosol monitor. Details of the device can be found here: https://www.tsi.com/dusttrak-ii-aerosol-monitor-8532/.

[87] One can obtain the data from the following link after selecting Lahore as a city from the drop-down menu: https://airnow.gov/index.cfm?action=airnow.global_summary.

[88] Details about the ECMWF model can be found here: https://www.ecmwf.int/en/forecasts.

[89] Details about the SPRINTARS model can be found here: https://sprintars.riam.kyushu-u.ac.jp/forecast.html.

D.5 Weather data

- **AccuWeather:** AccuWeather is a popular source of weather forecasts. It takes the U.S. National Oceanic and Atmospheric Administration's (NOAA) weather forecasts and transforms them for general consumers. Weather forecast data from Accuweather were scraped each day for the city of Lahore.[90] Data included temperature levels, precipitation levels and cloud cover.

D.6 Randomization details

Stratification and randomization were performed in R using the commands in the *blockTools* package (Moore 2012), which allows for blocking on a high-dimensional set of covariates and avoids discretizing continuous covariates. For robustness (in terms of block stability) to outliers, we generated multivariate location and spread using a Minimum Volume Elipsoid (MVE) estimator. Robustness to outliers was important in our setting because pilot surveys yielded very large forecast errors for some respondents. In computing the MVE, we weighted incentive-compatibly elicited baseline outcomes twice as heavily as other covariates. While the exact magnitudes of these weights were admittedly ad hoc, they made explicit our prior that baseline outcomes should predict endline outcomes better than other covariates. Per the recommendation of Athey and Imbens (2017), blocks contained eight subjects. We performed blocking using the optimal-greedy algorithm implemented in the *block* command. Within each block, we randomly assigned two subjects to each experimental condition (forecasts, training, forecasts and training, control).

D.6.1 Primary treatment

Subjects were stratified on risk aversion, air pollution forecast error (*t+1* and *t+3*), travel time forecast error (*t+1* and *t+3*), and willingness to pay for a particulate-filtering mask. We elicited these variables using incentive-compatible mechanisms as part of the baseline survey. We further stratified subjects on several self-reported variables: having rescheduled activities in response to air pollution in the past week, informedness about air pollution, household health risk from air pollution,[91] education, gender, age, and a dummy for having provided a subsequently verified phone number at baseline. For additional details on randomization, see Section D.6

[90] https://www.accuweather.com/en/pk/lahore/260622/daily-weather-forecast/260622.

[91] This measure was calculated as the first principal component of three indicators: presence of a household member with breathing problems, presence of children in the household, and presence of elderly people in the household.

E Intervention details

E.1 Day-ahead air pollution forecasts

We designed an ensemble model to forecast day-ahead (*t+1*) $PM_{2.5}$ air pollution: the concentration of particulates of diameter 2.5 microns or less, measured in micrograms per cubic meter ($\mu g/m^3$). Our ensemble forecast combined the following models.[92]

1. **Model based on data from our own air pollution monitors**
   This model used as inputs: (1) average daily $PM_{2.5}$ readings from one or both of our industry qualified air pollution monitors deployed in the Walton neighborhood (our study area) of Lahore; and (2) AccuWeather *t+1* forecasts for minimum temperature, maximum temperature, and precipitation in inches. The two monitors were: (1) an AQMesh; and (2) a Dusttrak II. We installed the AQMesh on the roof of a house in central Walton and it transmitted air pollution readings continuously via GSM. We then accessed these readings through an API. The Dusttrak II is a handheld device that a research assistant used to manually take readings in Walton 2 to 3 times a day under a fixed protocol. We predicted *t+1* $PM_{2.5}$ levels through an MA7 model with day of the week fixed-effects and weather forecast controls. The MA7 form was selected using a cross-validation exercise applied to our data.

2. **Model based on data from the US Consulate's air pollution monitor**
   This model was identical to the model based on our data, but used data from AirNow—a ground monitor located at the US consulate in Lahore.

3. **MeteoBlue and SPRINTARS models**
   These models offer publicly available air pollution forecasts based on satellite data. We accessed *t+1* forecasts at 5pm each day.

We combined the models above through a simple three step process: first, we designated retrospective data from our air pollution monitor(s) as the "ground truth" and we demeaned each of the other models (including our own prediction models) according to the differences between the predictions in these models and the ground truth over the prior week; second, we measured the root-mean squared error of each model relative to the ground truth over the prior week; and third, we took an average of the predictions for *t+1*, inversely weighted by each model's root-mean squared error.

[92] We describe the data sources listed below in greater detail later in Section D.4.

We employed an API-based SMS messaging service that used a short code to send SMS messages to our survey participants in Treatment Groups *1* and *3*.[93] The use of a short code allowed the participants to reply to our forecast messages with any queries, enabling some interaction on text messages as well. We sent our treatment group respondents two pieces of information: 1) an average $PM_{2.5}$ air pollution forecast for *t+1*; and 2) realized average $PM_{2.5}$ air pollution level for the previous day (*t-1*). The latter was intended to allow subjects to assess the accuracy of our forecasts.

E.2 Forecast Training

We implemented a one-hour forecast training based on the principles of Tetlock (2017) and Kahneman (2011). In particular we drew on the findings of Mellers et al. (2014) and Mauboussin and Callahan (2015), but no material was taken directly from this work. Broadly speaking, the training aimed to reduce behavioral and psychological mistakes that decrease the precision and accuracy of subjects' forecasts. Training took place in subjects' homes. A group of specially selected and trained enumerators conducted the trainings in Urdu.[94] Subjects received 150 PKR for their participation.

Each training session began with incentivized elicitations of air pollution forecasts. Over the course of the session, we elicited non-incentivized forecasts of the same outcomes to allow evaluation of individual training exercises. At the end of the session, we again elicited incentivized forecasts. This structure allows us to measure within-subject changes in forecast ability over the training session.

The first set of training exercises covered the concept of calibration. Participants provided 80 percent confidence intervals for $PM_{2.5}$ readings over the previous five days and then answered numerical questions about Pakistan's history and culture (for example, "what is the population of Islamabad?"). For each answer, the subjects provided a confidence level: the probability that their answer fell within a given range around the truth. In the third calibration exercise, the subjects answered "true or false" general knowledge questions and provided confidence levels for each answer. In pilot sessions, most subjects made large errors and demonstrated overconfidence, consistent with evidence from developed countries (Mellers et al., 2014). The calibration exercises were intended to show subjects that they had room for improvement and open their minds to subsequent lessons.

[93] A short code is a four digit telephone number (shorter than a full phone number) employed to send and receive SMS and MMS messages over mobile phones. In the local context, banks, public institutions, and accredited private organizations use short codes to share messages with their clients. The Pakistan Telecommunication Authority (PTA) follows a rigorous procedure to grant access to short codes. We obtained the short code "8755" to deliver SMS messages to our survey participants in groups *1* and *3*.

[94] Urdu is one of the primary local languages spoken in Lahore.

The next set of exercises taught subjects to combine *"outside"* and *"inside"* views when making a forecast (Kahneman and Lovallo, 1993; Lovallo, Clarke, and Camerer, 2012). The former denotes the base rate at which an event occurs in a reference class (for example, the long-run average level of $PM_{2.5}$ in Lahore). The latter denotes factors particular to a given forecast task (for example, subjects' knowledge that air pollution in Lahore is lower on weekends than on weekdays). The exercise taught subjects about choosing a good reference class and avoiding the tendency to give too much weight to the inside view in forecasting.

In the following set of exercises, we asked subjects to reflect on an earlier forecasting task. Subjects 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). Subjects then completed an exercise that encouraged them not to round their forecasts excessively. Previous work (Mellers et al., 2014) has found that most subjects round too much; that is, their initial rounded forecast does not incorporate all the information at their disposal.

The next 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. For example, if they wanted to forecast air pollution for three days ahead, they were told to consider at least three days of air pollution history.

The final exercise reminded subjects that people tend to allow their emotions and preferences to influence their forecasts. For example, a person who plans to spend the day outside tomorrow may underrate the chance of rain.

F Analysis details

F.1 Treatment on the treated details

The second-stage specification for within-subject analyses is as follows.

$$Y_i = \beta_F \hat{P}_{Fi} + \beta_T \hat{P}_{Ti} + \beta_{FT} \hat{P}_{FTi} + \gamma Y_{0i} + \boldsymbol{X_i'}\boldsymbol{\delta} + \varepsilon_i \tag{8}$$

In this equation i indexes subject. Y is the outcome and Y_0 is the corresponding baseline variable. The variables $\left\{\hat{P}_{Fi}, \hat{P}_{Ti}, \hat{P}_{FTi}\right\}$ represent instrumented takeup. Other controls and hypothesis testing are as in the ITT regressions. The three first-stage specifications for arm $A \in \{T, F, FT\}$ are as follows.

$$P_{Ai} = \varphi_F \mathit{Forecasts}_i + \varphi_T \mathit{Training}_i + \varphi_{FT} \mathit{Forecasts}_i \, \mathit{Training}_i + \nu_A Y_{0i} + \boldsymbol{X_i'}\boldsymbol{\theta_A} + \upsilon_{Ai} \tag{9}$$

Controls are naturally identical in both the first- and second-stage regressions.

F.2 Control variables: machine learning and missing values

As indicated in Section 4.1, we employ post-double selection LASSO to choose a precision-maximizing control set (Ahrens, Hansen, and Schaffer, 2018). This is consistent with the recommendation of Ludwig, Mullainathan, and Spiess (2019).[95] While we used enumerator training and survey design to minimize non-responses to specific questions, subjects were of course given the choice of not responding, or responding "don't know" to any question. We do not consider as potential control variables any questions with high non-response rates, as these may indicate confusion and higher likelihood of measurement error. In addition, to preserve sample size when controls are included, we handle missing values as follows: (i) create a dummy variable for whether the subject did not answer a given question; (ii) replace the control variable with zero instead of missing for non-responses; and iii) include both the control and the dummy in our regression. This is consistent with the recommendation in Gerber and Green (2012). Coefficients on these variables are not interpretable.

[95] According to Wager et al. (2016), ridge regression, LASSO, elastic net, and random forest procedures can all be used to improve efficiency without introducing bias into estimated treatment effects.

F.3 Meaningful deviations from the pre-analysis plan

- We employ asymptotic standard errors in the body of the paper, rather than standard errors from randomization inference, because the latter cannot readily be combined with our pre-specified algorithmic control selection using standard software tools.

- Avoidance index as a primary outcome. In version 1 of our PAP, we listed the avoidance index as a primary outcome. In version 2 we replaced it with outdoor time, failing to realize that effects on outdoor time cannot be analyzed using the same regression framework applied to the other primary outcomes. For example, forecasts may increase outdoor time on clean days and decrease it on polluted days, so the average effect is uninformative. We include the avoidance index as a primary outcome, and analyze the plausibly heterogeneous effects on outdoor time under secondary outcomes. This change has no impact on the number of results that are statistically significant at conventional thresholds (10, 5, or 1 percent).

- Hypothesis tests on willingness to pay for masks. In the PAP we made contradictory statements about alternative hypothesis (right- vs. two-tailed tests) for the avoidance index and willingness to pay for masks, even though both are qualitatively similar avoidance behaviors. We resolve the inconsistency in favor of right-tailed tests on both outcomes because our theoretical model predicts positive treatment effects (Section 2, Hypothesis 2). This change has no impact on the number of results that are statistically significant at conventional thresholds (10, 5, or 1 percent).

- Hypothesis tests on forecast error. In the PAP we specified a left-tailed test for the estimated interaction effect (forecasts and training) on air pollution forecast error. When we later completed our theoretical model, it became obvious that this was an error, as information and human capital can be either substitutes or complements in forecast production. Accordingly we present a two-tailed test of this coefficient estimate in our primary ITT results. This change has no impact on the number of results that are statistically significant at conventional thresholds (10, 5, or 1 percent).

G A Model for Risk Aversion

In this appendix section, we build a simple model with the sole purpose of showing that the effect of changes in risk aversion are ambiguous, in the absence of very strong assumptions (namely CARA). For simplicity of analysis, we abuse notation by reusing variables defined in the main text. In short, consider what follows to be independent of the model in the main text. All variables are re-defined.

First consider the case of air pollution in the absence of any mitigating behavior. The agent is faced with exposure to either high or low pollution. High levels of pollution reduce the agent's utility, and we model this as a reduction in her consumption. We normalize the agent's wealth on a low pollution day to C and model high pollution as damage X. Further assume that the probability of high pollution is $p \in [0, 1]$. The agent is assumed to be risk averse, and we model this by assuming that for consumption x, the agent receives utility $u(x)$, such that $u(0) = 0$, $u' > 0$ and $u'' < 0$. While later we wish to model the effects of risk aversion on the agent's behavior, for simplicity, we suppress any notation for risk preferences, until they are explicitly needed.

In the absence of any mitigating behavior, the agent faces expected utility (baseline)

$$B = pu(C - X) + (1 - p)u(C).$$

Now, assume that the agent may engage in avoidance behavior (e.g purchase a mask, reschedule activities, or remain indoors longer). Avoidance is not free, and comes at a cost (especially in the case of lost work), and we model it as a cost $a \leq X$.[96] By engaging in avoidance behavior pre-emptively (always avoid), the agent can mitigate all costs associated with high pollution; in essence she can guarantee the pay-off

$$A = u(C - a).$$

It is possible that the agent chooses not to avoid at all times. If the agent could predict high pollution, she could attempt to only avoid in such cases, and thereby save on the cost of unnecessary avoidance. In the absence of any forecast, we assume that the agent's naive belief that a given day is high pollution is equal to p as well.[97] Then every day, with a probability of p, she either chooses to avoid (in response to what she believes is a high pollution day), or not avoid with probability $(1-p)$. This probabilistic response to pollution would yield expected pay-off

$$N = p(pu(C - a) + (1 - p)u(C - X)) + (1 - p)(pu(C - a) + (1 - p)u(C)). \tag{10}$$

In the equation above, note that first nature decides whether a day is high or low pollution and then for each day, the agent naively predicts whether it is high or low pollution. Equation

[96] If $a > X$, the agent will never engage in avoidance, and the case is not of interest.

[97] If nature decides with probability p that there is high pollution, and the agent calculates her expected pay-off using the same, it is intuitive that the agent uses the same unconditional prediction.

10 can be rearranged and expressed as the weighted average of A and B,

$$\begin{aligned} N =& pu(C-a)(p+(1-p)) + (1-p)(pu(C-X) + (1-p)u(C)) \\ =& pA + (1-p)B. \end{aligned}$$

N is a convex combination of A and B, implying that the agent would never engage in naive predictions; she would either always avoid if $A \geq B$ or never avoid. We therefore only need to consider these two cases, and so introduce our forecast service case by case.

G.1 Introducing forecasting

Now assume there is a service available, which informs the agent whether she will face high or low pollution. This allows the agent to decide whether to avoid or not contingent on the additional information in the forecast. The price of the forecast is f and we are interested in finding the range of prices for which agents would purchase the service. The forecast is imperfect, that is with a probability π, it may incorrectly predict the level of pollution.[98] The forecast allows our agent to only avoid when the forecast predicts there is high pollution.

Then for an agent who purchases the forecast service (and follows it), her expected utility is given by

$$p[\pi u(C-a-f) + (1-\pi)u(C-X-f)] + (1-p)[\pi u(C-f) + (1-\pi)u(C-a-f)].$$

Note the implicit timing in the formulation. As with naive avoidance, nature first decides whether there is high or low pollution. Then based on the realized level of pollution, the forecast predicts the state of the world correctly or incorrectly, with probability π and $(1-\pi)$ respectively. If the agent buys the forecast, all consumption levels are reduced by the forecast price f. The forecast is introduced to two different types of agents; those who in the absence of a forecast were avoiding and those who were not. We consider these two cases separately.

G.2 Forecasting when avoidance is not too costly

In the case where avoidance is not too costly, the agent would purchase the forecast if

$$p[\pi u(C-a-f)+(1-\pi)u(C-X-f)]+(1-p)[\pi u(C-f)+(1-\pi)u(C-a-f)] \geq u(C-a). \quad (11)$$

[98] While these probabilities may be contingent on the realized level of pollution, for simplicity we assume that forecast reliability is constant.

We wish to model how behavior would change with changes in risk aversion. To do this, we focus on the threshold price of the forecast, f^a, such that for all $f \leq f^a$, an agent would purchase the forecast, and for those above they would continue to avoid at all times.

Remark. A threshold price f^a exists.

Proof. Note that we can re-write (11) as $T^a(f) = p[\pi u(C-a-f)+(1-\pi)u(C-X-f)]+(1-p)[\pi u(C-f)+(1-\pi)u(C-a-f)]-u(C-a)$. T^a is continuous as u is continuous. Further more, it is obvious that it is strictly decreasing in f (as u is strictly increasing). Finally, note that $T^a(X) \leq 0$ and if $T^a(0) \geq 0$ then by the mean value theorem, a $f^a \in [0, X]$, otherwise $f^a = 0$. Finally, f^a is unique as T^a is strictly decreasing. □

We are interested in how a non-trivial f^a, which can be interpreted as the highest willingness to pay for a forecast service, behaves as we change the agent's risk aversion. Intuitively, it should decrease as risk aversion increases, because the forecast service in essence offers a lottery, while always avoiding is a certain outcome. The intuition holds, and to see why we express the threshold function as

$$T^a(f) = u(\varphi) - u(c-a),$$

where φ is the certainty equivalent of a lottery with pay-offs of $(C-a-f, C-X-f, C-f)$ with respective probabilities $(p\pi + (1-p)(1-\pi), p(1-\pi), (1-p)\pi)$. Then by definition as risk aversion increases, φ decreases, shifting T^a downwards and decreasing the threshold price f^a.

Result 1. *For agents who, in the absence of a forecast would engage in avoidance, willingness to pay for a forecast is decreasing in their level of risk aversion .*

We can also derive other comparative static results using the geometric properties of T^a.

Result 2. *The threshold value f^a is:*

1. *Decreasing in p.*

2. *Increasing in π.*

Proof. As f^a is the fixed point of T^a, shifts in T^a would also shift its fixed point. As such we consider the partial derivatives of T^a with respect to each exogenous variable. $\frac{\partial T^a}{\partial p} = \pi[u(C-a-f) - u(C-f)] + (1-\pi)[u(C-X-f) - u(c-a-f)] \leq 0$ as $u' > 0$ and $0 \leq a \leq X$.

Similarly, $\frac{\partial T^a}{\partial \pi} = p[u(C-a-f) - u(C-X-f)] + (1-p)[u(C-f) - u(C-a-f)] \geq 0$. □

Both results are intuitive. As the probability of a high pollution event increases, the expected benefit of a sophisticated response gained through the forecast falls. Similarly, as the reliability of a forecast increases, so does demand for it.

G.3 Forecasting when avoidance is too costly

When avoidance is in itself too costly, a forecast product presents the agent with a choice between two lotteries: forecast-based avoidance and no avoidance. The agent would purchase the forecast if

$$\begin{aligned} T^n(f) =& p[\pi u(C-a-f) + (1-\pi)u(C-X-f)] + (1-p)[\pi u(C-f) + (1-\pi)u(C-a-f)] \\ & - pu(C-X) - (1-p)u(C) \geq 0. \end{aligned} \tag{12}$$

Once again, analogous to the previous case, the model yields a threshold price that is unique.

Remark. A threshold price f^n exists when avoidance is costly.

Before conducting comparative statics, let us consider the threshold at which the agent would consume a forecast even when it is given away for free. We set $f = 0$ and consider our agent's choice. She chooses to use the forecast service if

$$\begin{aligned} p[\pi u(C-a) + (1-\pi)u(C-X)] & \\ +(1-p)[\pi u(C) + (1-\pi)u(C-a)] &\geq pu(C-X) + (1-p)u(C), \\ p[\pi u(C-a) + (1-\pi)u(C-X) - u(C-X)] &\geq (1-p)[u(C) - \pi u(C) - (1-\pi)u(C-a)], \\ p\pi[u(C-a) - u(C-X)] &\geq (1-p)(1-\pi)[u(C) - u(C-a)]. \end{aligned} \tag{13}$$

This formulation provides intuition behind the agent's choice. The left-hand side in equation (13) captures the benefit of the forecast; it is the expected utility of avoiding when the forecast correctly predicts high pollution. Meanwhile the right hand side of the same equation reflects the expected cost of an incorrect forecast leading to unnecessary avoidance. The agent would only use a free forecast if the benefit is greater than costs. In essence, this shows that for a forecast to matter, its skill must exceed some lower bound.

We now move to comparative static analysis for our non-trivial case, i.e cases where equation 13 is satisfied, and the agent would have a non-zero threshold price. Basic comparative statics with respect to p and π can be derived as before, however analyzing changes with respect to the agent's risk preferences requires more assumptions. We therefore first establish the results with respect to the former, and then move to analyze risk separately.

Result 3. *The threshold value f^t is:*

1. *Decreasing in p.*
2. *Increasing in π.*

Proof. Analogous to f^a, f^n is the fixed point of T^n and shifts in T^n would also shift its fixed point. As such we consider the partial derivatives of T^n with respect to each exogenous variable. $\frac{\partial T^n}{\partial p} = \pi[u(C-a-f) - u(C-f)] + (1-\pi)(u(C-X-f) - u(C-a-f) \leq 0$ as $u' > 0$ and $0 \leq a \leq X$.

Similarly, $\frac{\partial T^n}{\partial \pi} = p[u(C-a-f) - u(C-X-f)] + (1-p)[u(C-f) - U(C-a-f)] \geq 0$. □

G.4 Risk aversion and willingness to pay.

To study the relationship between risk aversion and f^n, for tractability we need more structure. We assume that the forecast is perfectly reliable, i.e π =1 and further assume that the agent's utility exhibits constant absolute risk aversion (CARA). In particular we use the standard CARA formulation, and assume that when the agent consumes x units, her utility takes the form $u(x) = 1 - e^{-\alpha x}$, where α is her Arrow-Pratt coefficient of absolute risk aversion.

When the forecast is perfectly reliable, the agents choice simplifies to

$$pu(C-a-f) + (1-p)u(C-f) \geq pu(C-X) + (1-p)u(C). \tag{14}$$

The agent is comparing two simple lotteries, with the same binary probabilities over different outcomes. We therefore define the certainty equivalent for such binary lotteries. For CARA, the certainty equivalent is not a function of initial wealth, so we define the certainty equivalent based on spread. Let $ce(x, \alpha)$, be the certainty equivalent of a lottery that yields 0 with probability p and x with probability $(1-p)$, for an agent with an Arrow-Pratt coefficient of absolute risk aversion, α.

Then we can re-write equation (14), which defines the threshold value as

$$u(C - f - a + ce(a, \alpha)) \geq u(C - X + ce(X, \alpha)).$$

As u is strictly increasing in consumption, we can rewrite the above as $C - f - a + ce(a, \alpha) \geq C - X + ce(X, \alpha)$. So our threshold is equivalently defined by

$$f^t = (X - a) + ce(a, \alpha) - ce(X, \alpha).$$

Differentiating with respect to α yields

$$\frac{\partial f^t}{\partial \alpha} = ce_\alpha(a, \alpha) - ce_\alpha(X, \alpha).$$

To sign this we need to know the rate at which the slope of the certainty equivalent w.r.t. risk aversion changes w.r.t. the size of the lottery, i.e $ce_{\alpha x}$. So, let us focus on $ce(x, \alpha)$. Allowing for minor abuse of notation, we add risk aversion as a determinant of utility and express utility as $u(x, \alpha)$.

$$\begin{aligned} u(ce(x, \alpha), \alpha) &= pu(0, \alpha) + (1-p)u(x, \alpha), \\ &= (1-p)u(x, \alpha). \end{aligned} \tag{15}$$

We now differentiate both sides with respect to α, which yields

$$\begin{aligned} u_x(ce, \alpha)ce_\alpha(x, \alpha) + u_\alpha(ce, \alpha) &= (1-p)u_\alpha(x, \alpha), \\ u_\alpha(ce, A)ce_\alpha(x, \alpha) &= (1-p)u_\alpha(x, \alpha) - u_\alpha(ce, \alpha), \\ u_x(ce, \alpha)ce_\alpha(x, \alpha) &= [u_\alpha(x, \alpha) - u_\alpha(ce, \alpha)] - pu_\alpha(x, \alpha). \end{aligned}$$

Given our functional form for u, we know that $u_\alpha = \alpha e^{-\alpha x} \geq 0$, $u_\alpha = \alpha x e^{-\alpha x} \geq 0$, $u_{x\alpha} = -\alpha x e^{-\alpha x} \leq 0$ and $u_{xx} = -\alpha^2 e^{-\alpha x} \leq 0$. This coupled with the fact that $ce(x, \alpha) \leq x$ by construction, implies that the term in the square bracket is negative, and so $ce_\alpha \leq 0$ (as expected).

We are interested in $ce_{\alpha x} = ce_{x\alpha}$. To solve this, first differentiate (15) by x and then by α.

$$\begin{aligned} u(ce(x, \alpha), \alpha) =& (1-p)u(x, \alpha), \\ u_x(ce, \alpha)ce_x(x, \alpha) =& (1-p)u_x(x, \alpha), \\ ce_x(x, A) =& (1-p)\frac{u_x(x, \alpha)}{u_x(ce, \alpha)}, \\ ce_{xA} =& (1-p)\frac{u_x(ce, \alpha)u_{x\alpha}(x, \alpha) - u_x(x, \alpha)[u_{xx}(ce, \alpha)ce_\alpha(x, \alpha) + u_{x\alpha}(ce, \alpha)]}{u_x(ce, \alpha)^2}, \\ &\underset{sign}{\sim} u_x(ce, \alpha)u_{x\alpha}(x, \alpha) - u_x(x, \alpha)u_{xx}(ce, \alpha)ce_\alpha(x, \alpha) - u_x(x, \alpha)u_{x\alpha}(ce, \alpha), \\ &\underset{sign}{\sim} - u_x(x, \alpha)u_{xx}(ce, \alpha)ce_\alpha(x, \alpha) + [u_x(ce, \alpha)u_{x\alpha}(x, \alpha) - u_x(x, \alpha)u_{x\alpha}(ce, \alpha)]. \end{aligned}$$

The first term is negative given what we already know. Focusing on the term in the square bracket we have

$$\begin{aligned} u_x(ce,\alpha)u_{x\alpha}(x,\alpha) - u_x(x,\alpha)u_{x\alpha}(ce,\alpha) =& \alpha e^{-\alpha ce}(-\alpha x e^{-\alpha x}) - (\alpha e^{-\alpha x})(-\alpha(ce)e^{-\alpha ce}), \\ =& \alpha^2 e^{-\alpha(x+ce)}(ce - x) \leq 0. \end{aligned}$$

Therefore, $ce_{x\alpha} \leq 0$.

All this allows us to sign $\frac{\partial f^t}{\partial \alpha} = ce_\alpha(a,\alpha) - ce_\alpha(X,\alpha)$. As $X \geq a$ and $ce_{\alpha x} \leq 0$, we have that $\frac{\partial f^t}{\partial \alpha} \geq 0$.

Result 4. *When avoidance is costly, more risk averse agents are willing to pay higher prices for the (perfectly reliable) forecast service.*