Forecasts: Consumption, Production, and Behavioral Responses*

Husnain F. Ahmad, Matthew Gibson, Fatiq Nadeem,

Sanval Nasim, Arman Rezaee

November 2021

Abstract

Economic theory predicts forecasts are an important determinant of welfare. In developing countries, however, limited information and human capital may make it difficult for agents to produce accurate, precise forecasts. This plausibly limits the scope for optimal responses to uncertain future events. We study the effects on forecast consumption, production, and behavioral responses from two randomized interventions in Lahore, Pakistan: 1) provision of one-day ahead air pollution forecasts; and 2) general forecasting training aimed at reducing behavioral biases. On average, subjects exposed to forecasts were willing to pay roughly 60 percent of the cost of mobile internet access to continue receiving them. Both interventions reduced air pollution forecast error, and receipt of forecasts increased demand for protective masks. These results document substantial demand for forecasts among urban residents in the developing world. They suggest that modest educational interventions may durably improve forecasting-relevant human capital.

*Ahmad: Department of Economics, Sewanee: The University of the South. Gibson: Department of Economics, Williams College, and IZA. Nadeem: Bren School of Environmental Science & Management, University of California, Santa Barbara. Nasim: Department of Economics, Lahore University of Management Sciences (LUMS). Rezaee: Department of Economics, University of California, Davis, abrezaee@ucdavis.edu.
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); LUMS; Abdul Latif Jameel Poverty Action Lab (JPAL); Williams College and the University of the South. We would like to thank Airvisual, 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-preanalysis plan is registered with the AEA (RCTR-0004343).

1 Introduction

Economic theory predicts that forecasts are an important determinant of welfare. For example, an agent who relies on biased forecasts may not optimally smooth her consumption over time. An agent who does not forecast high air pollution may fail to undertake avoidance behavior that she would have chosen in the presence of an accurate forecast. Such phenomena may be especially common in developing countries, where third-party forecasts may be unavailable or of poor quality for a given agent's location (Rosenzweig and Udry, 2014a,b). Forecasting mistakes are common even among experts (Tetlock, 2017). In developing countries, underlying behavioral biases may interact with information scarcity and lower levels of human capital (North, 2003, Stiglitz, 2000, Hanushek, 2013). Forecasting mistakes may be particularly consequential for agents who face considerable risk in domains from health (Blakely et al., 2005) to employment and income (Munshi and Rosenzweig, 2016).

This paper studies how urban residents in the developing world solve forecasting problems in the presence of constraints on information and human capital. It is concerned with the following broad questions. Is there positive demand for forecast products among residents of developing cities? How does consumption of forecasts influence behavior, especially avoidance of environmental harm? Can urban residents form useful forecasts? Finally, can their forecasting ability be improved? The answers to these questions shed light on the behavior of developing-country urbanites and may be a useful input to government policy.

To address these research questions, we implemented a randomized controlled trial in which 999 subjects acted as both consumers and producers of forecasts. The research design includes two orthogonal treatments: first, provision of day-ahead air pollution forecasts delivered by text message; and second, general in-person training designed to reduce behavioral biases in forecasts. Broadly, we are interested in three types of outcomes: 1) consumption, e.g. demand for our air pollution forecast product; 2) production, e.g. error in forecasting one's own future labor supply; and 3) behavioral responses, e.g. demand for particulate-filtering face masks. In theory, these two treatments could be gross complements or gross substitutes. While our research design allows us to measure this interaction, it is of secondary interest.

Our experiment took place in Lahore, Pakistan. In 2019 Lahore was the 12th most polluted city in the world, with air roughly comparable to that of Delhi and Dhaka (IQAir, 2019, Riaz and Hamid, 2018, Zahra-Malik, 2017). While the pollution problem is acute, residents of Lahore seeking to make accurate forecasts (i.e., to form unbiased expectations) face a challenging information environment. There is some provision of retrospective information, but such efforts remain spotty. Measurements posted online by the Environmental Protection Department (EPD) of the Punjab provincial government are incomplete in space and time.[1] The US consulate in Lahore recently began providing information online, but this represents one point in a city with an area of more than 680 square miles. Air pollution forecasts are not readily available, particularly to the majority of residents who do not speak English.

Average levels of human capital in Lahore may also contribute to the difficulty of making accurate forecasts. Citywide, the average level of educational attainment is between 6.2 and 6.5 years (NIPS and ICF, 2019). In our subject population, it is 9.3 years. The nationwide level of educational attainment in Pakistan (4.8 years) is a year lower than in India, and roughly comparable to Uganda, Ethiopia, and Nigeria (World Bank, 2017). Urban residents in these countries may face skill constraints similar to those of our subjects. Moreover residents of Lahore may confront the same behavioral biases that generate forecasting errors even in highly educated populations (Kahneman and Tversky, 1973).

Using incentive-compatible elicitations, we find subjects exposed to our one-day ahead air pollution forecasts were willing to pay an average of 93.22 Pakistani Rupees (PKR) to continue receiving forecasts for 90 days.[2] On a monthly basis, this is roughly 60 percent of the cost of 4G mobile internet access. It stands in some 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 by roughly one-tenth of a standard deviation, or 5 $\mu g/m^3$.[3] This is approximately 20 percent of the World Health Organization's corresponding maximum safe 24-hour standard.[4] Forecast provision increased willingness to pay for particulate-filtering masks by 6.8 PKR, roughly five percent of the retail price.[5] While the estimated effect of training is positive (3.9 PKR), it is

[1] According to the government of Punjab: "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, 2017).

[2] Willingness to pay for both forecasts and masks was elicited using a Becker-DeGroot-Marschak mechanism, as described in Section 3.3.2.

[3] Forecasts were incentivized using payments for being within 5, 10, or 20 percent of realized particulate pollution.

[4] Here "fine particulates" denotes $PM_{2.5}$ air pollution: the concentration of particulates of diameter 2.5 microns or less, measured in micrograms per cubic meter ($\mu g/m^3$). The World Health Organization $PM_{2.5}$ standards are 25 $\mu g/m^3$ for 24 hours, and 10 $\mu g/m^3$ annually (Organization et al., 2006).

[5] N95 masks filter 95 percent of small particles. According to the mask manufacturer 3M, the retail price of a genuine N95 mask was 130-140 PKR in November 2019, while our experiment was in progress. Endline surveys were completed prior to the outbreak of Covid-19 in Pakistan, which would have significantly

not statistically significant. We find no treatment effects on an index of avoidance behavior or a proxy for the variance of utility.

In secondary empirical results, we find imprecise evidence that training reduces subjects' errors in forecasting their own labor supply. Forecast confidence was unaffected by either treatment. Secondary outcomes also illuminate the mechanisms behind treatment-driven reductions in air pollution forecast errors. Subjects in the training arm increased takeup of free weather forecasts prior to making an incentivized air pollution forecast. This is consistent with training increasing demand for inputs to forecast production. Subjects in the SMS forecast arm reported increased consumption of air quality and weather information, but did not increase takeup of weather forecasts just prior to making incentivized air pollution forecasts. These subjects had larger information sets, but there is no evidence they adjusted their forecasting processes.

Our project contributes to several literatures, of which the first is forecasting in developing countries. Previous work has focused on farmers and responses to weather forecasts, especially precipitation forecasts (Rosenzweig and Udry, 2014a,b, Kala, 2017). While we also study responses to forecasts, this paper makes novel contributions to this literature on several dimensions. We elicit beliefs directly using incentive-compatible mechanisms, rather than inferring them within a structural model. Our urban experimental subjects face uncertainty in different oucomes, including air pollution, travel times, labor supply, expenditure, and happiness. Studying such outcomes is increasingly important, as rural citizens in the developing world continue moving to cities (Henderson, 2002). These outcomes also enable us to ask somewhat different research questions. For example, the commonly studied response to precipitation forecasts is the timing of planting. Our setting allows for study of frequently repeated choices, such as time spent outside, where learning is plausibly easier.

Second, our results add to the literature on training interventions in developing countries. Previous work has focused on business and entrepreneurship skills (Karlan and Valdivia, 2011, McKenzie and Woodruff, 2014, Valdivia, 2015), or job training (Card et al., 2011, Acevedo et al., 2017). Our paper instead considers training in forecast skills, with a focus on reducing behavioral biases. Scholars have sought to reduce such biases in wealthy country settings (Mellers et al., 2014, Morewedge et al., 2015, Soll, Milkman, and Payne, 2015). To the best of our knowledge, ours is the first study to adapt such techniques to the constraints of a developing city.

confounded our elicitations had it occurred earlier.

The third relevant literature is on pollution avoidance behavior.[6] There is a substantial empirical literature on avoidance behavior in developed countries. Prominent examples include Neidell (2004) , Graff Zivin and Neidell (2009) , and Moretti and Neidell (2011) .[7] We provide evidence from the developing world, where preferences may differ, incomes are lower, and the scope for avoidance may differ (e.g. because of available technologies or intra-day patterns in air pollution). Previous work on avoidance largely relies on natural experiments for identification, which limits the questions it can ask. For example, it is common to observe an avoidance behavior, such as a canceled trip to an outdoor zoo, but agents' air pollution expectations are unobserved. So far as we are aware, ours is the first experimental study in this literature. Our finding that average willingness to pay for masks was approximately 77 percent of the retail price offers a potential explanation for low takeup in some developing cities with high air pollution.[8]

Our project speaks to the literature on economic effects of air pollution in developing countries. Previous work has examined mortality and several dimensions of health (Edwards and Langpap, 2012, Chen et al., 2013, Ebenstein et al., 2015, Arceo, Hanna, and Oliva, 2016).[9] A recent body of work has estimated effects on labor productivity (Adhvaryu, Kala, and Nyshadham, 2019, He, Liu, and Salvo, 2019, Chang et al., 2019) and labor supply (Hanna and Oliva, 2015). We broaden this literature to include the entire 24-hour time allocation for both adults and children, including home production and leisure.

Lastly we contribute to the body of work on demand for environmental information. Barwick et al. (2019) estimate effects of air pollution information in China on a variety of outcomes, including shopping and housing markets, and use these responses to bound the value of air pollution information. Our study differs in eliciting the value of air pollution information directly, using an incentive-compatible mechanism. The most closely related study is Barnwal et al. (2017) , which randomized prices for arsenic testing in Bihar, India.

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 outlines data collection. Section 5 evaluates balance and presents estimating equations. Section 6 discusses estimates for both primary and secondary outcomes, and Section 7 concludes.

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

[7] A thorough review, including a brief theoretical foundation, is in (Graff Zivin and Neidell, 2013).

[8] Mean endline willingness to pay for an N95 mask in our control group was 104 PKR. Median WTP was 100 PKR. According to 3M Pakistan, the average retail price of such a mask prior to the covid-19 pandemic was approximately 135PKR.

[9] Other important work in this area includes: Alberini et al. (1997) , Cropper et al. (1997) ; and Jeuland, Pattanayak, and Bluffstone (2015) .

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 ($t = 1$).[10] 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 or rescheduling of planned outdoor activities.[11] 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[12] and d^s is the state-dependent damage function,[13] assumed to be decreasing and strictly convex in the sum of avoidance purchased.[14] 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 $d^h_1(A) \leq d^l_1(A)\ \forall A$.[15] 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.[16] Mathematically this implies that globally, $c_1 \leq c_2$.

[10]This day-ahead model allows us to simplify our analysis and ignore inter-temporal considerations.

[11]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.

[12]We assume the endowment is large enough to avoid any credit constraints.

[13]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 the appendix.

[14]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. 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.

[15]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.

[16]Very far ahead of time, all other time uses in an individual's allocation are potentially substitutes for outdoor time. Just beforehand, however, substitution may be constrained, or so expensive an individual

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, $c_{11} \leq c_{12} \leq c_{22}$.[17]

The level of pollution is unknown at time 0 but revealed at time 1. The probability of high pollution is $P(h) = \pi$.[18] 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, $P(H|h, \iota, \tau) = P(L|l, \iota, \tau) = \rho$, and assume she is equally good at predicting high and low pollution.[19] 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 $\rho \geq \max\{\pi, 1 - \pi\}$.

2.1 Avoidance purchased after pollution is realized (2nd period)

We begin solving backward and consider the problem at time $t = 1$ (the second 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

$$u^s(x) = \max_y \{E - d^s(x + y) - c(x, y)\}. \tag{1}$$

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

$$-d_1^s(x + y^s) - c_2(y^s) = 0. \tag{2}$$

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

---

would never undertake it.

[17]As costs are convex in each period's avoidance, making this assumption ensures that at any (x, y), buying more x would increase the marginal cost of x by less than the marginal cost of more y ($c_{11} \leq c_{21}$). Similarly buying more y would raise marginal cost of y by more than that of x ($c_{22} \geq c_{12}$). This is perhaps easier to see if we assume $c(x + \beta y)$, where $\beta > 1$. 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 requirements are $c(0,0) = \alpha$, $c_1(0,0) = \beta$ and $c_2(0,0) = \gamma$, and $\gamma \geq \beta$.

[18]The parameter π can also be interpreted as the agent's unbiased prior before she begins optimizing.

[19]Note the abuse of notation. ρ is a function of (ι, τ), and not a constant. We drop the arguments for notational simplicity.

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 2nd 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.

2.2 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

$$V(\rho, \pi) = \max_{x^H, x^L} \{\pi[\rho u^h(x^H) + (1-\rho)u^h(x^L)] + (1-\pi)[\rho u^l(x^L) + (1-\rho)u^l(x^H)]\}. \qquad (3)$$

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 realised, 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 period 0 that is a function of the agent's forecast.[20] 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

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

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

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

$$\max_x \left\{ \begin{matrix} q^H[E - d^h(x + y^h(x)) - c(x, y^h(x))] \\ +(1 - q^H)[E - d^l(x + y^l(x)) - c(x, y^l(x))] \end{matrix} \right\}. \tag{4}$$

Before continuing, we note that the best-case scenario for the agent is a 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 $x \geq 0$. Then we can re-write equation 4 as

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

The first-order condition yields[21]

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

Rearranging and substituting in the first-order condition for the period 1 problem (equation 2) yields

$$q^H[-d_1^h] + (1 - q^H)[-d_1^l] - c_1 = 0. \tag{5}$$

We do not need to check the second-order condition as this is a simple case of partial minimisation. 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 5 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

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

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.

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.[22] 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.[23]

2.3 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 3. 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

[22] In particular, if $\pi \geq 0.5$, $q^H \geq q^L$, while the reverse is true otherwise.
[23] This is driven by the fact that $\rho \geq \max\{\pi, 1-\pi\}$, which implies $q^L, q^F \geq 0.5$.

$q^H[-d_1^h]+(1-q^H)[-d_1^l]=c_1$ 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, $u_1^l \leq 0$.[24] 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 ($x^H(q^H) \leq \text{argmax}_x\, u^h(x)$) and over-invests when it is low ($x^L(q^L) \geq \text{argmax}_x\, u^l(x)$).[25] The result is intuitive, as when the agent forecasts high pollution, she under-invests to benefit 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.

2.4 Effects of information and training

As a final step, we now model the effects of our SMS forecast and training services. 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.[26] 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.

2.5 Hypotheses

The assumption that our treatments increase an agent's forecast skill yields the following hypothesis.

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

---

[24]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.

[25]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.

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

We now turn to avoidance behavior. Lahore experienced high air pollution throughout our study. If subjects forecasted 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, collected through phone surveys and at endline, 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 air pollution forecast.

Hypothesis 3. *Avoidance (e.g. rescheduling of outdoor activities, reduced outdoor time) is expected to be higher among recipients of the SMS service on days where the the level of pollution forecast was higher.*

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 little structure on the agent's forecast function ρ.[27]

Hypothesis 5. *Among participants who received the SMS service, 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

In this section we describe our interventions, sampling, and randomization. Figure 1 shows the division of our sample into treatment and control groups. We delivered SMS air pollution forecast messages to respondents in Treatment Groups 1 and 3 (*T1* and *T3*) every day over a period of eight months. We implemented the forecast training once for every household in Treatment Groups 2 and 3 (*T2* and *T3*) over a period of three months.

[27] 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 beyond the scope of this paper.

3.1 Interventions

3.1.1 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.[28]

1. A model based on data from our own air pollution monitors. $PM_{2.5}$ levels for *t+1* were predicted using a 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 third-party forecasts based on satellite data.

For additional detail on how these models were estimated and disseminated, see Appendix C. We sent our treatment group (*T1*) respondents two pieces of information: 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 forecasts.

3.1.2 Forecast Training

We implemented a one-hour forecast training based on the principles of Tetlock (2017) and Kahneman (2011) . Broadly speaking, the training aimed to reduce behavioral and psychological mistakes that decrease the precision and accuracy of subjects forecasts. A group of specially selected and trained enumerators conducted the trainings in Urdu in subjects' homes.[29] Subjects received 150 PKR for their participation.

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 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.

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). Subjects

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

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

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. 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.

3.2 Study site and sampling frame

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.[30] Table A3 reports results. 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.[31]

3.3 Assignment to treatment

3.3.1 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.

[30] The MICS data does not distinguish Model Town (our other selected Tehsil) from other Tehsils. This prevents us from including data from Model Town in Table A3

[31] 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.

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,[32] education, gender, age, and a dummy for having provided a subsequently verified phone number at baseline.

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).

3.3.2 Secondary treatments

Our experimental design included a second treatment orthogonal to our primary treatments. At both baseline and endline, we elicited willingness to pay for a particulate-filtering mask via a Becker-DeGroot-Marschak mechanism (Becker, DeGroot, and Marschak 1964), with the price in Pakistani Rupees (PKR) drawn from a uniform distribution (baseline $[0, 150]$, endline $[0, 200]$). Conditional on a subject's bid b, receipt of the mask at baseline was random. The probability of receiving the mask at baseline was $b/150$ for $b \in [0, 150]$ and 1 for $b > 150$. This provides an additional source of exogenous variation. Subjects with $b \leq 150$ can be matched on b and their outcomes can be compared. This is not arbitrary (and was prespecified). If masks are experience goods, preferences for them may change over time among individuals who use them. This may, in turn, influence willingness to pay, forecast ability, and other outcomes.

[32] 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.

4 Data

Below we describe original data collection and third-party data sets employed in the experiment. For details, see Appendix D.

4.1 Sampling and subjects

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

Following the sampling frame described in Section 3.2, we used 7 Charges (sub-subdistricts). 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. For each household, respondent gender was chosen using a pre-generated random list.

4.2 Survey frequency

There were three primary data-collection efforts. First, in-person baseline and endline surveys of all respondents were conducted. Second, following a one week transitionary period from the start of our SMS intervention, all respondents were surveyed weekly by phone. The weekly phone data collection lasted approximately 12 weeks. Third, for individuals who received the forecasting training (groups T2 and T3 in Figure 1), data were recorded during the training session.

4.3 Third-party data

- AQMesh and Dusttrak II: We used two industrial-grade air pollution monitors: (1) the AQMesh; and (2) the Dusttrak II.
- AirNow International: The U.S. EPAs AirNow International is a global version of the agency's U.S.-based air quality data management and display system. It provides hourly data on $PM_{2.5}$ levels.

- meteoblue: meteoblue uses forecasts of global atmospheric conditions from the Copernicus Atmosphere Monitoring Service (CAMS) to generate air quality forecasts. Forecasts are issued at a 40 km resolution and combined from a set of different global datasets from different sources, variables, spatial and temporal resolution, reaching back to 1984. This data is combined using an ensemble model consisting of calculating the median value of the individual outputs.

- SPRINTARS: The Spectral Radiation-Transport Model for Aerosol Species uses aerosols from both natural and anthropogenic sources to estimate categories for SPM, PM_{10} and $PM_{2.5}$.

- AccuWeather: AccuWeather takes the U.S. National Oceanic and Atmospheric Administration's (NOAA) weather forecasts and transforms them for general consumers.

- Google Maps: Travel time data were collected from Google Maps from one month before the baseline survey through the endline survey.

5 Empirical strategy

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

5.1 Treatment effects

The estimation strategy for primary outcomes is given explicitly below. The strategy for some secondary outcomes is given explicitly, but for others we proceed by analogy with the primary outcomes.

5.1.1 Intent to treat

We estimate willingness to pay for three months of SMS forecasts between subjects.

$$Y_i = \alpha + \boldsymbol{Z_i'}\boldsymbol{\beta} + \varepsilon_i \tag{6}$$

In this equation i indexes subject and Y is the outcome. $\boldsymbol{Z}$ is the vector of dummies denoting random assignment to SMS forecasts (Z_F) and training (Z_T), plus their interaction ($Z_{FT} = Z_F * Z_T$).

Heteroskedasticity-robust asymptotic standard errors appear in the body of the paper, with standard errors from randomization inference in the appendix. For this outcome, our pre-specified hypothesis test is one-tailed: $\alpha + \beta_F > 0$. That is, we test whether mean willingness to pay is positive among subjects in the forecast-only group.[33] This is the test that will be included in our multiple testing correction procedure. 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.

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 the avoidance index. The estimating equation is as follows.

$$Y_i = \boldsymbol{Z_i'}\boldsymbol{\beta} + \gamma Y_{0i} + \boldsymbol{X_i'}\boldsymbol{\delta} + \varepsilon_i \tag{7}$$

Notation for outcomes and treatments is as in Equation 6.[34] Y_0 is the baseline variable corresponding to the outcome Y_i. $\boldsymbol{X}$ is a vector of controls, including include block dummies. As pre-specified, other elements of $\boldsymbol{X}$ were chosen using post-double-selection LASSO applied separately to each primary outcome.[35]

Again as pre-specified, hypothesis testing on $\hat{\boldsymbol{\beta}}$ varies by outcome. For the air pollution forecast error index, theory predicts that more information and better forecast training should both weakly improve forecast quality. The tests are one-tailed, against the alternatives $\beta_F < 0$, $\beta_T < 0$, and $\beta_{FT} < 0$. Theory predicts that both treatments should improve subjects' ability to smooth utility over time, so tests in the model of self-reported happiness variance are one-tailed ($\beta_F < 0$, $\beta_T < 0$, and $\beta_{FT} < 0$). Finally we expect both treatments to increase avoidance, but have no strong prior on their interaction, so tests for mask demand and the avoidance index are against the following alternatives: $\beta_F < 0$, $\beta_T < 0$, and $\beta_{FT} \neq 0$. Theory predicts that the sign of the ITT effects may vary, both within and across subjects, because of differences in expectations. We discuss this important heterogeneity in Section 5.2 below.

For the time-use questions in our telephone surveys, dynamic effects may be present and there are multiple ways to estimate them. We report several, either in the primary tables

[33] Note that because block dummies are not included in Equation 6, 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] All treatment regressions include a constant term, but we omit it from most equations in this document in the interest of clarity.

[35] See Section 5.3 for more discussion.

or the appendix. One possibility is to estimate a separate effect for each date on which a telephone survey is conducted. The estimating equation is as follows.

$$Y_{id} = \boldsymbol{\alpha_i} + \boldsymbol{\delta_d} + \sum_{a=F,T,FT} Z_{ai}\boldsymbol{\beta_{ad}} + \boldsymbol{X'_{id}}\gamma + \varepsilon_{id} \tag{8}$$

The vectors $\boldsymbol{\beta_{ad}}$ are the estimands of interest; each one contains a separate parameter for each date (d). This is the most flexible way to model dynamic effects, but may suffer from poor statistical power. We also estimate regressions pooling these outcomes at the week and month level. Finally we estimate a regression in which ITT effects are assumed to evolve linearly over dates.

5.1.2 Treatment on the treated

For the training arm ($Z_T = 1$) we observe participation in the training session ($P_T = 1$). For the forecast arm ($Z_F = 1$) takeup meant looking at our SMS forecast. This was not directly observable. Moreover it plausibly varied, both across individuals and within individual over time. We construct a takeup measure using telephone survey responses to the question: "How many times in the last week have you seen our pollution forecast message?"[36] Denote the response of subject i on date d as R_{id}.[37] Then a subject's takeup for the week w preceeding date d is defined as $P_{Fiw} = \frac{1}{7}R_{id}$. The subject's aggregate takeup is $P_{Fi} = \frac{1}{W_i}\sum_{w=1}^{W_i} P_{Fiw}$, where W_i is the total number of telephone survey responses for individual 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 $\boldsymbol{Z_i}$. For subjects in the arm assigned to both treatments $P_{FT} = P_F P_T$. Let the vector $\boldsymbol{P}$ contain all three takeup variables.

Effects of treatment on the treated are estimated using 2SLS, with $\boldsymbol{Z}$ instrumenting for $\boldsymbol{P}$. In particular, the second-stage specification for cross-subject analyses is as follows.

$$Y_i = \alpha + \widehat{\boldsymbol{P'_i}}\boldsymbol{\beta} + \gamma Y_{0i} + \boldsymbol{X'_i}\boldsymbol{\delta} + \varepsilon_i \tag{9}$$

In this equation i indexes subject. Y is the outcome and Y_0 is the corresponding baseline

[36] This question will be asked only of subjects assigned to the forecast treatment.

[37] Subjects who responded "not sure" are assigned $R_{id} = 0$.

variable. $\widehat{\boldsymbol{P}}$ is instrumented takeup. Other controls and hypothesis testing are as in the ITT regressions. The first-stage specifications are as follows.

$$\begin{aligned} P_{Ti} &= \eta_T + \boldsymbol{Z_i'}\boldsymbol{\varphi_T} + \nu_T Y_{0i} + \boldsymbol{X_i'}\boldsymbol{\theta_T} + \upsilon_{Ti} \\ P_{Fi} &= \eta_F + \boldsymbol{Z_i'}\boldsymbol{\varphi_F} + \nu_F Y_{0i} + \boldsymbol{X_i'}\boldsymbol{\theta_F} + \upsilon_{Fi} \\ P_{FTi} &= \eta_{FT} + \boldsymbol{Z_i'}\boldsymbol{\varphi_{FT}} + \nu_{FT} Y_{0i} + \boldsymbol{X_i'}\boldsymbol{\theta_{FT}} + \upsilon_{FTi} \end{aligned} \tag{10}$$

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

The high-frequency avoidance measurements again require a different approach. The second- and first-stage specifications are as follows. In contrast to the immediately preceding specifications, they include individual fixed effects $\boldsymbol{\alpha_i}$ and date fixed effects $\boldsymbol{\delta_d}$.

$$\begin{aligned} Y_{id} &= \boldsymbol{\alpha_2 i} + \boldsymbol{\delta_{2d}} + \widehat{\boldsymbol{P_i'}}\boldsymbol{\beta} + \boldsymbol{X_{id}'}\gamma + \varepsilon_{id} \\ P_{Tid} &= \boldsymbol{\alpha_{1Ti}} + \boldsymbol{\delta_{1Td}} + \boldsymbol{Z_i'}\boldsymbol{\varphi_T} + \boldsymbol{X_{id}'}\boldsymbol{\theta_T} + \upsilon_{Tid} \\ P_{Fid} &= \boldsymbol{\alpha_{1Fi}} + \boldsymbol{\delta_{1Fd}} + \boldsymbol{Z_i'}\boldsymbol{\varphi_F} + \boldsymbol{X_{id}'}\boldsymbol{\theta_F} + \upsilon_{Fid} \\ P_{FTid} &= \boldsymbol{\alpha_{1FTi}} + \boldsymbol{\delta_{1FTd}} + \boldsymbol{Z_i'}\boldsymbol{\varphi_{FT}} + \boldsymbol{X_{id}'}\boldsymbol{\theta_{FT}} + \upsilon_{FTid} \end{aligned} \tag{11}$$

We also estimate effects by collapsing the data to the subject level and employing the same 2SLS specifications as for other outcomes.

Effects are estimated between subjects for willingness to pay for three months of SMS forecasts (PKR). The first- and second-stage estimating equations are as follows.

$$\begin{aligned} Y_i &= \alpha + \widehat{\boldsymbol{P_i'}}\boldsymbol{\beta} + \boldsymbol{X_i'}\boldsymbol{\gamma} + \varepsilon_i \\ P_{Ti} &= \eta_T + \boldsymbol{Z_i'}\boldsymbol{\varphi_T} + \boldsymbol{X_i'}\boldsymbol{\theta_T} + \upsilon_{Ti} \\ P_{Fi} &= \eta_F + \boldsymbol{Z_i'}\boldsymbol{\varphi_F} + \boldsymbol{X_i'}\boldsymbol{\theta_F} + \upsilon_{Fi} \\ P_{FTi} &= \eta_{FT} + \boldsymbol{Z_i'}\boldsymbol{\varphi_{FT}} + \boldsymbol{X_i'}\boldsymbol{\theta_{FT}} + \upsilon_{FTi} \end{aligned} \tag{12}$$

One- and two-tailed hypothesis tests for primary outcomes are analogous to those in our ITT regressions.

In addition, a variant of the between-subjects TOT specification is used to estimate the effect of baseline mask takeup on all endline outcomes. The instrument Z_M is the randomly drawn BDM price, the treatment P_M is mask takeup, and the control set $\boldsymbol{X}$ includes a set of dummies for subjects' BDM bids. These estimated mask treatment effects are of secondary interest. They are not included in corrections for multiple hypothesis testing.

Finally, a similar approach is used to estimate the effect of winning the endline BDM for our SMS service on several outcomes measured from a phone survey one week after the endline. For this analysis, we limit to the sample of households that were in the control group for the SMS service treatment and thus had never received it before. The instrument Z_S is the randomly drawn BDM price, the treatment P_S is SMS takeup, and the control set $\boldsymbol{X}$ includes a set of dummies for subjects' BDM bids. These estimated SMS service treatment effects are of secondary interest, as they reflect less than a week of experience with our forecasts. They are not included in corrections for multiple hypothesis testing.

5.2 Heterogenous effects

The theory of Section 2 predicts several important dimensions of heterogeneity: 1) risk aversion; 2) perceived forecast quality; and 3) expectations. First, Rosenzweig and Udry (2014a) point out that forecasts serve an insurance function in that they allow agents to take costly variance-reducing action before uncertainty is resolved. A model of this type predicts that more risk-averse agents may respond differently to forecast provision than less risk-averse agents. Second, treated subjects who perceive the forecasts to be low-quality will likely respond less than subjects who perceive them to be high quality. Third, treatment effects may vary both within and across subjects because of differences in expectations. For example, receiving a forecast of high pollution may increase avoidance for a subject with low expectations but decrease it for a subject with high expectations.

Several other dimensions of heterogeneity have an *ex ante* foundation, but are not the focus of our study and results are reported in the appendix. Subjects who are the primary carriers of a single household cell phone may be more exposed to our SMS forecasts and so exhibit larger responses. Subjects with higher baseline concern about air pollution may respond more to treatment. The interaction of treatment with baseline knowledge of air pollution is not obvious. Highly informed subjects may respond less if there are ceiling effects on avoidance, but if such effects are absent they may respond more. A large economics literature has found larger health effects of air pollution for the very young, the very old, and people with respiratory conditions. The presence of a household member in one of these categories may be an important dimension of heterogeneity. Subjects who are more informed or concerned about air pollution and weather at baseline may exhibit different effects. More educated or numerate subjects may exhibit larger effects, as hypothesized in Chang et al. (2016) , because they are more sophisticated producers and consumers of forecasts. Subjects of different ages may respond differently, e.g. because of differing preferences or remaining years of life. Finally, our letter of support from the government of Punjab requires that we examine heterogeneous effects by gender. All estimates of heterogeneity are excluded from corrections for multiple hypothesis testing.

5.3 Control variables: machine learning and missing values

As indicated in Section 5.1.1, we employ post-double selection LASSO to choose a precision-maximizing control set (Ahrens, 2018). This is consistent with the recommendation of Ludwig, Mullainathan, and Spiess (2019) .[38] Further improvements in precision are possible, as we have not yet used the entire set of available controls as inputs to the double-LASSO selection procedure.

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.

5.4 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 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.

[38]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.

6 Results

6.1 Baseline balance and attrition

To assess treatment-control balance at baseline, Table 1 reports means for all four experimental groups. For each treatment mean in columns two through four, stars signal the result of a hypothesis test of the difference relative to the control group. All differences are small, and only 3 of 54 (5.6 percent) are statistically significant at the ten percent level. To test the joint significance of these differences, we regress group indicators on all listed variables, then conduct F tests of the joint null hypothesis that all coefficients are zero. Results appear at the bottom of Table 1. In all four columns we fail to reject the null hypothesis at any conventional significance level. This suggests our randomization procedure worked as designed.

The overall attrition rate between baseline and endline was 8.2 percent (89 subjects). This was modestly lower than the 12.5 percent prediction of our pre-analysis plan. Table A1 presents attrition rates by group. The maximum pairwise difference is 3.7 percentage points and no difference is statistically significant. Table 2 reports group means of baseline variables for non-attriting subjects. F tests provide no evidence of attrition-induced imbalance. Endline imbalance remains possible on variables used for blocking or stratification during randomization. To investigate, Table A2 reports group means for baseline measures of primary outcomes among non-attritors, and again groups are strongly similar. Taken together, these balance and attrition checks are consistent with our identifying assumption: equality of average potential outcomes in the control state of the world across all four groups.

6.2 Primary outcomes, intent to treat

We begin by examining demand for our SMS air pollution forecasts. Willingness to pay was elicited from all subjects at endline using a Becker-DeGroot-Marschak mechanism (Becker, DeGroot, and Marschak, 1964), in which subjects bid on 90 additional days of our SMS air pollution forecasts.[39] The maximum bid was 200 PKR. As pre-specified, our analysis focuses on subjects exposed only to the forecast treatment. Forecasts are plausibly an experience good, and the demand of these subjects reflects months of interaction and learning.[40] This

[39]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.

[40]For a discussion of how experienece with our SMS forecast treatment affected willingness to pay for forecasts, see Section 6.4.

informed demand is the relevant estimand for a policymaker contemplating distribution of government forecasts and conducting a benefit-cost analysis. Figure 2 presents a histogram of willingness to pay 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 A4). This is consistent with Hypothesis 3 from the model in Section 2, that willingness to pay for useful forecasts is non-zero.

On a monthly basis, mean WTP of 93.22 PKR represents roughly 60 percent of the cost of 4G mobile internet access.[41] Considering a different benchmark, 93.22 PKR is approximately 19 percent of a day's earnings for an unskilled laborer.

In our low-income subject population, it was by no means obvious *ex ante* that mean willingness to pay would be appreciably positive. Barnwal et al. (2017) found that demand for arsenic testing of wells in Bihar, India was low and elastic. More broadly, a large body of work in development economics has found that demand for preventative health care is both low and strongly elastic (Kremer and Glennerster, 2011). Thornton (2008) finds that even at a zero price, only 34 percent of subjects pick up HIV test results. Even small incentives double 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 forecasts may stem from several factors. First, because the forecasts are delivered by text message, subjects do 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's air was the 12th most polluted in the world based on $PM_{2.5}$ (IQAir, 2019). Our results may not be very informative about settings like Accra or Santiago, where air quality is substantially better. They do potentially shed light on cities

[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.

with air pollution similar to Lahore's, e.g. Delhi and Dhaka, and on past periods of acute fine particulate pollution in cities like Beijing.

Our other primary hypotheses pertain to regression estimates of treatment effects, which are presented in Table 3. Column headings indicate dependent variables and shaded cells denote 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 forecasts reduced forecast error by .08 standard deviations, while training reduced forecast error by .11 standard deviations. The former estimate is statistically significant at the five percent level, while the latter is statistically significant at the one percent level. As discussed in Section 5, we pre-specified one-tailed tests for some estimates. In Table 3, stars * denote two-tailed significance, daggers † denote left-tailed significance, and double daggers ‡ denote right-tailed significance.

The estimate for the interaction of treatments is positive, so the effect on the group that received both treatments was $-.08 + -.11 + .11 = -.08\sigma$. While the estimated interaction effect is only marginally statistically significant, it is consistent with gross substitutibility of information and human capital in the production of forecasts. A similar pattern obtains in all columns of Table 3, with estimated interaction effects taking the sign opposite that of the forecast and training effects. This pattern is also potentially consistent with constraints on recall and/or cognition. Treatment interactions were not the focus of our experimental design. As we cannot falsify any of the potential mechanisms behaind them, we do not discuss interaction effects on subsequent outcomes.

The reductions in forecast error from forecast provision and training are practically large. Estimating effects in concentration rather than standard deviations, 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.[42] 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.[43]

Comparisons to other experiments in which treatment is designed to reduce forecast error require care, due to differences in setting, time horizon, and forecast scoring. Mellers et al. (2014) find that probability training improved mean standardized Brier score, a measure

[42] Both the United States and the European Union employ more stringent standards.
[43] The standard deviation of training duration was 15 minutes.

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. (2016) find a six to 12 percent improvement in Brier scores, again roughly similar to our estimated effects. While participants were drawn from 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. (2016) contained substantially more material and was more complex than ours. It is striking that a modestly shorter, simpler training, conducted with less educated subjects, yielded a coarsely similar improvment in forecast performance for air pollution.

Column 2 of Table 3 presents effects on the variance of happiness, as reported by subjects on a five-point Likert scale. The question at endline was, "How variable has your level of happiness been from day to day over the past week?"[44] 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).

Column 3 reports effects on willingness to pay for N95 particulate-filtering masks, elecited by a BDM mechanism with a maximum bid of 200 PKR.[45] The forecast intervention increased WTP by 6.79 PKR and this estimate is statistically significant at the five percent level. This result is consistent with Hypothesis 1 from the model in Section 2. The estimated effect of training is positive at 3.92 PKR, but not statistically significant. This point estimate could stem from complementarity of masks and forecasts in the production of avoidance—that is, subjects in the forecast group may have higher WTP for masks because they are able to produce forecasts with smaller errors.[46] 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 behaviors related to drinking water in developing countries. Madajewicz et al. (2007) find a large increase in the probability of switching wells when a household is informed of arsenic contamination, while Jalan and Somanathan (2008) find households informed of fecal water contamination are more likely to begin purifying their water.

As discussed in Section 5.4, p values corresponding to parimary hypotheses are adjusted for multiple hypothesis testing using the procedure of Benjamini, Krieger, and Yekutieli

[44] At baseline, we asked "How variable has your level of happiness been over the past month?" While not identical, we will still consider our baseline measure as a control at endline for precision.

[45] At baseline the maximum bid was 150 PKR. Despite this difference in censoring, we employ baseline WTP as a control corresponding to Y_{0i} in Equation 7.

[46] Zhang and Mu (2018) find that mask demand in China increases when pollution is high, and the same may be true in Lahore, but our design holds pollution fixed across the treatment and control groups.

(2006) , which controls the false discovery rate. Corrected values are presented in Table 4. 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), $p = .06$ for forecasts and $p = .04$ 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, $p = .06$ and we reject a zero null hypothesis at the ten percent level. While we fail to reject the null for the effect of training on willingness to pay for masks, the adjusted value is subjectively small ($p = .12$). Ongoing work with algorithmic control selection, which was pre-specified, may improve available precision.

6.3 Mechanisms: weather forecasts

In Table 3 there is evidence that both the text message forecasts and the forecasting training reduced air pollution forecast error at endline. Broadly there are two possible mechanisms for these effects: i) larger information sets; and ii) improved information processing. Both of our treatments could plausibly have operated through both channels, influencing information-seeking behavior and the "technology" by which subjects generated forecasts from their information. To shed some light on these mechanisms, subjects were given the opportunity to view a weather forecast prior to making their incentivized air pollution forecasts. Weather forecasts are predictive of air pollution. For example, rain produces very large reductions in pollutant concentrations. In essence, subjects were offered a useful forecast input at a zero price.

Table 5 assesses the information set and information processing mechanisms empirically. Column 1 presents effects on baseline air pollution forecast error, interacting treatment with weather forecast takeup. While this analysis was prespecified, weather forecasts were not randomly assigned and we interpret the regression results with caution. Baseline air pollution forecasts were made prior to any treatment, and as expected coefficients on treatment variables are small and not statistically significant. Because weather forecasts are a useful input to air pollution forecasting, there was the possibility that weather forecast takeup would be associated with smaller air pollution forecast errors. Column 1 shows no evidence of this: there is no large or statistically significant association between weather forecast takeup and air pollution forecast error in any experimental group.

Column 2 presents intent-to-treat effects on weather forecast takeup, constructed as the average of dummies for viewing the one-day-ahead and three-day-ahead forecasts at endline. Effects on this variable can be interpreted as changes in the average probability of weather forecast takeup. Estimated coefficients are small and not statistically significant at conventional thresholds. Corresponding 95 percent confidence intervals rule out a large increase in information seeking from either treatment. There is no evidence that the reduced air pollution forecast error in the primary estimates of Table 3 arises from larger information sets (mechanism i). We cannot exclude this mechanism entirely, as treated subjects may have expanded their information sets in ways we do not observe. But in Table 6 below we find no strong evidence of treatment-driven increases in information about pollution and weather.

Column 3 of Table 5 presents effects on endline air pollution forecast error. To begin, note that in the control group weather forecast takeup is surprisingly associated with larger forecast errors. For trained subjects, however, this heterogeneity looks quite different. Trained subjects who did not take up the weather forecast produced air pollution forecast errors roughly .19 standard deviations larger than the control group, with the difference statistically significant at the ten percent level. In contrast, the interaction of training and weather forecast takeup has a coefficient equal to -.34 standard deviations (statistically significant at the one percent level). The marginal effect of training on subjects who did take up the weather forecast is an error reduction of roughly .15 standard deviations.

This difference within the trained group on the dimension of weather forecast takeup (column 3) potentially reflects both selection and causal effects of viewing the weather forecasts. There are several factors, however, that make large selection effects substantially less likely. First, the LASSO control selection procedure chooses no variables as predictive of the interaction of weather forecast takeup and training once randomization block dummies are partialed out. This indicates there is no strong differential selection on observables in the trained group.[47] Second, there is no evidence of reduced air pollution forecast error from selection effects at baseline (column 1). Third, we find no meaningful treatment effects on weather forecast takeup rates (column 2). Training-driven changes in selection would have to operate through a takeup-preserving change in group composition. Fourth, the association of air pollution forecast error with weather forecast takeup in the control group has the opposite (positive) sign, and one can reject a null hypothesis of equality at the one percent level. For selection in the trained group to explain the observed negative association, it would have to dramatically alter the selection seen in the control group. Taken together, the evidence in Table 5 is consistent with training improving subjects' ability to process information when making air pollution forecasts (mechanism ii from the beginning of this section). As selection cannot be ruled out entirely, however, this pattern of results is only

[47] In the control group the LASSO procedure chooses only enumerator dummies.

suggestive.

6.4 Secondary outcomes, intent to treat

Estimated intent-to-treat effects on secondary outcomes are reported in Table 6. Outcomes in Panel A are what one might call forecast outputs: forecasting errors for several outcomes and an index of subjective confidence. In column 1, the effect of training on forecast error in travel time was negligible. This may be because subjects are already very good at forecasting travel time; the mean absolute error in the control group was 6.3 minutes. There is imprecise evidence in column 2 that training reduces error in forecasting one's own labor supply. Relative to control subjects, subjects in the training arm reduced forecast error by 1.55 hours, but this estimate is not statistically significant at conventional thresholds. Average absolute forecast error in the control group was 6.2 hours, so the error reduction from treatment was proportionally large at 25 percent. Column 3 reports null effects on subjective forecast confidence. For the forecasting training this is surprising, as confidence levels did change within the training session. In ongoing work, we are using panel data from telephone surveys to investigate dynamic effects. Our pre-analysis plan also calls for analyzing heterogeneity by gender, which may be particularly important for subjective forecast confidence. Taken together, the results in Panel A are consistent with a durable improvement in forecast ability beyond the air pollution domain, but available precision limits the weight one can place on this evidence.

Outcomes in Panel B of Table 6 are measures of informedness. Column 1 reports effects on an air pollution awareness index, comprised of responses to questions like "To what degree do you agree or disagree with the following statement? I am aware of the air quality in Lahore." Estimated marginal effects are negative, but small relative to the control-group mean and imprecise. Column 2 reports effects on the number of times a respondent viewed any source of air quality information. The effect of our training treatment is negative and statistically significant at the five percent level, suggesting that subjects sought out less air pollution information once they had been trained. Finally column 3 presents effects on the number of times a respondent viewed weather information. Marginal effects are not statistically distinguishable from zero. Taken together, the results in Panel B do not suggest large changes in informedness about air pollution and weather.

Lastly, in Appendix Table A4 we report treatment effects on willingness to pay for forecasts. The estimated effect of the SMS forecast treatment, 5.80PKR, is not statistically significant at conventional thresholds ($p = .11$). The point estimate is consistent with Hypothesis 4 from the model of Section 2, that experience will increase willingness to pay for forecasts. Hypothesis 5 from the model predicts that the interaction of SMS forecasts and training depends on whether information and forecasting skill are complements or substitutes in belief formation (formation of the agent's internal forecast). The estimated interaction effect is noisy, with the 95 percent confidence interval including substantially positive and negative values, but the negative point estimate is consistent with substitutability.

6.5 Primary outcomes, effect of treatment on the treated

Table 7 reports estimated effects of treatment on the treated, instrumenting for takeup with treatment assignment as shown in Equation 9. In the text message forecast condition, endogenous takeup is defined as the average share of forecasts viewed, ranging from zero to one. In the training condition, takeup is a dummy for participation in training. First-stage F statistics are far above critical values for five percent size and bias control. Pre-specified LASSO control selection and other details are just as in Table 3. More than 96 percent of subjects assigned to training took up training, so TOT estimates are not meaningfully different from their ITT counterparts. Subjects receiving text messages viewed them a slightly less than half the time, so TOT estimates are roughly twice as large as their ITT counterparts. As a result, the 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 debiasing training.[48] To put the point another way, the apparent advantage of training in ITT estimates arises largely from high takeup rates, rather than large effects on compliers. Perfect compliers in the text message group increased their willingness to pay for masks by approximately 14 percent of the control-group mean.

6.6 Robustness

6.6.1 Experimenter demand

One might worry that some subject responses, especially related to air pollution avoidance, might have been affected by experimenter demand effects. 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

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

in several ways. First, all of our enumerators were trained to distance themselves from the implementation of treatment activities 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 and selected outcomes to try to mitigate experimenter demand effects. 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 A11. 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, but there is no evidence of it in Table A11.

6.6.2 Spillovers

Given the ease of relaying our forecasts, spillovers may be a first-order concern for our text message forecast treatment. The sampling was designed to mitigate these concerns by separating subjects in space, but some networks may include both treatment and control subjects despite this. We also asked subjects not to share pollution forecasts outside their household. We sought to measure those spillovers we could not eliminate. At endline, subjects in the forecast treatment group were asked directly whether they had shared our forecasts and if so, where. Control group members 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 A10 reports the complete set of spillover frequencies. This evidence on spillovers does not raise substantial bias concerns.

7 Conclusion

In this paper we present evidence of meaningful willingness to pay for air pollution forecasts among developing-country urbanites. Even assuming our estimate of WTP is double the mean among Lahore's 11 million residents, the implied annual aggregate WTP is roughly

2.1 billion PKR, or US$12.5 million.[49] While capital and operating costs for reference-quality air pollution monitors are considerable—the equipment alone typically costs more than US$10,000—this level of demand 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.

In addition, we show that our one-hour forecast training, which aimed to reduce behavioral biases, 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 expensive to administer, other work has demonstrated successful de-biasing from videos and video games, which scale much more cheaply (Morewedge et al., 2015).

We find that exposure to pollution forecasts increased willingness to pay for protective masks. This suggests that in areas where mask-wearing is not yet commonplace, information provision may be an important spur to mask adoption. Our finding that mean WTP for masks was roughly 77 percent of the retail price suggests that modest subsidies may produce large changes in takeup, with concomitant health benefits.

Masks are a private response to environmental information. Somanathan (2010) has hypothesized 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.

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 welfare effects of providing forecasts or training agents to produce more accurate forecasts, they are plausibly considerable, and warrant future research.

[49] $(.5 * 93.22PKR) * (365/90) * 11000000 = 2.079 * 10^9$PKR.

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. 2019. "Management and shocks to worker productivity." Working paper, National Bureau of Economic Research.

Ahrens, H.C.S.M., A. 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.

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.

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.

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.

Chang, T.Y., 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.

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

Chen, Y., A. Ebenstein, M. Greenstone, and H. Li. 2013. "Evidence on the impact of sustained exposure to air pollution on life expectancy from China's Huai River policy." *Proceedings of the National Academy of Sciences* 110:12936–12941.

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.

Ebenstein, A., M. Fan, M. Greenstone, G. He, P. Yin, and M. Zhou. 2015. "Growth, pollution, and life expectancy: China from 1991-2012." *American Economic Review* 105:226–31.

Edwards, J.H., and C. Langpap. 2012. "Fuel choice, indoor air pollution and children's health." *Environment and Development Economics* 17:379–406.

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

Graff Zivin, J., and M. Neidell. 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., 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.

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

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

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.

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.

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.

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.

Munshi, K., and M. Rosenzweig. 2016. "Networks and misallocation: Insurance, migration, and the rural-urban wage gap." *American Economic Review* 106:46–98.

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

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.

North, D.C. 2003. "The new institutional economics and third world development." In *The new institutional economics and third world development*. Routledge, pp. 31–40.

Organization, W.H., et al. 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.

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

Punjab, E. 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*, pp. .

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

Soll, J.B., K.L. Milkman, and J.W. Payne. 2015. "A users 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?-New Edition*. Princeton University Press.

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

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.

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

Zahra-Malik, M. 2017. "In Lahore, Pakistan, Smog Has Become a 'Fifth Season'."

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.

8 Figures

Figure 1: Treatment Groups

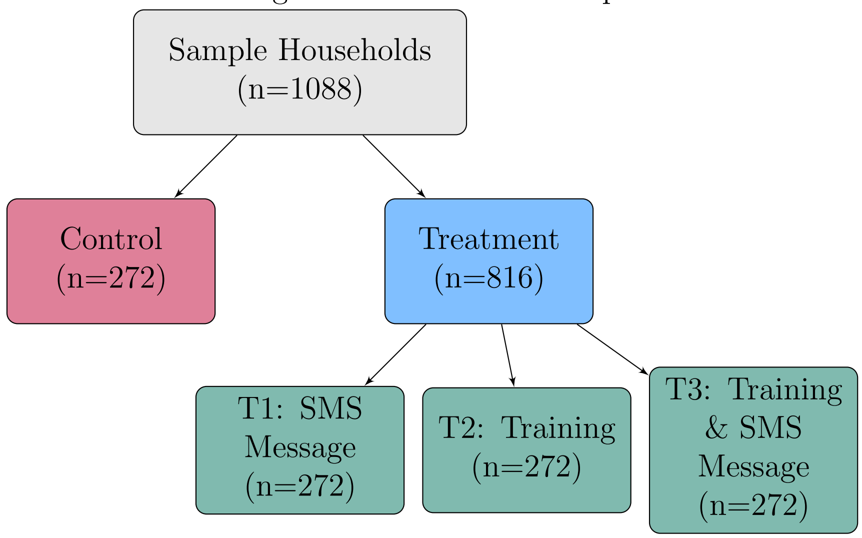
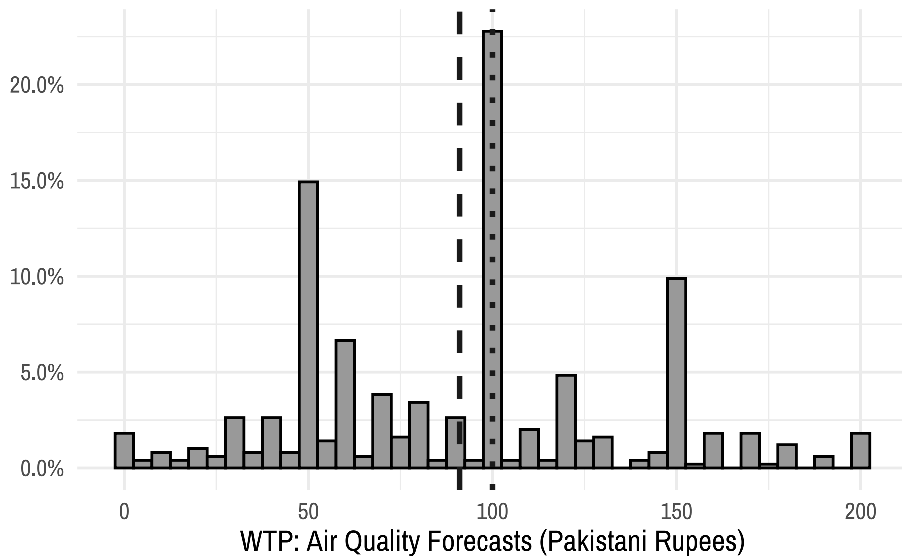


Figure 2: Willingness to pay for air pollution forecasts, forecast-only group



Note: Willingness to pay was elicited from all subjects at endline using a Becker-DeGroot-Marschak mechanism (Becker, DeGroot, and Marschak, 1964), in which subjects bid on 90 additional days of our SMS air pollution forecasts. The maximum bid was 200 PKR. The histogram reflects the forecast-only treatment group (496 subjects), as explained in Section 6.2. The vertical dashed line marks the mean at 93.22 PKR, while the vertical dotted line marks the median at 100 PKR. For a formal hypothesis test of the mean against a zero null, see Table A4.

9 Tables

Table 1: Treatment-control balance

	Control	Forecast	Training	Forecasts + Training
Risk aversion	0.426	0.38	0.38	0.37
	(0.030)	(0.029)	(0.029)	(0.029)
Age (yrs)	31.643	30.56	30.56	31.78
	(0.663)	(0.608)	(0.633)	(0.647)
Travel time forecast (mins)	42.349	43.06	43.57	42.37
	(1.070)	(1.299)	(1.104)	(1.146)
No of people in HH	5.493	5.6	5.68	5.86
	(0.161)	(0.145)	(0.143)	(0.182)
Elderly in HH	0.404	0.4	0.41	0.44
	(0.041)	(0.042)	(0.042)	(0.042)
Children in HH	1.680	1.95*	1.75	1.94*
	(0.099)	(0.112)	(0.103)	(0.116)
HH members employed	1.728	1.69	1.85	1.82
	(0.060)	(0.054)	(0.061)	(0.062)
Female	0.515	0.5	0.5	0.48
	(0.030)	(0.030)	(0.030)	(0.030)
Cares about air quality	0.757	0.76	0.76	0.8
	(0.026)	(0.026)	(0.026)	(0.024)
Aware of city's air quality	0.526	0.52	0.53	0.57
	(0.030)	(0.030)	(0.030)	(0.030)
Aware of county's air quality	0.518	0.53	0.55	0.56
	(0.030)	(0.030)	(0.030)	(0.030)
Accessed weather forecast last week	1.220	1.32	1.38	1.47*
	(0.087)	(0.094)	(0.101)	(0.100)
Air quality masks work	0.939	0.92	0.92	0.93
	(0.016)	(0.018)	(0.018)	(0.016)
Respiratory disease in HH	0.143	0.15	0.18	0.15
	(0.021)	(0.022)	(0.023)	(0.022)
Reduced working hours due to air quality	3.518	3.61	3.42	10.17
	(0.072)	(0.070)	(0.070)	(4.590)
Car/Jeep	0.045	0.06	0.06	0.05
	(0.013)	(0.014)	(0.014)	(0.013)
Mobile Phone	0.208	0.17	0.17	0.24
	(0.032)	(0.030)	(0.031)	(0.037)
Enumerator ID	125.592	125.3	125.34	125.42
	(0.228)	(0.206)	(0.210)	(0.233)
N	272	272	272	272
F	1.10	1.30	0.796	0.723

Note: For each treatment mean, a two-tailed test of the difference from the control was conducted against a zero null. F statistics correspond to regressions of group dummies on all listed covariates. Two-tailed significance: $p<0.1$*; $p<0.05$**; $p<0.01$***.

Table 2: Treatment-control balance, non-attritors

	Control	Forecast	Training	Forecasts + Training
Risk aversion	0.453	0.415	0.417	0.404
	[0.499]	[0.494]	[0.494]	[0.492]
Age (yrs)	31.746	30.671	30.308	31.580
	[10.965]	[10.178]	[10.447]	[10.436]
Travel time forecast (mins)	42.047	43.098	43.065	42.484
	[17.053]	[21.141]	[18.672]	[18.608]
No of people in HH	5.527	5.606	5.563	5.892
	[2.704]	[2.276]	[2.239]	[3.075]
Elderly in HH	0.410	0.382	0.405	0.436
	[0.680]	[0.676]	[0.691]	[0.698]
Children in HH	1.656	1.919	1.741	1.924
	[1.623]	[1.528]	[1.710]	[1.896]
HH members employed	1.734	1.687	1.838	1.824
	[1.002]	[0.845]	[0.987]	[1.030]
Female	1.500	1.472	1.482	1.484
	[0.501]	[0.500]	[0.501]	[0.501]
Cares about air quality	2.387	2.305	2.377	2.289
	[1.019]	[0.935]	[0.963]	[0.927]
Aware of city's air quality'	2.773	2.679	2.718	2.651
	[1.132]	[0.985]	[1.027]	[1.010]
Aware of county's air quality	2.774	2.753	2.751	2.754
	[1.105]	[1.066]	[1.094]	[1.128]
Accessed weather forecast last week	3.612	3.541	3.802	3.688
	[1.823]	[1.797]	[1.773]	[1.736]
Air quality masks work	1.060	1.088	1.081	1.074
	[0.238]	[0.284]	[0.273]	[0.262]
Respiratory disease in HH	1.852	1.846	1.826	1.856
	[0.356]	[0.362]	[0.380]	[0.352]
Reduced working hours due to air quality	2.496	2.398	2.583	2.343
	[1.185]	[1.173]	[1.165]	[1.083]
Car/Jeep	1.952	1.938	1.942	1.952
	[0.215]	[0.241]	[0.234]	[0.215]
Mobile Phone	2.645	2.565	2.838	2.815
	[1.278]	[1.161]	[2.792]	[1.844]
Enumerator ID	125.523	125.033	125.154	125.428
	[3.797]	[3.345]	[3.379]	[3.861]
N	256	246	247	250
F	0.580	1.237	1.104	0.927

Note: Standard deviations appear in square brackets. F statistics correspond to regressions of group dummies on all listed covariates.

Table 3: Primary outcomes, intent to treat

	Forecast error index	Happiness variance	WTP: Masks	Avoidance idx.
Forecasts	-0.08††	0.01	6.79‡‡	0.02
	(0.05)	(0.07)	(3.53)	(0.06)
Training	-0.11†††	0.06	3.92	0.01
	(0.05)	(0.07)	(3.54)	(0.06)
Forecasts + Training	0.11	-0.07	-7.50	-0.01
	(0.06)	(0.09)	(5.03)	(0.08)
Observations	997	994	998	998
Control mean	0	1.86	104.14	0

Note: Estimates correspond to Equation 7, with the dependent variable indicated in the column heading. Shaded cells denote primary hypotheses. All columns include 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 left-, right-, or two-tailed test was conducted for each estimate in accordance with the pre-analysis plan. Two-tailed significance: $p<0.1^{*}$; $p<0.05^{**}$; $p<0.01^{***}$. Left-tailed significance: $p<0.1^{\dagger}$; $p<0.05^{\dagger\dagger}$; $p<0.01^{\dagger\dagger\dagger}$. Right-tailed significance: $p<0.1^{\ddagger}$; $p<0.05^{\ddagger\ddagger}$; $p<0.01^{\ddagger\ddagger\ddagger}$

Table 4: MHT-adjusted p-values, primary outcomes (ITT)

	WTP: Forecasts	Forecast error index	Happiness variance	WTP: Masks	Avoidance idx.
Forecasts	-	0.06	-	0.06	0.27
Training	-	0.04	0.43	0.12	0.27
Forecasts + Training	-	-	-	-	-
Mean, forecast-only group	0.001	-	-	-	-

Note: The p value in column 1 corresponds to the test illustrated in Figure 2 and formalized at the bottom of column 1, Appendix Table A4. The p values in columns 2 through 5 correspond to the tests in Table 3. The MHT correction is performed using the procedure of Benjamini, Krieger, and Yekutieli (2006) , which controls the false discovery rate.

Table 5: Mechanisms: weather forecasts

	Forecast error index (baseline)	Weather forecast takeup (endline)	Forecast error index (endline)
Forecasts	-0.0326	0.00693	-0.0179
	(0.116)	(0.0227)	(0.113)
Training	0.126	-0.0378	0.186*
	(0.116)	(0.0247)	(0.0968)
Forecasts + Training	0.0426	0.00711	-0.0277
	(0.171)	(0.0342)	(0.145)
Weather forecast takeup	0.0996		0.404***
	(0.103)		(0.104)
Forecasts=1 × Weather forecast takeup	0.0605		-0.0615
	(0.141)		(0.133)
Training=1 × Weather forecast takeup	-0.0345		-0.335***
	(0.142)		(0.124)
Forecasts + Training=1 × Weather forecast takeup	-0.164		0.154
	(0.207)		(0.178)
Observations	998	998	997

Note: Estimates correspond to Equation 7, with the dependent variable indicated in the column heading. All columns include block indicators. A pre-specified LASSO procedure was used to select additional controls separately for each outcome. Heteroskedasticity-robust standard errors are in parentheses. Two-tailed significance: $p<0.1$*; $p<0.05$**; $p<0.01$***.

Table 6: Secondary outcomes, intent to treat

	Panel A: Forecast errors and Confidence		
	Travel time	Labor Supply (paid)	Confidence idx
Forecasts	0.56	-1.12	0.34
	(0.44)	(1.13)	(1.28)
Training	0.02	-1.55	-0.35
	(0.35)	(1.30)	(1.30)
Forecasts + Training	-0.28	-0.79	-1.90
	(0.60)	(1.51)	(1.80)
Observations	997	504	998
Control mean	6.33	6.23	69.45

	Panel B: Information		
	Awareness idx	Air quality info.	Weather info.
Forecasts	-0.13	0.06	0.00
	(0.10)	(0.13)	(0.16)
Training	-0.16	-0.28**	-0.23
	(0.10)	(0.13)	(0.16)
Forecasts + Training	0.24	0.35*	0.45**
	(0.15)	(0.19)	(0.23)
Observations	1088	978	993
Control mean	3.13	1.53	2.72

Note: Estimates correspond to Equation 7, with the dependent variable indicated in the column heading. All columns include block indicators. A pre-specified LASSO procedure was used to select additional controls separately for each outcome. Heteroskedasticity-robust standard errors are in parentheses. Two-tailed significance: p<0.1*; p<0.05**; p<0.01***.

Table 7: Primary outcomes, effect of treatment on the treated

	Forecast error index	Happiness variance	WTP: Masks	Avoidance idx.
% Forecasts seen	-0.172†	0.000	15.443‡‡	0.055
	(0.106)	(0.001)	(8.142)	(0.134)
Attended training	-0.091††	0.000	4.454	0.015
	(0.049)	(0.000)	(3.810)	(0.062)
% Forecasts seen	0.239	-0.001	-16.327	-0.032
× Attended training	(0.156)	(0.001)	(12.199)	(0.199)
Observations	997	994	998	998
Control mean	-0.00	-0.25	104.14	0.00
1st stage F-stat	168.61	195.17	173.84	175.97

Note: Estimates correspond to Equation 9, with the dependent variable indicated in the column heading. Shaded cells denote primary hypotheses. All columns include 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 left-, right-, or two-tailed test was conducted for each estimate in accordance with the pre-analysis plan. Two-tailed significance: p<0.1*; p<0.05**; p<0.01***. Left-tailed significance: p<0.1†; p<0.05††; p<0.01†††. Right-tailed significance: p<0.1‡; p<0.05‡‡; p<0.01‡‡‡

A Additional figures

B Additional tables

Table A1: Attrition rates by experimental condition

	Control	Forecast	Training	Forecasts + Training
Attrited from endline dummy	0.059	0.096	0.092	0.081
	[0.236]	[0.295]	[0.289]	[0.273]
N	272	272	272	272

Note: Standard deviations appear in square brackets.

Table A2: Balance, non-attritors, primary outcomes at baseline

	Control	Forecast	Training	Forecasts + Training
Forecast error index (baseline)	-0.002	0.017	0.108	0.014
	[0.953]	[0.976]	[1.023]	[0.978]
WTP: Masks (baseline)	90.000	89.110	89.615	89.640
	[35.852]	[34.647]	[36.704]	[35.886]
Avoidance idx (baseline)	0.005	-0.014	0.094	-0.043
	[0.788]	[0.778]	[0.816]	[0.778]
Happiness variance (baseline)	2.803	2.746	2.703	2.611
	[0.978]	[0.987]	[1.046]	[0.926]
N	256	246	247	250
F	1.156	0.202	2.664	1.986

Note: Standard deviations appear in square brackets. 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 A3: Statistical Balance between Individuals in Lahore Cantonment (Walton) and Rest of Lahore, MICS 2011.

Variable	(1) Walton Mean/SE	(2) Rest of Lahore Mean/SE	T-test P-value (1)-(2)
Primary Education (Household Head)	0.106	0.121	0.041**
	(0.007)	(0.002)	
Secondary Education (Household Head)	0.198	0.213	0.102
	(0.008)	(0.002)	
Higher Education (Household Head)	0.270	0.185	0.000***
	(0.009)	(0.002)	
Age	25.620	24.951	0.104
	(0.407)	(0.113)	
Wealth Index	1.258	0.960	0.000***
	(0.014)	(0.005)	
Cough Problems	0.007	0.016	0.001***
	(0.002)	(0.001)	
Tuberclosis	0.004	0.003	0.165
	(0.001)	(0.000)	
Child in Household	0.369	0.369	0.980
	(0.010)	(0.003)	
Elders in Household	0.058	0.047	0.013**
	(0.005)	(0.001)	
N	2214	26847	

Notes: The value displayed for t-tests are p-values. All missing values in balance variables are treated as zero.***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table A4: Willingness to pay for air pollution forecasts

	WTP: Forecast (1)	WTP: Forecast (2)
Forecasts	4.46	5.80
	(4.22)	(3.63)
Training	-1.22	2.43
	(4.08)	(3.61)
Forecasts + Training	-3.08	-5.12
	(5.72)	(5.00)
Constant	88.76***	108.81
	(2.96)	(15.99)
Observations	999	998
Forecasts group mean	93.22‡‡‡	
	(3.00)	

Note: Column 1 reports a regression of forecast WTP 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. 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 7, with forecast WTP as the outcome. Block dummies are included. A pre-specified LASSO procedure was used to select additional controls. Heteroskedasticity-robust standard errors are in parentheses. Right-tailed significance: $p<0.1^{‡}$; $p<0.05^{‡‡}$; $p<0.01^{‡‡‡}$

Table A5: Primary results (without baseline controls)

	Forecast error index	Happiness variance	WTP: Masks	Avoidance idx.
Forecasts	-0.05	-0.01	6.79‡‡	-0.01
	(0.07)	(0.08)	(3.80)	(0.07)
Training	-0.12††	-0.01	3.92	-0.06
	(0.06)	(0.08)	(3.82)	(0.06)
Forecasts + Training	0.08	-0.03	-7.58	0.07
	(0.09)	(0.12)	(5.41)	(0.09)
Observations	998	995	999	1088
Control mean	0	1.86	104.14	0

Note: Estimates correspond to Equation 7 without the baseline variable for controls or the vector of controls. All columns include block indicators. Shaded cells denote primary hypotheses. Heteroskedasticity-robust standard errors are in parentheses. A left-, right-, or two-tailed test was conducted for each estimate in accordance with the pre-analysis plan. Two-tailed significance: p<0.1*; p<0.05**; p<0.01***. Left-tailed significance: p<0.1†; p<0.05††; p<0.01†††. Right-tailed significance: p<0.1‡; p<0.05‡‡; p<0.01‡‡‡

Table A7: Effects on forecast error denominated in $\mu g/m^3$

	Forecast error
Forecasts	-4.32††
	(2.62)
Training	-5.35††
	(2.63)
Forecasts + Training	5.98
	(3.66)
Observations	998
Control mean	63.78

Note: Specification is as in column 1 of Table 3, but with average error (t+1 and t+3) denominated in $\mu g/m^3$, rather than control-group standard deviations. As in column 1 of Table 3, tests are left-tailed. Left-tailed significance: $p<0.1^{\dagger}$; $p<0.05^{\dagger\dagger}$; $p<0.01^{\dagger\dagger\dagger}$

Table A8: Secondary results - II

	Income variance	Expenditure variance	Children's avoidance
Forecasts	-0.06	-0.08	0.05
	(0.06)	(0.07)	(0.05)
Training	-0.08	-0.08	0.06
	(0.06)	(0.07)	(0.05)
Forecasts + Training	0.11	0.20*	-0.10
	(0.08)	(0.10)	(0.06)
Observations	952	993	483
Control mean	1.49	2.18	0.33

Note: Estimates correspond to Equation 7, with the dependent variable indicated in the column heading. Pre-specified LASSO procedures were used to select controls: risk aversion elicitation, date of survey, enumerator, avoidance behaviours and demographics. Heteroskedasticity-robust standard errors are in parentheses. Two-tailed significance: p<0.1*; p<0.05**; p<0.01***.

Table A10: Spillover frequencies, non-SMS group

# spillover messages	# control HHs
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, comprised of 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 A11: Effects on a social desirability index

	Soc. Desirability Idx
Forecasts	-0.139
	(0.126)
Training	-0.00642
	(0.123)
Forecasts + Training	0.0388
	(0.179)
Observations	998

Note: Estimates correspond to Equation 7, with the dependent variable indicated in the column heading. All columns include block indicators. A pre-specified LASSO procedure was used to select additional controls separately for each outcome. Heteroskedasticity-robust standard errors are in parentheses. Two-tailed significance: p<0.1*; p<0.05**; p<0.01***.

C Intervention details

C.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.[50]

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.

[50] 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*.[51] 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.

C.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.[52] Subjects received 150 PKR for their participation.

Each training session began with incentivized elicitations of air pollution and travel time forecasts. Over the course of the session, we elicited non-incentivized forecasts of the same outcomes (air pollution and travel time) 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 Pakistans 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

[51] 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*.

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

et al., 2014). The calibration exercises were intended to show subjects that they had room for improvement and open their minds to subsequent lessons.

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.

D Data and sampling details

D.1 Sampling and subjects

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 SurveyCTOs Open Data Kit (ODK) server.

Following the sampling frame described in Section 3.2, 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 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;[53] 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 nth household member. The enumerator then asked to speak with the nth listed eligible individual to

[53] 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.

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. Forecast elicitation (air pollution and travel time);
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 three different stages of the experiment.

1. **In-person surveys:** In-person baseline and endline surveys of all respondents were conducted.
2. **Telephone surveys:** Following a one week transitionary period from the start of our SMS intervention, all respondents were surveyed weekly by phone. To do this, respondents were placed in random order at the beginning of the first week and surveyed in that order every week. Enumerators attempted to contact every respondent at two significantly spaced-out times on a particular day and on $t + 1$. If in all three attempts of reaching out, the respondent did not answer the phone call, the respondent was recorded as non-responsive for that week. The weekly phone data collection lasted approximately 12 weeks.
3. **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.

D.4 Air pollution data

1. **AQMesh and Dusttrak II:** We used two industrial-grade monitors: (1) the AQMesh; and (2) the Dusttrak II.[54] 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. EPAs 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.[55]

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).[56] 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.[57] 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.

---

[54] 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/.

[55] 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.

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

[57] 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.[58] Data included temperature levels, precipitation levels and cloud cover.

D.6 Traffic data

- **Google Maps:** Travel time data were collected from Google Maps from one month before the baseline survey through the endline survey. Since Google Maps does not allow data scraping, this information was manually collected by a research assistant.

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

E 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 the 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 not interesting.) We use the avoidance index as a primary outcome, and analyze the plausibly heterogeneous effects on outdoor time under secondary outcomes. None of the three estimated treatment effects for the avoidance index are statistically significant, so this change has no impact on the number of statistically significant results.

- Hypothesis tests on willingness to pay for masks. In the PAP we made contradictory claims about alternative hypothesis (one- vs. two-tailed tests) for the avoidance index and willingness to pay for masks, even though both are qualitatively similar avoidance behaviors. We resolved this inconsistency in favor of one-tailed tests on both outcomes.

F 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 behaviour. 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 normalise 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 behaviour, the agent faces expected utility (baseline)

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

Now, assume that the agent may engage in avoidance behaviour (e.g purchase a mask, re-schedule 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 $a \leq X$.[59] By engaging in avoidance behaviour 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.[60] Then every day, with a probability of p, she either chooses to avoid (in response to what she believes is a high

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

[60] 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.

pollution day), or not avoid with probability $(1-p)$. This probabilistic response to pollution would yield expected pay-off

$$N = p(pu(C-a) + (1-p)u(C-X)) + (1-p)(pu(C-a) + (1-p)u(C)). \quad (13)$$

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 13 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.

F.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.[61] 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 realised 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

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

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.

F.2 Forecasting when avoidance is not too costly

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

$$p[\pi u(C-a-f)+(1-\pi)u(C-X-f)]+(1-p)[\pi u(C-f)+(1-\pi)u(C-a-f)] \geq u(C-a). \quad (14)$$

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 (14) 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.

F.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

$$
\begin{aligned}
T^n(f) =& p[\pi u(C-a-f) + (1-\pi)u(C-X-f)] + (1-p)[\pi u(C-f) + (1-\pi)u(C-a-f)] \\
& - pu(C-X) - (1-p)u(C) \geq 0. \qquad (15)
\end{aligned}
$$

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

$$
\begin{aligned}
p[\pi u(C-a) + (1-\pi)u(C-X)] & \\
+(1-p)[\pi u(C) + (1-\pi)u(C-a)] &\geq pu(C-X) + (1-p)u(C), \\
p[\pi u(C-a) + (1-\pi)u(C-X) - u(C-X)] &\geq (1-p)[u(C) - \pi u(C) - (1-\pi)u(C-a)], \\
p\pi[u(C-a) - u(C-X)] &\geq (1-p)(1-\pi)[u(C) - u(C-a)]. \qquad (16)
\end{aligned}
$$

This formulation provides intuition behind the agent's choice. The left-hand side in equation (16) 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 16 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 analyse 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 derivates 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, $\frac{\partial T^n}{\partial \pi} = p[u(C-a-f) - u(C-X-f)] + (1-p)[u(C-f) - U(C-a-f)] \geq 0$. □

F.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

$$pu(C-a-f) + (1-p)u(C-f) \geq pu(C-X) + (1-p)u(C). \tag{17}$$

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 (17), 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 $u(x, \alpha)$.

$$\begin{aligned} u(ce(x, \alpha), \alpha) &= pu(0, \alpha) + (1-p)u(x, \alpha), \\ &= (1-p)u(x, \alpha). \end{aligned} \tag{18}$$

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 (18) 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.*