Eliciting Human Judgment for Prediction Algorithms

Rouba Ibrahim
School of Management, University College London, rouba.ibrahim@ucl.ac.uk

Song-Hee Kim
Marshall School of Business, University of Southern California, songheek@marshall.usc.edu

Jordan Tong
Wisconsin School of Business, University of Wisconsin-Madison, jordan.tong@wisc.edu

May 13, 2020

Even when human point forecasts are less accurate than data-based algorithm predictions, they can still help boost performance by being used as algorithm inputs. Assuming one uses human judgment indirectly in this manner, we propose changing the elicitation question from the traditional direct forecast (DF) to what we call the private information adjustment (PIA): how much the human thinks the algorithm should adjust its forecast to account for the information that only the human has. Based on a behavioral model, we theoretically prove that, when there is human random error in the forecast, eliciting the PIA leads to more accurate predictions than eliciting the DF; however, this DF-PIA gap does not exist for perfectly-consistent forecasters. The DF-PIA gap is increasing in the random error people make while incorporating public information (data that the algorithm has access to) but is decreasing in the random error that people make while incorporating private information (data that only the human has access to). In controlled experiments with students and Amazon Mechanical Turk workers, we find support for these hypotheses and demonstrate the flexibility to conduct elicitation in multiple ways to enhance performance.

Key words: laboratory experiments, behavioral operations, random error, elicitation, forecasting, prediction, discretion, expert input, private information, judgment, aggregation

1. Introduction

Because of increased access to data and advancements in machine-learning algorithms, a common operational improvement initiative is to replace human forecasters with data-driven prediction algorithms. For example, in our motivating setting, a hospital needs surgery duration forecasts to schedule operating room use, which costs $\$2,190$ per hour on average (Childers and Maggard-Gibbons 2018). Using surgery duration data from that hospital, Ibrahim and Kim (2019) showed that physicians' mean absolute percent forecast error was 33%, whereas algorithms based on available patient and surgery data reduced that error to 29%.

Nevertheless, even if humans are not better than algorithms head-to-head, their judgments can still help. In the above example, the hospital could improve predictive accuracy even further to under 27% by using the physicians' forecasts as an input (along with the other data) to the algorithm. In other words, the best forecasts often come not from replacing humans with algorithms, but from combining them.

In this research, we ask the following question: If we know that we are going to use human judgment not directly but *indirectly* in an algorithm, should we elicit something else besides point forecasts? If so, what alternative information should we elicit, and why might it work better?

We theorize that the primary reason why humans add value to algorithms is that they have access to private information that the algorithm does not have access to, for example, because it is not in the database or is too unstructured to use in the algorithm. Therefore, we consider whether, rather than asking for a human's *direct forecast* (DF), it may be better to ask about her private information instead (even if we do not know what this private information is ahead of time). Specifically, we propose the idea of eliciting the *private information adjustment* (PIA)—how much the human thinks the algorithm should adjust its forecast to account for the information that only the human has.

Using a stylized behavioral model (§2), we theorize that the PIA leads to more accurate predictions than DF only if there is human random error. That is, from a predictive accuracy perspective, there is no difference between eliciting the DF or the PIA if people are perfectly consistent in how they use information to make a forecast. However, if they are inconsistent, then the PIA should help algorithms more than the DF. Furthermore, the model sheds light on which environmental conditions would lead to greater differences in performance. Namely, it shows that the PIA's advantage, relative to DF, is larger when "public" data—the data that the algorithm also has access to—is complicated for the human to process, but smaller when the human's private information is complicated to process instead.

To test these hypotheses regarding the difference in performance between DF and PIA, we conducted controlled experiments in which we elicited human judgments for 50 simulated surgery durations based on predictive data. We told our participants that the hospital's algorithm had access to only some of the data ("public information"), but the other data only the participant had access to ("private information"). In one condition, we elicited judgment by asking for the participant's DF for each surgery while, in the other, we elicited their PIA for each surgery. Then, for each condition, we calibrated prediction algorithms using the first 35 surgeries, and tested their predictive performance using the last 15 surgeries.

In Experiment 1 (§3), conducted with university students and replicated with Amazon Mechanical Turk (MTurk) workers, we find that prediction algorithms performed significantly better when they had access to the participants' PIA as inputs as opposed to their DF – their average root mean squared error (RMSE) in the test sets was 21% lower. In Experiment 2 (§4), we manipulate random error magnitudes by making the public or private information more or less complex: subjects must aggregate multiple factors when the information is complex, but are provided one equivalent factor when the information is not complex. Consistent with our theoretical development, we find that the RMSE for PIA is 48% lower than for DF when the public information is complex, but only 6% lower than DF when the private information is complex.

In Experiment 3 (§5), we explore a setting where the PIA is not numerical: We test whether a lower-effort multiple choice version of the PIA question can enhance PIA's performance and its advantage over DF. Finally, in §6, we highlight opportunities for future research.

We contribute to three main bodies of research. Management science researchers have long recognized the potential value in *integrating human judgment with forecasting algorithms* (see Bunn and Wright 1991, Arvan et al. 2019 for reviews). The two most common integration approaches are to make judgmental adjustments to an algorithm's point forecast (e.g., see Carbone et al. 1983, Lim and O'Connor 1995, Fildes et al. 2009, Sanders and Ritzman 2001) or to combine separate human and algorithm point forecasts (e.g., see Lawrence et al. 1986, Blattberg and Hoch 1990, Goodwin 2000). We contribute by examining a different human elicitation question from the point forecast. Notably, our proposed method is *not* equivalent to judgmental adjustments because we use the PIA as an *input* for the prediction algorithm. In fact, in our experiments, using PIA responses to adjust algorithm forecast outputs yields poor predictive performance.

A stream of behavioral operations management research studies the *system design implications of human random error*. For example, the best way to design contracts (e.g., Su 2008, Ho and Zhang 2008), queues (e.g., Huang et al. 2013), or auctions (e.g., Davis et al. 2014) changes once the system designer considers human random error. Most closely related to our paper is Kremer et al. (2016). They show that human random error causes eliciting human forecasts in a top-down fashion to be more effective in some environments, but bottom-up forecasting to be more effective in others. We contribute by showing how a forecasting system's elicitation design impacts performance once one considers human random error, even if it has no effect without human random error.

Finally, researchers in judgment and decision making have made advancements in developing *strategies to improve human judgment accuracy*. Perhaps the most well-known idea is to harness the "wisdom of crowds" (e.g., see Surowiecki 2005) through averaging multiple peoples' judgments. Interestingly, because people are so inconsistent (Kahneman et al. 2016), even averaging multiple judgments by the same person separated by time (Vul and Pashler 2008) or with a prompt to think differently (Herzog and Hertwig 2009) helps, albeit only about half as much as averaging two different people (Mannes et al. 2012). Most closely related to our work in this stream is Palley and Soll (2019), who develop a new elicitation method that improves the wisdom of crowds strategy by estimating the amount of shared information between individuals. Our elicitation strategy also seeks to improve an aggregation strategy by addressing the issue of disentangling the shared information between the human and the algorithm.

2. Theory Development

We present simple models of rational (no random error) and behavioral (random error) forecasters. Then, we characterize the performance of prediction models that use as inputs direct forecasts (DF) or private

information adjustments (PIA). Finally, we use these results to motivate two hypotheses about whether and how eliciting the PIA will be more effective than DF.

2.1. Surgery Duration Assumptions

We assume an actual surgery duration, Y, is a random variable defined by the linear model:

$$Y = v + \sum_{i \in \mathcal{P} \cup \mathcal{I}} w_i X_i + \epsilon, \tag{1}$$

where we separate the public factors, denoted by the index set $\mathcal{P}$, from the private factors, denoted by the index set $\mathcal{I}$. In (1), ϵ is an error term, with $\mathbb{E}[\epsilon] = 0$, which represents true environmental random shocks i.e., random variations that are impossible to predict even with all public and private information. We assume that ϵ and $(X_i)_{i \in \mathcal{P} \cup \mathcal{I}}$ are mutually independent.

2.2. DF and PIA are Equivalent under Rational Forecasting

We define the DF and PIA for the *rational forecaster* as follows:

$$DF^* = v^* + \sum_{i \in \mathcal{P} \cup \mathcal{I}} w_i^* X_i \text{ and } PIA^* = \sum_{i \in \mathcal{I}} w_i^* X_i. \tag{2}$$

In this rational model, we assume that v^* and w_i^*, for $i \in \mathcal{P} \cup \mathcal{I}$, are *deterministic*, though not necessarily known by the algorithm a priori . (Note, they can be any constants and are not necessarily "optimal".) Define the best fitting models of Y given the public factors and DF^* or PIA^* using linear regression:

$$\text{(Model-}DF^*\text{)} \quad M_{DF^*} = \alpha_0 + \sum_{i \in \mathcal{P}} \alpha_i X_i + \beta_{DF^*} DF^*, \tag{3}$$

$$\text{(Model-}PIA^*\text{)} \quad M_{PIA^*} = \gamma_0 + \sum_{i \in \mathcal{P}} \gamma_i X_i + \beta_{PIA^*} PIA^*. \tag{4}$$

Then, the following proposition holds. We relegate the proofs of all propositions to the appendix.

PROPOSITION 1. *Model-DF^* and Model-PIA^* yield the same predictions.*

That is, with rational (deterministic) forecasters, predicting surgery durations using DF as model inputs yields the same predictions as using PIA as model inputs; the two elicitation methods are equivalent from the algorithm's perspective.

2.3. PIA Outperforms DF under Behavioral Forecasting

Next, we consider the following *behavioral forecasting model* for the DF and PIA:

$$DF^b = v^b + \sum_{i \in \mathcal{P} \cup \mathcal{I}} W_i^b X_i \text{ and } PIA^b = \sum_{i \in \mathcal{I}} W_i^b X_i. \tag{5}$$

Here, we assume that W_i^b are *random variables* with $\mathbb{E}[W_i^b] = \bar{w}_i^b$ and $\text{Var}[W_i^b] > 0$. We also assume that W_i^b and X_i are all mutually independent, for $i \in \mathcal{P} \cup \mathcal{I}$. In this way, in contrast to the rational model in (2),

the behavioral model in (5) captures inconsistencies or random error in assigning weights to each factor. One can interpret these random weights as capturing randomness in a variety of cognitive processes, such as in the encoding of information, memory retrieval, aggregation of multiple factors, or the translation from one domain to another.

Define the best fitting models of Y given the public factors and DF^b or PIA^b using linear regression:

$$\text{(Model-}DF^b\text{)} \quad M_{DF^b} = \alpha_0 + \sum_{i \in \mathcal{P}} \alpha_i X_i + \beta_{DF^b} DF^b, \tag{6}$$

$$\text{(Model-}PIA^b\text{)} \quad M_{PIA^b} = \gamma_0 + \sum_{i \in \mathcal{P}} \gamma_i X_i + \beta_{PIA^b} PIA^b. \tag{7}$$

In contrast to the equivalence result in Proposition 1 for the rational model, the following proposition demonstrates the benefit of eliciting PIA compared to eliciting DF under the behavioral model.[1]

PROPOSITION 2. *The mean squared error (MSE) for predictions under Model-DF^b is strictly larger than that under Model-PIA^b, i.e.,* $\mathbb{E}[(Y - M_{DF^b})^2] > \mathbb{E}[(Y - M_{PIA^b})^2]$.

The intuition is that, from the algorithm's perspective, the value of the human input is the private information—the algorithm already has the public information. The algorithm can infer the private information equally well from DF or PIA responses without human random error. However, when there is human random error, the algorithm can more accurately infer the private information from the PIA. Based on this result, we formulate our first hypothesis.

Hypothesis 1 *All else equal, a prediction model that is calibrated using DF yields less accurate predictions than a prediction model that is calibrated using PIA.*

2.4. The DF-PIA Gap Magnitude Depends on Random Error Location

Proposition 2 establishes our main result that, because of human random errors, using PIA yields more accurate predictions than using DF. We now investigate how the "location" of these random errors (i.e., whether they occur incorporating public versus private factors) affects the performance difference between DF and PIA. To do so, we study the behavior of the *DF-PIA gap*, which we define as the difference in the MSEs from Proposition 2, $\mathbb{E}[(Y - M_{DF^b})^2] - \mathbb{E}[(Y - M_{PIA^b})^2]$.

Random Error Incorporating Public Factors. To examine the effect of random error when incorporating public factors on the DF-PIA gap, we consider two cases that are identical except for the degree of variability in W_i^b for $i \in \mathcal{P}$. Specifically, we define $\widetilde{W}_i^b$ to be a mean preserving spread of W_i^b (Rothschild and Stiglitz 1970). The following result establishes that increasing the variability in how people incorporate public factors increases the DF-PIA gap:

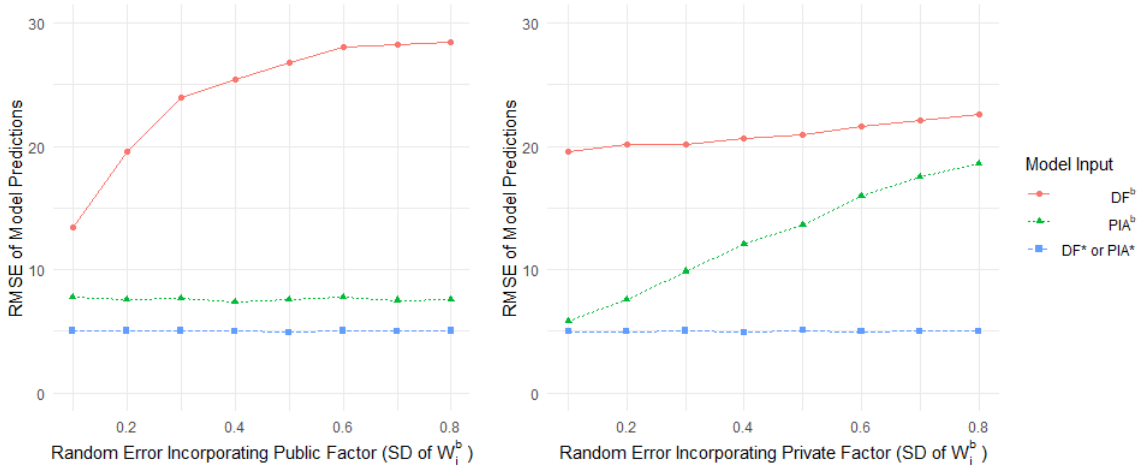
[1] Note that all propositions hold for both MSE and RMSE; we use RMSE to report our experiment results in §3–5.

PROPOSITION 3. *The DF-PIA gap is larger when $\widetilde{W}_i^b$ is used in (5), for $i \in \mathcal{P}$, instead of W_i^b.*

The idea behind Proposition 3 is as follows. Model-PIA^b remains the same when we add variability to $W_i^b, i \in \mathcal{P}$ because PIA responses are unaffected by the random error incorporating public factors. However, Model-DF^b is less accurate when $\widetilde{W}_i^b$ is used instead of $W_i^b, i \in \mathcal{P}$. DF responses are more variable when $\widetilde{W}_i^b$ is used, which makes it harder for the algorithm to learn the private factors. Combining these two observations implies that the DF-PIA gap increases.

Figure 1 shows the results from numerical simulations (see Appendix B for details). The left figure corresponds to Proposition 3. It varies the standard deviation of the public factor random weight, holding constant the standard deviation of the private factor random weight. Observe that the DF-PIA gap increases with the variability in the public factor weight because the RMSE associated with Model-PIA^b remains constant while the RMSE associated with Model-DF^b increases.

Figure 1 Numerical Simulation Illustrations of Propositions 3 and 4.



Random Error Incorporating Private Factors. We now turn to examining the effect of random error incorporating private factors on the DF-PIA gap. We proceed as above, by considering two cases that are identical except for the degree of variability in W_i^b for $i \in \mathcal{I}$. We have:

PROPOSITION 4. *The DF-PIA gap is smaller when $\widetilde{W}_i^b$ is used in (5) for $i \in \mathcal{I}$ instead of W_i^b.*

In contrast to Proposition 3, Proposition 4 shows that adding variability to how people incorporate private factors reduces the DF-PIA gap. Both Model-PIA^b and Model-DF^b lose accuracy as we add variability to $W_i^b, i \in \mathcal{I}$. However, the loss is more dramatic for Model-PIA^b. PIA's advantage of more directly eliciting the private information decreases as the random error incorporating private information increases.

Figure 1 (right) is the corresponding figure for Proposition 4. Observe that the DF-PIA gap decreases in the standard deviation of the private factor random error term because, while random error incorporating the private factor increases the RMSE under both Model-DF^b and Model-PIA^b, the increase is steeper in the latter.

Summary and Hypothesis. Proposition 3 shows that the DF-PIA gap increases in the random error incorporating public factors. Proposition 4 shows that the DF-PIA gap decreases in the random error incorporating private factors. Combined, they imply that the DF-PIA will be greater when adding random error incorporating public information than when adding the same amount of random error incorporating private information. Based on these results, we formulate our second hypothesis:

Hypothesis 2 *The location of random error moderates the DF-PIA gap. Specifically,*

(a) Settings that induce more random error incorporating public information increase the DF-PIA gap.

(b) Settings that induce more random error incorporating private information decrease the DF-PIA gap.

(c) Random error incorporating public information increases the DF-PIA gap more than random error incorporating private information.

3. Experiment 1: Elicitation via Direct Forecast (DF) vs. Private Information Adjustment (PIA)

Experiment 1 is a simple direct test of Hypothesis 1.

3.1. Experimental Design

3.1.1. Task. Participants first reviewed 30 historical surgeries, each with information about the number of procedures, anesthesia complexity score, and the resulting surgery duration. They then completed 50 rounds of surgery duration prediction. In each round, they were shown a new surgery's number of procedures and anesthesia complexity score. Then, they were asked a question about predicting its duration.

3.1.2. Conditions. Subjects were randomly assigned to one of two conditions: direct forecast ("DF") versus private information adjustment ("PIA"). The only difference between these two conditions is that, in each of the 50 rounds, DF participants answered the question: "What is your forecast for the duration of this surgery? I think this surgery duration will be _____ minutes.", whereas PIA participants answered the question: "The hospital system only has the first piece of information about this surgery—the number of procedures. You have additional information. To account for your additional information, how would you advise the hospital system to adjust its forecast for the duration of this surgery? I would advise the hospital system to increase/decrease (choose one) its forecasted surgery duration by _____ minutes." See Appendix, Figure D.1 for screenshots.

3.1.3. Simulating Surgery Duration. We used the following equation to simulate surgery durations: $Y_s = 60 + 20X_s^P + 10X_s^I + \epsilon_s$. Here, Y_s is the duration of surgery s, X_s^P denotes the number of procedures, an integer-valued public factor that has a uniform distribution between 1 and 10, inclusive, and X_s^I denotes the anesthesia complexity score, a private factor that has a uniform distribution between -5 and 5. Finally, ϵ_s follows a normal distribution with mean 0 and standard deviation 5. All participants observed the same 30 simulated historical surgeries. However, each participant observed a unique sequence of randomly-generated surgeries for their 50 prediction rounds.

3.1.4. User Interface and Instructions. We programmed the user interface using the online software *SoPHIE* (Hendriks 2012). After written instructions describing the task, participants were required to pass a three-question comprehension test before starting the experiment. They could review the instructions and retake the test until they answered all questions correctly. See Appendix, Figure D.2 for full instructions.

3.1.5. Pre-registration. For all experiments, we set our target sample sizes, exclusion criteria, and analysis plans a priori. We pre-registered to exclude participants who (1) did not complete the experiment or (2) put the same answer more than 90% of the time. We also pre-registered our dependent variable and analyses. We calibrate prediction algorithms using data from the participants' train set (first 35 rounds), then use the algorithms to generate predictions on the test set (last 15 rounds). Our performance criterion is the root mean squared error (RMSE) of the predictions generated on the test set. The full pre-registration document for Experiment 1 is available at `https://aspredicted.org/blind.php?x=3e427n`.

3.2. Results

3.2.1. Participants and Responses Summary Statistics. Undergraduate and graduate students at a large research university in the US were invited to participate through a behavioral laboratory subject pool recruitment system. Each participant received a $10 electronic gift card for completing the online study.

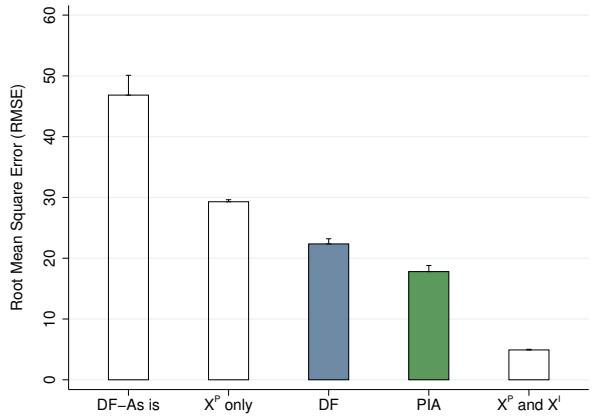
A total of 120 students participated. Following our exclusion criteria, we removed 8 participants who did not complete the experiment, leaving 112 for analysis (56 in each condition). Among the 112 participants, 75% were female, 88% were 18 to 24 years old and 12% were 25 to 34 years old. The average of mean response was 152.6 minutes (SD 30.8) in the DF condition, and 12.1 minutes (SD 23.7) in the PIA condition.

3.2.2. Algorithm Calibration and Prediction Accuracy Calculation. For each condition, we used the number of procedures, actual surgery duration, and participant response from all participants' first 35 rounds to calibrate prediction algorithms for surgery duration. The pre-registered linear regression model included participant dummies, number of procedures interacted with participant dummies, and participant response interacted with participant dummies. Table D.1 in the Appendix summarizes the prediction algorithms calibrated for each condition. We used the calibrated prediction algorithms to generate the predictions, $\hat{Y}_s$, for

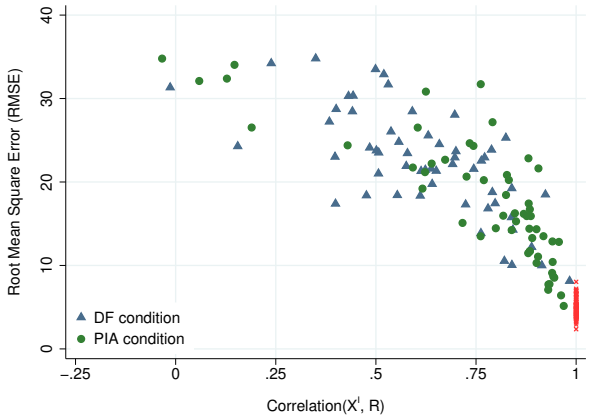
each surgery s in the last 15 rounds of each participant. We then computed $RMSE = \sqrt{\frac{1}{15}\sum_{s=1}^{15}(Y_s - \hat{Y}_s)^2}$ for each participant.

3.2.3. Testing Hypothesis 1. The average RMSE (averaged across all participants) was 22.4 (SD 6.3) in the DF condition and 17.8 (SD 7.6) in the PIA condition; see Figure 2(a). This difference of 4.6 is significant ($p = 0.0008$) and represents a 21% decrease. This result supports Hypothesis 1.

Figure 2 Experiment 1: Performance Comparison.



(a) RMSE comparison. Means and standard errors are shown. X^P is public factor and X^I is private factor.



(b) Correlation(X^I, R) versus RMSE. Each dot is one participant. R is defined as the residual of response after regressing it on X^P. Red x marks show the performance of the "X^P and X^I" model.

3.2.4. Other Benchmarks. Figure 2(a) also shows the performance of three other benchmarks:

"DF-As is" corresponds to using participant DF condition responses without any algorithms. Doing so results in an average RMSE of 46.8 (SD 24.5), significantly worse than when we use participant responses as inputs to algorithms.

"X^P only" corresponds to using *only* the public information in the algorithm, without the use of any participant responses. Across the 112 participants, such an algorithm leads to an average RMSE of 29.3 (SD 3.5)—an improvement over "DF-As is" even though participants had access to the private information, in addition to the public information. However, it is worse than the average RMSE of both the DF condition ($p < 0.0001$) and the PIA condition ($p < 0.0001$). In other words, participant responses added predictive value in the experiment.

Lastly, "X^P and X^I" corresponds to allowing prediction algorithms to directly observe the private information X^I, and include it in prediction algorithms. In that sense, it is a benchmark for the best performance possible. Doing so results in an average RMSE across the 112 participants of 4.9 (SD 1.1).

3.2.5. Mechanism Evidence. The theorized mechanism driving Hypothesis 1 is that PIA responses help the algorithm account for the private information better than the responses from the DF condition. To examine this mechanism, we calculate the correlation of the PIA responses with X^I and compare them with the correlation of the DF responses with X^I (after taking out the effect of X^P).

Naturally, because the PIA asks directly about the private information, the correlation between X^I and response was lower in the DF condition than in the PIA condition (0.39 versus 0.74). Next, we consider the correlation between X^I and R, where we define R as the residual of participant response after regressing it on X^P. That is, we take out the effect of public factor from each response to construct R. Note that if participants did not suffer from random error, then R would be perfectly correlated with X^I in both DF and PIA. In contrast, we find it is less than 1 in both conditions. However, it is significantly higher in the PIA condition than in the DF condition (0.76 versus 0.62, $p = 0.0008$). In other words, PIA responses provide better information about X^I than DF responses.

Figure 2(b) illustrates the predictive accuracy versus the correlation value above, for each participant. As expected, higher correlation between X^I and R leads to better prediction performance. There are more participants with high correlation in the PIA condition than in the DF condition, which contributes to the better performance of the PIA condition overall. The red "x" marks indicate the hypothetical perfect-rationality benchmark with no random error for each participant, which yields perfect correlation for both DF and PIA. The deviation of the PIA and DF dots from the red marks illustrate the effect of human random error in participant responses.

3.3. MTurk Replication

We replicated the same experiment with MTurk workers. See `https://aspredicted.org/blind.php?x=yv2vs7` for the pre-registration document. While, overall, the predictions from the experiment with MTurk workers were less accurate, the between-condition results replicated, providing evidence of robustness across different populations. Appendix C provides details on the replication as well as a comparison between the performances of university students and MTurk workers.

3.4. Discussion

Consistent with Hypothesis 1, Experiment 1 provides evidence that eliciting the PIA information instead of DF leads to better prediction algorithm performance. Participant responses, after the effect of public factor is taken out, are more correlated with the private factor. This tighter relationship helps prediction algorithms to incorporate private information, leading to better predictive performance. These results were replicated across university students and MTurk workers.

4. Experiment 2: Manipulating Information Complexity of Public vs. Private Factors

Experiment 2 was designed to test Hypothesis 2 on how the DF-PIA gap established in Experiment 1 is moderated by random error in incorporating public versus private factors. In addition, it provides a replication test of Hypothesis 1 using different surgery duration equations.

4.1. Experimental Design

The task was similar to that in Experiment 1. However, we changed the surgery duration equation, and we varied the number of factors by condition. We conjectured that greater information complexity induces higher random error. Therefore, we created higher complexity to induce more human random error by requiring that subjects aggregate multiple factors. Otherwise, to create lower complexity to induce less random error, we automatically pre-aggregated multiple factors into a single representative factor for the participant.

Specifically, in the *Baseline* case we pre-aggregated information so that there is only one public and one private factor as in Experiment 1. However, we required that participants account for two public factors in the *Public Info Complex* case, or two private factors in the *Private Info Complex* case. Thus, the experiment was a 2 (DF, PIA) by 3 (*Public Info Complex*, *Private Info Complex*, *Baseline*) between-subject experimental design.

The equations below show the underlying model we used for all conditions and the pre-aggregations we constructed to manipulate complexity by condition:

$$\begin{aligned}
Y_s &= 150 + 10X_s^{P1} + 10X_s^{P2} + 10X_s^{I1} + 10X_s^{I2} + \epsilon_s \quad \text{(Underlying Model)} \\
&= 150 + 10X_s^{P1} + 10X_s^{P2} \quad + (50 + 20X_s^{I}) + \epsilon_s \quad \text{(Public Info Complex)} \\
&= 150 \quad + (50 + 20X_s^{P}) + 10X_s^{I1} + 10X_s^{I2} + \epsilon_s \quad \text{(Private Info Complex)} \\
&= 150 \quad + (50 + 20X_s^{P}) \quad + (50 + 20X_s^{I}) \quad + \epsilon_s \quad \text{(Baseline)}
\end{aligned}$$

Here, X_s^{P1} and X_s^{P2} represent the two public factors. In the experimental task, they are the "procedure set-up requirements" and the "procedure complexity score," respectively. Symmetrically, X_s^{I1} and X_s^{I2} represent the two private factors. In the experimental task, they are the "anesthesia set-up requirements" and the "anesthesia complexity score," respectively. The random generation process for public and private factors was symmetric. For every surgery s, X_s^{P1} and X_s^{I1} were uniform random integers between 0 and 10, inclusive. X_s^{P2} and X_s^{I2} were uniform random numbers between -5 and 5 (rounded to the nearest tenth). We set $X_s^{P} = (X_s^{P1} - 5)/2 + X_s^{P2}/2$ and $X_s^{I} = (X_s^{I1} - 5)/2 + X_s^{I2}/2$, which establishes the above equalities. In the experimental task, they are a generic "procedure score" and "anesthesia score," respectively. The pre-registration document for Experiment 2 is available at `https://aspredicted.org/blind.php?x=9uq8dw`.

4.2. Results

4.2.1. Participants, Participant Responses, and Prediction Algorithm. MTurk workers who were located in the US, had a Human Intelligence Task (HIT) approval rate of 99% or higher, and had 100 or more HITs approved, were qualified to participate in the experiment. Participants who completed the experiment were paid $2 for participation. A total of 480 MTurk workers participated. Following the pre-registered exclusion criteria, we removed 174 individuals who did not complete the experiment and 54 participants who failed to correctly answer a four-question comprehension test on their first attempt. Among the 252 remaining participants, 42% were female, and 8% were 18 to 24 years old; 36%, 25 to 34; 31%, 35 to 44; 13%, 45 to 54; and 11%, 55 or over. Columns (1) and (2) of Table 1 provide the number of participants and average response in each condition. We developed prediction algorithms in the same way as Experiment 1 (see §3.2.2). Table D.2 in Appendix summarizes the prediction algorithms.

Table 1 Experiment 2: Summary of experiment results.

Information	Question	(1)	(2)	(3)	(4)	(5)	(6)
Type	Type	N	Response	Corr(X^I, response)	Corr(X^I, R)	RMSE of test set	DF-PIA gap
Baseline	DF	47	236.6 (30.7)	0.52 (0.21)	0.63 (0.25)	35.2 (13.9)	12.5***
	PIA	42	4.0 (19.1)	0.76 (0.23)	0.79 (0.22)	22.8 (11.4)	
Public info	DF	41	241.9 (24.5)	0.28 (0.18)	0.42 (0.26)	41.2 (11.2)	19.9***
complex	PIA	38	13.8 (33.9)	0.80 (0.27)	0.80 (0.27)	21.3 (14.0)	
Private info	DF	47	237.7 (31.6)	0.66 (0.15)	0.70 (0.15)	30.6 (8.8)	1.7
complex	PIA	37	48.8 (46.1)	0.72 (0.17)	0.73 (0.17)	28.9 (7.0)	

Note. Means and standard deviations (in parenthesis) are shown. X^P is public factor and X^I is private factor. In column (4), R is defined as the residual of response after regressing it on X^P. In column (6), DF-PIA gap is defined as the difference between the mean RMSEs of DF and PIA conditions. Column (6) also shows DF-PIA gap's statistical significance from a two-sample t-test for difference of means. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

4.2.2. Robustness of Hypothesis 1. Columns (5)-(6) of Table 1 summarize the prediction performance in each of the six conditions. Consistent with Hypothesis 1, the average RMSE of all PIA participants was 32% lower than the average RMSE of all DF participants (35.4 versus 24.2, $p < 0.0001$). As shown in column (6) of Table, 1, the DF-PIA gap was statistically significant at the 5% level in 2 of the 3 information conditions. In §3.2.5, we found that better performance is associated with higher correlation between participant response and X^I—after the effect of X^P in the responses are taken out. Columns (3) and (4) of Table 1 provide the average correlation between X^I and response and average correlation between X^I and R, defined as the residual of response after regressing it on X^P. As expected, the correlations are higher in the PIA conditions than in the DF conditions, which again provides mechanism evidence for Hypothesis 1.

4.2.3. Testing Hypothesis 2. Hypothesis 2(a) predicts the DF-PIA gap to be greater under public-information-complex conditions than under baseline conditions. Consistent with this hypothesis, the gap was 19.9 under the public-information-complex conditions and 12.5 under the baseline conditions. This 7.4 difference was statistically significant ($p = 0.036$; see Table D.4).

Hypothesis 2(b) predicts the DF-PIA gap to be smaller under private-information-complex conditions than under baseline conditions. Consistent with this hypothesis, the gap was 1.7 under the private-information-complex conditions and 12.5 under the baseline conditions. This 10.8 difference was statistically significant ($p = 0.002$; see Table D.4).

Hypothesis 2(c) predicts the DF-PIA gap to be greater under public-information-complex conditions than under private-information-complex conditions. Consistent with this hypothesis, the gap was 19.9 under public-information-complex conditions and 1.7 under private-information-complex conditions. This 18.2 difference was statistically significant ($p < 0.001$; see Table D.4). These findings provide evidence that the benefit of PIA over DF is greater when public information is complex, more so than when private information is complex.

4.3. Discussion

In addition to replicating Hypothesis 1 under different simulation parameters and information variables, Experiment 2 provided evidence that the location of random error matters in a manner consistent with Hypothesis 2. Specifically, eliciting human judgment via PIA instead of DF is most helpful when there is random error incorporating public information, but less helpful when there is random error incorporating private information.

5. Experiment 3: Numeric vs. Multiple-Choice Responses

Because we use human judgments as *inputs* to algorithms, responses need not be in the same domain as the ultimate forecast (in our case, minutes). Any linear transformation of DF and PIA responses defined in (5) are equivalent from the algorithm's perspective. This observation is practically relevant because it broadens the range of possible formats that the system designer can use to elicit the PIA. A designer may search this broader range of possibilities to try to reduce human random error, lower the effort required, and enhance PIA's performance.

The purpose of Experiment 3 is therefore to test whether Hypothesis 1 is robust even when the PIA question is formatted as a non-numeric multiple choice question. We further conjecture that structuring the PIA question in a way that reduces the cognition required to translate the private information into the domain requested can potentially reduce random error incorporating the private information, leading to enhanced PIA performance.

5.1. Experimental Design

The experiment was similar to Experiment 1, but with the additional manipulation of a 5-point scale multiple-choice question format for both DF and PIA. Thus, the experiment had a 2 answer types (numeric or multiple choice) by 2 question types (DF or PIA) between-subject design.

Participants assigned to the Multiple choice-DF condition were asked: "I think this surgery duration will be: a) a lot shorter than average, b) a little shorter than average, c) about average, d) a little longer than average, e) a lot longer than average." On the other hand, participants assigned to the Multiple choice-PIA condition were asked: "Of the above two characteristics of this future surgery, the hospital system only knows about the number of procedures. Compared to surgeries with the same number of procedures, I would advise the hospital system that this surgery duration will be: a) a lot shorter than average, b) a little shorter than average, c) about average, d) a little longer than average, e) a lot longer than average."

We conjectured that even though the multiple-choice format can convey less precise information than the numeric format, it may be cognitively easier (thereby reducing random error) for participants to translate the private information (a score between -5 and 5) to this multiple choice format as opposed to attempt to convert to minutes in the numeric format. The pre-registration document for Experiment 3 is available at `https://aspredicted.org/blind.php?x=mx2if3`. [2]

5.2. Results

5.2.1. Participants, Participant Response, and Prediction Algorithms. The MTurk worker qualification and payment settings were the same as in Experiment 2 (see §4.2.1). A total of 200 MTurk workers participated. Following the pre-registered exclusion criteria, we removed 30 individuals who did not complete the experiment. Among the 170 remaining participants, 45% were female, 9% were 18 to 24 years old; 45%, 25 to 34; 30%, 35 to 44; 13%, 9 to 54; and 6%, 55 or over. Columns (1) and (2) of Table 2 provide the number of participants and the average response in each condition.

We developed prediction algorithms in the same way as Experiment 1 (see §3.2.2). As was pre-registered, we used the first 20 rounds of each participant to develop prediction algorithms and the last 20 rounds to evaluate prediction performance. Table D.3 in the Appendix summarizes the prediction algorithms.

5.2.2. Performance Comparison. Columns (5)-(6) of Table 2 summarize the prediction performance of the four conditions. Consistent with Hypothesis 1, the results show robust support for the benefit of eliciting human judgment via PIA over DF. The DF-PIA gap was 5.0 ($p = 0.0121$) in numeric conditions and 8.6 ($p = 0.0001$) in multiple-choice conditions. The DF-PIA gap was directionally larger under multiple

[2] There were a few other minor changes compared to Experiment 1. For example, each participant observed 20 historical surgeries instead of 30, each participant completed 40 rounds instead of 50, and all participants observed the same sequence of surgeries.

Table 2 Experiment 3: Summary of experiment results.

Answer Type	Question Type	(1) N	(2) Response	(3) Corr(X^I, response)	(4) Corr(X^I, R)	(5) RMSE of test set	(6) DF-PIA gap
Numeric	DF	46	159.9 (30.1)	0.30 (0.18)	0.52 (0.26)	28.4 (5.4)	5.0*
	PIA	32	26.7 (29.2)	0.56 (0.35)	0.62 (0.33)	23.4 (11.6)	
Multiple choice	DF	44	11%, 15%, 24%, 28%, 21%	0.34 (0.27)	0.48 (0.31)	27.5 (6.2)	8.6**
	PIA	48	17%, 15%, 21%, 28%, 19%	0.64 (0.37)	0.71 (0.35)	18.9 (12.5)	

Note. Means and standard deviations (in parenthesis) are shown. For multiple choice conditions, the average percentage of each response is provided in the following order: 'a lot shorter than average'; 'a little shorter than average'; 'about average'; 'a little longer than average'; and 'a lot longer than average.' We coded 'a lot shorter than average' as 1, 'a little shorter than average' as 2, 'about average' as 3, 'a little longer than average' as 4, and 'a lot longer than average' as 5 for correlation computation and prediction algorithm development. X^P is public factor and X^I is private factor. In column (4), R is defined as the residual of response after regressing it on X^P. In column (6), DF-PIA gap is defined as the difference between the mean RMSEs of DF and PIA conditions. Column (6) also shows DF-PIA gap's statistical significance from a two-sample t-test for difference of means. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

choice, but the difference was not statistically significant ($p = 0.2152$; see Table D.5). In addition, columns (3)-(4) of Table 2 show the average correlation between X^I and the response and the average correlation between X^I and R, respectively. The higher correlation values between X^I and R in PIA conditions show, once more, strong mechanism evidence for Hypothesis 1.

The average RMSEs in numeric versus multiple choice conditions were not statistically different within DF conditions (28.4 vs 27.5, $p = 0.5016$). Within the PIA conditions, RMSEs were directionally better under multiple choice vs. numeric, although the difference was not statistically significant at the 5% level (23.4 vs 18.9, $p = 0.1138$). The RMSEs in the best performing condition, Multiple choice-PIA, were significantly better than the worst performing condition, Numeric-DF (18.9 vs 28.4, $p < 0.0001$).

5.3. Completion Times

The multiple-choice questions were less time-consuming for participants than the numeric questions. The mean of each participant's average time to complete one round was 16.1 seconds (SD 9.5) for numeric conditions, which is significantly greater ($p < 0.0001$) than that for multiple choice conditions (Mean 10.8 seconds and SD 7.0). Within numeric conditions, there was no statistical difference in the mean of average time to complete one round between DF and PIA conditions (15.3 seconds vs. 17.1 seconds; $p = 0.4155$). Similarly, within multiple choice conditions, there was no statistical difference in the mean of average time to complete one round between DF and PIA conditions (10.4 seconds vs. 11.1 seconds; $p = 0.6410$).

5.4. Discussion

Experiment 3 shows that the benefit of PIA over DF persists when using a 5-point multiple-choice scale response that was not in the same domain as the ultimate predictions and was also less time-consuming to answer. In this experiment, the multiple choice format also directionally outperformed the numerical format. We do *not* interpret this pattern as evidence of the general superiority of the multiple choice format.

Rather, we conclude that a fruitful strategy to enhance PIA's performance is to investigate the type of private information people have and how they store it in their memory, then format the question to make it cognitively easier for the participant to translate from their private information to the question's domain.

6. Conclusion

Our theoretical and experimental results suggest that there is opportunity to substantially improve prediction algorithm performance by applying a new private information adjustment (PIA) elicitation method to incorporate human judgment, instead of the traditional method of using direct forecast (DF). We formalize how judgmental random error drives a difference between the performance of prediction models using DF versus PIA (DF-PIA gap). We experimentally manipulate random error magnitude by varying the number of factors subjects must aggregate to show that random error incorporating public information increases the gap but random error incorporating private information decreases it. Finally, we demonstrate that the flexibility to write the PIA question in a domain different from the forecast's domain can be leveraged to potentially further enhance its predictive performance.

Perhaps the most natural next question our results help raise is whether and to what extent the PIA method leads to improvement in the field. Nevertheless, our exploration of the topic in this paper also highlights the importance of addressing a broader set of research questions: How does the DF-PIA gap behave for nonlinear machine learning algorithms? Beyond the number of factors, what other features of the environment induce greater random error? What is the best way to write the PIA question for common important forecasting contexts? How should system designers cope with situations where calibration data is sparse? Are there algorithm aversion or incentive issues that need to be addressed before implementation? We believe that, in addition to field work, laboratory experiments and behavioral models can help answer these questions that are all important for driving improvement in practice.

References

Arvan M, Fahimnia B, Reisi M, Siemsen E (2019) Integrating human judgement into quantitative forecasting methods: A review. *Omega* 86:237–252.

Blattberg RC, Hoch SJ (1990) Database models and managerial intuition: 50% model+ 50% manager. *Management Science* 36(8):887–899.

Bunn D, Wright G (1991) Interaction of judgemental and statistical forecasting methods: Issues & analysis. *Management science* 37(5):501–518.

Carbone R, Andersen A, Corriveau Y, Corson PP (1983) Comparing for different time series methods the value of technical expertise individualized analysis, and judgmental adjustment. *Management Science* 29(5):559–566.

Childers CP, Maggard-Gibbons M (2018) Understanding costs of care in the operating room. *JAMA Surgery* 153(4):e176233–e176233.

Davis AM, Katok E, Kwasnica AM (2014) Should sellers prefer auctions? A laboratory comparison of auctions and sequential mechanisms 60(4):990–1008.

Fildes R, Goodwin P, Lawrence M, Nikolopoulos K (2009) Effective forecasting and judgmental adjustments: An empirical evaluation and strategies for improvement in supply-chain planning. *International journal of forecasting* 25(1):3–23.

Goodwin P (2000) Correct or combine? Mechanically integrating judgmental forecasts with statistical methods. *International Journal of Forecasting* 16(2):261–275.

Hendriks A (2012) SoPHIE - Software platform for human interaction experiments. *Working paper, University of Osnabrueck* .

Herzog SM, Hertwig R (2009) The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psychological Science* 20(2):231–237.

Ho TH, Zhang J (2008) Designing pricing contracts for boundedly rational customers: Does the framing of the fixed fee matter? *Management Science* 54(4):686–700.

Huang T, Allon G, Bassamboo A (2013) Bounded rationality in service systems. *Manufacturing & Service Operations Management* 15(2):263–279.

Ibrahim R, Kim SH (2019) Is expert input valuable? The case of predicting surgery duration. *Seoul Journal of Business* 25(2).

Kahneman D, Rosenfield A, Gandhi L, Blaser T (2016) Noise: How to overcome the high, hidden cost of inconsistent decision making. *Harvard Business Review* 36–43.

Kremer M, Siemsen E, Thomas DJ (2016) The sum and its parts: Judgmental hierarchical forecasting. *Management Science* 62(9):2745–2764.

Lawrence MJ, Edmundson RH, O'Connor MJ (1986) The accuracy of combining judgemental and statistical forecasts. *Management Science* 32(12):1521–1532.

Lim JS, O'Connor M (1995) Judgemental adjustment of initial forecasts: Its effectiveness and biases. *Journal of Behavioral Decision Making* 8(3):149–168.

Mannes AE, Larrick RP, Soll JB (2012) The social psychology of the wisdom of crowds. Krueger JI, ed., *Frontiers of social psychology. Social judgment and decision making*, 227–242 (Psychology Press).

Palley AB, Soll JB (2019) Extracting the wisdom of crowds when information is shared. *Management Science* 65(5):2291–2309.

Rothschild M, Stiglitz JE (1970) Increasing risk: I. A definition. *Journal of Economic Theory* 2(3):225–243.

Sanders NR, Ritzman LP (2001) Judgmental adjustment of statistical forecasts. *Principles of Forecasting*, 405–416 (Springer).

Su X (2008) Bounded rationality in newsvendor models. *Manufacturing & Service Operations Management* 10(4):566–589.

Surowiecki J (2005) *The wisdom of crowds* (Anchor).

Vul E, Pashler H (2008) Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science* 19(7):645–647.

Appendix A: Proofs of Propositions

A.1. Proof of Proposition 1

Proof. To find the best fitting linear model of Y given X_i for $i \in \mathcal{P}$ and DF^*, we need to solve the following optimization problem:

$$\min_{\alpha_0,(\alpha_i)_{i\in\mathcal{P}},\beta_{DF^*}} \mathbb{E}[(Y-\alpha_0-\sum_{i\in\mathcal{P}}\alpha_i X_i-\beta_{DF^*}DF^*)^2]. \tag{8}$$

Using the definition of DF^* in (2), (8) is equivalent to

$$\min_{\alpha_0,(\alpha_i)_{i\in\mathcal{P}},\beta_{DF^*}} \mathbb{E}[(Y-(\alpha_0+\beta_{DF^*}v^*)-\sum_{i\in\mathcal{P}}(\alpha_i+\beta_{DF^*}w_i^*)X_i-\sum_{i\in\mathcal{I}}\beta_{DF^*}w_i^*X_i)^2]. \tag{9}$$

Similarly, to find the best fitting linear model of Y given X_i for $i \in \mathcal{P}$ and PIA^*, we need to solve the following optimization problem:

$$\min_{\gamma_0,(\gamma_i)_{i\in\mathcal{P}},\beta_{PIA^*}} \mathbb{E}[(Y-\gamma_0-\sum_{i\in\mathcal{P}}\gamma_i X_i-\beta_{PIA^*}PIA^*)^2]. \tag{10}$$

Using the definition of PIA^* in (2), (10) is equivalent to

$$\min_{\gamma_0,(\gamma_i)_{i\in\mathcal{P}},\beta_{PIA^*}} \mathbb{E}[(Y-\gamma_0-\sum_{i\in\mathcal{P}}\gamma_i X_i-\sum_{i\in\mathcal{I}}\beta_{PIA^*}w_i^*X_i)^2]. \tag{11}$$

We note that by letting

$$\alpha_0+\beta_{DF^*}v^*=\gamma_0,\ \ \alpha_i+\beta_{DF^*}w_i^*=\gamma_i \ \text{ for } i\in\mathcal{P},\ \text{ and } \ \beta_{DF^*}=\beta_{PIA^*},$$

there is a one-to-one correspondence between all feasible solutions for problems (9) and (11). That is, the two optimization problems are equivalent, and M_{DF^*} and M_{PIA^*} must yield the same predictions. ∎

A.2. Proof of Proposition 2

Proof. To find the best fitting linear model of Y given X_i for $i \in \mathcal{P}$ and DF^b, we need to solve the following optimization problem:

$$\min_{\alpha_0,(\alpha_i)_{i\in\mathcal{P}},\beta_{DF^b}} \mathbb{E}[(Y-(\alpha_0+\beta_{DF^b}v^b)-\sum_{i\in\mathcal{P}}(\alpha_i+\beta_{DF^b}W_i^b)X_i-\sum_{i\in\mathcal{I}}\beta_{DF^b}W_i^bX_i)^2] \tag{12}$$

Similarly, to find the best fitting linear model of Y given X_i for $i \in \mathcal{P}$ and PIA^b, we need to solve the following optimization problem:

$$\min_{\gamma_0,(\gamma_i)_{i\in\mathcal{P}},\beta_{PIA^b}} \mathbb{E}[(Y-\gamma_0-\sum_{i\in\mathcal{P}}\gamma_i X_i-\sum_{i\in\mathcal{I}}\beta_{PIA^b}W_i^bX_i)^2]. \tag{13}$$

We now show that for any feasible solution $(\bar{\alpha}_0,(\bar{\alpha}_i)_{i\in\mathcal{P}},\bar{\beta}_{DF^b})$ for problem (12), we can construct a feasible solution $(\bar{\gamma}_0,(\bar{\gamma}_i)_{i\in\mathcal{P}},\bar{\beta}_{PIA^b})$ for (13) that yields a *strictly* smaller objective value by letting:

$$\bar{\gamma}_0=\bar{\alpha}_0+\bar{\beta}_{DF^b}v^b,\ \ \bar{\gamma}_i=\bar{\alpha}_i+\bar{\beta}_{DF^b}\bar{w}_i^b \ \text{ for } i\in\mathcal{P},\ \text{ and } \ \bar{\beta}_{PIA^b}=\bar{\beta}_{DF^b}, \tag{14}$$

where we recall that $\mathbb{E}[W_i^b]=\bar{w}_i^b$.

Note that the objective value for problem (12) with $(\bar{\alpha}_0,(\bar{\alpha}_i)_{i\in\mathcal{P}},\bar{\beta}_{DF^b})$ is:

$$\mathbb{E}[(Y-(\bar{\alpha}_0+\bar{\beta}_{DF^b}v^b)-\sum_{i\in\mathcal{P}}(\bar{\alpha}_i+\bar{\beta}_{DF^b}W_i^b)X_i-\sum_{i\in\mathcal{I}}\bar{\beta}_{DF^b}W_i^bX_i)^2] \tag{15}$$

We define V and U as follows:

$$V = Y - \bar{\gamma}_0 - \sum_{i\in\mathcal{P}} \bar{\gamma}_i X_i - \sum_{i\in\mathcal{I}} \bar{\beta}_{PIA^b} W_i^b X_i \text{ and } U = \sum_{i\in\mathcal{P}} (\bar{\alpha}_i + \bar{\beta}_{DF^b} W_i^b) X_i - \sum_{i\in\mathcal{P}} (\bar{\alpha}_i + \bar{\beta}_{DF^b} \bar{w}_i^b) X_i.$$

Then, utilizing (14) and the definitions of V and U, (15) can be written as follows:

$$\begin{aligned}
&\mathbb{E}\left[\left(Y - (\bar{\alpha}_0 + \bar{\beta}_{DF^b} v^b) - \sum_{i\in\mathcal{P}} (\bar{\alpha}_i + \bar{\beta}_{DF^b} W_i^b) X_i - \sum_{i\in\mathcal{I}} \bar{\beta}_{DF^b} W_i^b X_i + \sum_{i\in\mathcal{P}} (\bar{\alpha}_i + \bar{\beta}_{DF^b} \bar{w}_i^b) X_i - \sum_{i\in\mathcal{P}} (\bar{\alpha}_i + \bar{\beta}_{DF^b} \bar{w}_i^b) X_i\right)^2\right] \\
&= \mathbb{E}\left[\left((Y - \bar{\gamma}_0 - \sum_{i\in\mathcal{P}} \bar{\gamma}_i X_i - \sum_{i\in\mathcal{I}} \bar{\beta}_{PIA^b} W_i^b X_i) - (\sum_{i\in\mathcal{P}} (\bar{\alpha}_i + \bar{\beta}_{DF^b} W_i^b) X_i + \sum_{i\in\mathcal{P}} (\bar{\alpha}_i - \bar{\beta}_{DF^b} \bar{w}_i^b) X_i)\right)^2\right] \\
&= \mathbb{E}[(V-U)^2] \\
&= \mathbb{E}[V^2] + \mathbb{E}[U^2] - 2\mathbb{E}[VU] \\
&= \mathbb{E}[V^2] + \mathbb{E}[U^2] \\
&> \mathbb{E}[V^2].
\end{aligned}$$

Note that in the fifth equation we use:

$$\begin{aligned}
\mathbb{E}[VU] &= \mathbb{E}[\mathbb{E}[VU | X_i \in \mathcal{P}]] \\
&= \mathbb{E}[\mathbb{E}[V | X_i \in \mathcal{P}] \mathbb{E}[U | X_i \in \mathcal{P}]] \\
&= 0,
\end{aligned}$$

since $\mathbb{E}[U|X_i, i \in \mathcal{P}] = 0$ and U and V are conditionally independent on $X_i \in \mathcal{P}$. Also note that $\mathbb{E}[U^2] = \text{Var}[U] > 0$ since $\text{Var}[W_i^b] > 0$.

Notice that $\mathbb{E}[V^2]$ is the objective value for problem (13) with $(\bar{\gamma}_0, (\bar{\gamma}_i)_{i\in\mathcal{P}\cup\mathcal{I}}, \bar{\beta}_{PIA^b})$. The reasoning above holds for any feasible solution for problem (12). In particular, it holds for the optimal solution at the optimal value. Thus, the mean squared error (MSE) for predictions under Model-DF^b is strictly larger than that under Model-PIA^b. ∎

A.3. Proof of Proposition 3

Proof. Recall that we defined $\widetilde{W}_i^b$ to be a mean preserving spread of W_i^b. By the definition of mean preserving spread (Rothschild and Stiglitz 1970), we can let $\widetilde{W}_i^b = W_i^b + \Gamma_i$, where we assume $\mathbb{E}[\Gamma_i] = 0$ and $\text{Var}[\Gamma_i] > 0$.

Note that when $\widetilde{W}_i^b$ is used in (5) for $i \in \mathcal{P}$ instead of W_i^b, Model-PIA^b does not change because PIA responses are unaffected by the random error incorporating public factors. Thus, we only need to compare the performance of Model-DF^b when $\widetilde{W}_i^b$ is used in (5) for $i \in \mathcal{P}$ versus W_i^b.

To find the best fitting linear model of Y given X_i for $i \in \mathcal{P}$ and DF^b with $\widetilde{W}_i^b$ for $i \in \mathcal{P}$, we need to solve the following optimization problem:

$$\min_{\widetilde{\alpha}_0, (\widetilde{\alpha}_i)_{i\in\mathcal{P}}, \widetilde{\beta}_{DF^b}} \mathbb{E}[(Y - (\widetilde{\alpha}_0 + \widetilde{\beta}_{DF^b} v^b) - \sum_{i\in\mathcal{P}} (\widetilde{\alpha}_i + \widetilde{\beta}_{DF^b} (W_i^b + \Gamma_i)) X_i - \sum_{i\in\mathcal{I}} \widetilde{\beta}_{DF^b} W_i^b X_i)^2] \tag{16}$$

Similarly, to find the best fitting linear model of Y given X_i for $i \in \mathcal{P}$ and DF^b with W_i^b for $i \in \mathcal{P}$, we need to solve the following optimization problem:

$$\min_{\alpha_0, (\alpha_i)_{i\in\mathcal{P}}, \beta_{DF^b}} \mathbb{E}[(Y - (\alpha_0 + \beta_{DF^b} v^b) - \sum_{i\in\mathcal{P}} (\alpha_i + \beta_{DF^b} (W_i^b)) X_i - \sum_{i\in\mathcal{I}} \beta_{DF^b} W_i^b X_i)^2] \tag{17}$$

We now show that for any feasible solution $(\widetilde{\alpha}_0, (\widetilde{\alpha}_i)_{i\in\mathcal{P}}, \widetilde{\beta}_{DF^b})$ for problem (16), we can construct a feasible solution $(\alpha_0, (\alpha_i)_{i\in\mathcal{P}}, \beta_{DF^b})$ for (17) that yields a *strictly* smaller objective value by letting:

$$\widetilde{\alpha}_0 + \widetilde{\beta}_{DF^b} v^b = \alpha_0 + \beta_{DF^b} v^b, \;\; \widetilde{\alpha}_i = \alpha_i \;\; \text{for } i \in \mathcal{P}, \;\; \text{and} \;\; \widetilde{\beta}_{DF^b} = \beta_{DF^b}. \tag{18}$$

Let $V = Y - (\alpha_0 + \beta_{DF^b} v^b) - \sum_{i\in\mathcal{P}} (\alpha_i + \beta_{DF^b} W_i^b) X_i - \sum_{i\in\mathcal{I}} \beta_{DF^b} W_i^b X_i$. Then, utilizing (18) and the definition of V, the objective value for problem (16) with $(\widetilde{\alpha}_0, (\widetilde{\alpha}_i)_{i\in\mathcal{P}}, \widetilde{\beta}_{DF^b})$ can be written as follows:

$$\begin{aligned}
&\mathbb{E}\left[\left(Y - (\widetilde{\alpha}_0 + \widetilde{\beta}_{DF^b} v^b) - \sum_{i\in\mathcal{P}} (\widetilde{\alpha}_i + \widetilde{\beta}_{DF^b}(W_i^b + \Gamma_i)) X_i - \sum_{i\in\mathcal{I}} \widetilde{\beta}_{DF^b} W_i^b X_i\right)^2\right] \\
&= \mathbb{E}\left[\left(Y - (\alpha_0 + \beta_{DF^b} v^b) - \sum_{i\in\mathcal{P}} (\alpha_i + \beta_{DF^b} W_i^b) X_i - \sum_{i\in\mathcal{I}} \beta_{DF^b} W_i^b X_i - \sum_{i\in\mathcal{P}} \beta_{DF^b} \Gamma_i X_i\right)^2\right] \\
&= \mathbb{E}[(V - \sum_{i\in\mathcal{P}} \beta_{DF^b} \Gamma_i X_i)^2] \\
&= \mathbb{E}[V^2] + \mathbb{E}[(\sum_{i\in\mathcal{P}} \beta_{DF^b} \Gamma_i X_i)^2] - 2\mathbb{E}[V \cdot \sum_{i\in\mathcal{P}} \beta_{DF^b} \Gamma_i X_i] \\
&> \mathbb{E}[V^2],
\end{aligned}$$

since

$$\begin{aligned}
\mathbb{E}[V \cdot \sum_{i\in\mathcal{P}} \beta_{DF^b} \Gamma_i X_i] &= \mathbb{E}[\mathbb{E}[V \sum_{i\in\mathcal{P}} \beta_{DF^b} \Gamma_i X_i \big| X_i \in \mathcal{P}]] \\
&= 0,
\end{aligned}$$

because of conditional independence on $X_i, i \in \mathcal{P}$, and the fact that $\mathbb{E}[\Gamma_i] = 0$. Note that $\mathbb{E}[V^2]$ is the objective value for problem (17) with $(\alpha_0, (\alpha_i)_{i\in\mathcal{P}}, \beta_{DF^b})$. ∎

A.4. Proof of Proposition 4

In the interest of algebraic tractability, we present here the proof for the case where there is one public factor and one private factor only. It is straightforward to generalize the proof for cases with more than one public factor and one private factor using the same approach.

Proof. We let X_1 be a public factor and X_2 be a private factor. By the definition of mean preserving spread (Rothschild and Stiglitz 1970), we can let $\widetilde{W}_i^b = W_i^b + \Gamma_i$, where we assume $\mathbb{E}[\Gamma_i] = 0$ and $\text{Var}[\Gamma_i] > 0$.

We first define a general optimization problem that finds the best fitting linear model of Y given X_1 and responses (DF or PIA):

$$Z(\gamma, W) = \min_{\alpha_0, \alpha_1, \beta} \mathbb{E}\big[Y - (\alpha_0 + \gamma\beta v^b) - (\alpha_1 + \gamma\beta W_1^b) X_1 - \beta W X_2\big]^2. \tag{19}$$

Here, α_0 is the calibrated intercept, α_1 is the coefficient for X_1, and β is the coefficient for response (DF or PIA). By choosing γ and W appropriately, we can define the following four models for different combinations of using DF versus PIA and using W_2^b versus $\widetilde{W}_2^b$:

$$\begin{aligned}
A = Z(1, W_2^b) &= \min_{\alpha_0, \alpha_1, \beta} \mathbb{E}\big[Y - (\alpha_0 + \beta v^b) - (\alpha_1 + \beta W_1^b) X_1 - \beta W_2^b X_2\big]^2, \\
B = Z(0, W_2^b) &= \min_{\alpha_0, \alpha_1, \beta} \mathbb{E}\big[Y - \alpha_0 - \alpha_1 X_1 - \beta W_2^b X_2\big]^2,
\end{aligned}$$

$$C = Z(1, \widetilde{W}_2^b) = \min_{\alpha_0,\alpha_1,\beta} \mathbb{E}\big[Y - (\alpha_0 + \beta v^b) - (\alpha_1 + \beta W_1^b)X_1 - \beta \widetilde{W}_2^b X_2\big]^2,$$
$$D = Z(0, \widetilde{W}_2^b) = \min_{\alpha_0,\alpha_1,\beta} \mathbb{E}\big[Y - \alpha_0 - \alpha_1 X_1 - \beta \widetilde{W}_2^b X_2\big]^2,$$

where $\widetilde{W}_2^b = W_2^b + \Gamma$ and $\mathbb{E}[\Gamma_i] = 0$ and $\text{Var}[\Gamma_i] > 0$. In general:

We assume that the actual surgery duration Y is defined as follows:

$$Y = \delta_0 + \delta_1 X_1 + \delta_2 X_2 + \varepsilon.$$

We can then show:

$$\begin{aligned}
Z(\gamma, W) =& \min_{\alpha_0,\alpha_1,\beta} \mathbb{E}\big[Y - (\alpha_0 + \gamma\beta v^b) - (\alpha_1 + \gamma\beta W_1^b)X_1 - \beta W X_2\big]^2 \\
=& \min_{\alpha_1,\beta} \text{Var}\big[Y - \gamma\beta v^b - (\alpha_1 + \gamma\beta W_1^b)X_1 - \beta W X_2\big] \\
=& \min_{\alpha_1,\beta} \text{Var}\big[Y - (\alpha_1 + \gamma\beta W_1^b)X_1 - \beta W X_2\big] \\
=& \min_{\alpha_1,\beta} \text{Var}\big[\varepsilon + (\delta_1 - \alpha_1 - \gamma\beta W_1^b)X_1 + (\delta_2 - \beta W)X_2\big] \\
=& \min_{\lambda_1,\beta} \text{Var}\big[\varepsilon + (\lambda_1 - \gamma\beta W_1^b)X_1 + (\delta_2 - \beta W)X_2\big] \\
=& \min_{\lambda_1,\beta} \text{Var}\big[\varepsilon + \lambda_1 X_1 - \gamma\beta W_1^b X_1 + \delta_2 X_2 - \beta W X_2\big] \\
=& \min_{\lambda_1,\beta} \text{Var}[\varepsilon] + \lambda_1^2 \text{Var}[X_1] + \gamma^2\beta^2 \text{Var}[W_1^b X_1] + \delta_2^2 \text{Var}[X_2] + \beta^2 \text{Var}[W X_2] \\
& -2\lambda_1\gamma\beta \text{Cov}(X_1, W_1^b X_1) - 2\delta_2\beta \text{Cov}(X_2, W X_2) \\
=& \text{Var}[\varepsilon] + \delta_2^2 \text{Var}[X_2] \\
& + \min_{\lambda_1,\beta} \Big[\lambda_1^2 \text{Var}[X_1] + \gamma^2\beta^2 \text{Var}[W_1^b X_1] + \beta^2 \text{Var}[W X_2] - 2\lambda_1\gamma\beta \text{Var}(X_1)\mathbb{E}[W_1^b] - 2\delta_2\beta \text{Cov}(X_2, W X_2)\Big] \\
=& \text{Var}[\varepsilon] + \delta_2^2 \text{Var}[X_2] \\
& + \min_{\beta} \Big[\min_{\lambda_1} \Big[\text{Var}[X_1]\lambda_1^2 - 2\gamma\beta \text{Var}(X_1)\mathbb{E}[W_1^b]\lambda_1\Big] + \beta^2(\gamma^2 \text{Var}[W_1^b X_1] + \text{Var}[W X_2]) - 2\delta_2\beta \text{Cov}(X_2, W X_2)\Big] \\
=& \text{Var}[\varepsilon] + \delta_2^2 \text{Var}[X_2] \\
& + \min_{\beta} \Big[\gamma^2\beta^2 \text{Var}[W_1^b X_1] - \gamma^2\beta^2 \mathbb{E}[W_1^b]^2 \text{Var}[X_1] + \beta^2 \text{Var}[W X_2] - 2\delta_2\beta \text{Var}(X_2)\mathbb{E}[W]\Big] \\
=& \text{Var}[\varepsilon] + \delta_2^2 \text{Var}[X_2] + \min_{\beta} \Big[\Big(\gamma^2 \text{Var}[W_1^b X_1] - \gamma^2 \mathbb{E}[W_1^b]^2 \text{Var}[X_1] + \text{Var}[W X_2]\Big)\beta^2 - 2\Big(\delta_2 \text{Var}(X_2)\mathbb{E}[W]\Big)\beta\Big] \\
=& \text{Var}[\varepsilon] + \delta_2^2 \text{Var}[X_2] - \frac{(\delta_2 \text{Var}(X_2)\mathbb{E}[W])^2}{\gamma^2 \text{Var}[W_1^b X_1] - \gamma^2 \mathbb{E}[W_1^b]^2 \text{Var}[X_1] + \text{Var}[W X_2]} \\
=& \text{Var}[\varepsilon] + \delta_2^2 \text{Var}[X_2] - \frac{(\delta_2 \text{Var}(X_2)\mathbb{E}[W])^2}{\gamma^2 \text{Var}[W_1^b]\mathbb{E}[X_1^2] + \text{Var}[W X_2]}.
\end{aligned}$$

Note that $\mathbb{E}[W_2^b] = \mathbb{E}[\widetilde{W}_2^b]$. Denote $M \equiv (\delta_2 \text{Var}[X_2]\mathbb{E}[W_2^b])^2 = (\delta_2 \text{Var}[X_2]\mathbb{E}[\widetilde{W}_2^b])^2$. Then, we obtain:

$$\begin{aligned}
(A - B) - (C - D) =& (-\frac{(\delta_2 \text{Var}(X_2)\mathbb{E}[W_2^b])^2}{\text{Var}[W_1^b]\mathbb{E}[X_1^2] + \text{Var}[W_2^b X_2]} + \frac{(\delta_2 \text{Var}(X_2)\mathbb{E}[W_2^b])^2}{\text{Var}[W_2^b X_2]}) \\
& - (-\frac{(\delta_2 \text{Var}(X_2)\mathbb{E}[\widetilde{W}_2^b])^2}{\text{Var}[W_1^b]\mathbb{E}[X_1^2] + \text{Var}[\widetilde{W}_2^b X_2]} + \frac{(\delta_2 \text{Var}(X_2)\mathbb{E}[\widetilde{W}_2^b])^2}{\text{Var}[\widetilde{W}_2^b X_2]}) \\
=& (-\frac{M}{\text{Var}[W_1^b]\mathbb{E}[X_1^2] + \text{Var}[W_2^b X_2]} + \frac{M}{\text{Var}[W_2^b X_2]}) \\
& - (-\frac{M}{\text{Var}[W_1^b]\mathbb{E}[X_1^2] + \text{Var}[\widetilde{W}_2^b X_2]} + \frac{M}{\text{Var}[\widetilde{W}_2^b X_2]}) \\
=& \frac{M \text{Var}[W_1^b]\mathbb{E}[X_1^2]}{(\text{Var}[W_1^b]\mathbb{E}[X_1^2] + \text{Var}[W_2^b X_2])(\text{Var}[W_2^b X_2])} - \frac{M \text{Var}[W_1^b]\mathbb{E}[X_1^2]}{(\text{Var}[W_1^b]\mathbb{E}[X_1^2] + \text{Var}[\widetilde{W}_2^b X_2])(\text{Var}[\widetilde{W}_2^b X_2])} \\
=& M \text{Var}[W_1^b]\mathbb{E}[X_1^2]\Big[\frac{1}{(\text{Var}[W_1^b]\mathbb{E}[X_1^2] + \text{Var}[W_2^b X_2])(\text{Var}[W_2^b X_2])}
\end{aligned}$$

$$
\begin{aligned}
&- \frac{1}{(\mathrm{Var}[W_1^b]\mathbb{E}[X_1^2] + \mathrm{Var}[\widetilde{W}_2^b X_2])(\mathrm{Var}[\widetilde{W}_2^b X_2])}\Bigg] \\
\geq &0,
\end{aligned}
$$

where the last inequality follows from $\mathrm{Var}[\widetilde{W}_2^b X_2] = \mathrm{Var}[(W_2^b + \Gamma)X_2] = \mathrm{Var}[W_2^b X_2] + \mathrm{Var}[\Gamma X_2] \geq \mathrm{Var}[W_2^b X_2]$. ■

Appendix B: Numerical Simulation Details

To construct Figure 1, we programmed a numerical simulation in R (script available from the authors upon request). As in Experiment 1 (see §3), we simulated true surgery durations according to $Y_s = 60 + 20X_s^P + 10X_s^I + \epsilon_s$, where X_s^P are uniform random numbers between 1 and 10 (inclusive), X_s^I are uniform random numbers between -5 and 5 (rounded to the tenths position), and ϵ_s are normally distributed with mean 0 and standard deviation 5.

We simulated the behavioral forecaster's coefficients, W_i^bs in (5), by adding normally distributed random error with mean 0 to the rational forecaster's coefficients, w_i^*s in (2). Specifically, to create the left figure, we varied the standard deviation of the normally distributed random error added to the behavioral forecaster's coefficient for the public factor, holding constant the standard deviation of the normally distributed random error added to the behavioral forecaster's coefficient for the private factor at 0.2. Similarly, for the right figure, we varied the standard deviation of the normally distributed random error added to the behavioral forecaster's coefficient for the private factor, holding constant the standard deviation of the normally distributed random error added to the behavioral forecaster's coefficient for the public factor at 0.2.

For each point in the figure, the script executed the following steps:

1. Simulate $10,000$ "actual" surgery durations.
2. Fit a linear model based on the simulated public and private factors. We used these coefficients to define the w_i^*s in (2) for a rational forecaster.
3. Calculate the rational forecaster's DF and PIA values for each surgery based on these w_i^*s.
4. Simulate the behavioral forecaster's coefficients W_i^bs in (5) for each surgery by adding normal random error with mean 0 to the w_i^*s.
5. Define the behavioral forecaster's DF and PIA for each surgery based on these random W_i^bs.
6. Split the dataset in half to define a train and test set.
7. Calibrate Model-DF^*, Model-PIA^*, Model-DF^b, and Model-PIA^b in (3), (4), (6), and (7), respectively, using the train set data.
8. Predict surgery durations onto the test set using these calibrated models.
9. Evaluate the RMSE for each model's predictions on test set.

Appendix C: Replicating Experiment 1 with MTurk Workers

The experiment was identical to Experiment 1 except for the study population; we recruited MTurk workers instead of university students. The pre-registration document for this experiment is available at `https://aspredicted.org/blind.php?x=yv2vs7`.

MTurk workers who were located in the US, had a Human Intelligence Task (HIT) approval rate of 99% or higher, and had 100 or more HITs approved were qualified to participate in the experiment. Participants who completed the experiment were paid $2 for participation. A total of 160 MTurk workers participated. Following the pre-registered exclusion criteria, we removed 37 individuals who did not complete the experiment, leaving 123 participants for analysis. Among the 123 participants, 47% were female and 9% were 18 to 24 years old; 43%, 25 to 34; 30%, 35 to 44; 11%, 45 to 54; and 7%, 55 or over.

Table C.1 provide a summary of experiment results and Table C.2 summarizes the prediction algorithms. The average RMSE was 26.0 in the DF condition and 22.5 in the PIA condition. The difference of 3.5 is significant ($p = 0.0085$) and represents a 13% decrease. As in Experiment 1, the correlation between X^I and R—defined as the residual of participant response after regressing it on X^P—was significantly higher in the PIA condition (0.61 versus 0.49, $p = 0.0191$).

Table C.1 Replicating Experiment 1 with MTurk Workers: Summary of experiment results.

	(1)	(2)	(3)	(4)	(5)	(6)
Condition	N	Response	Corr(X^I, response)	Corr(X^I, R)	RMSE of test set	DF-PIA gap
Direct Forecast (DF)	68	154.0 (37.3)	0.29 (0.22)	0.49 (0.24)	26.0 (5.8)	3.5**
Private Information Adjustment (PIA)	55	20.2 (35.3)	0.57 (0.33)	0.61 (0.30)	22.5 (8.7)	

Note. Means and standard deviations (in parenthesis) are shown. X^P is public factor and X^I is private factor. In column (4), R is defined as the residual of response after regressing it on X^P. In column (6), DF-PIA gap is defined as the difference between the mean RMSEs of DF and PIA. Column (6) also shows DF-PIA gap's statistical significance from a two-sample t-test for difference of means. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table C.2 Replicating Experiment 1 with MTurk workers: Prediction algorithms for each condition.

Condition	Coefficient for participant dummies	Coefficient for X^P interacted w/ participant dummies	Coefficient for response interacted w/ participant dummies
Direct Forecast (DF)	40.6 (20.6)	12.1 (7.0)	0.4 (0.3)
Private Information Adjustment (PIA)	66.5 (14.4)	17.8 (3.3)	0.5 (0.4)

Table C.3 provides detailed summary statistics comparing students and MTurk workers in Experiment 1 and this experiment. In addition, Table C.4 provides a more detailed comparison of RMSE. Although students and MTurk workers did not differ in their responses or in their time to provide the responses, the comparison results suggest that students were significantly better at our task than MTurk workers. An important takeaway, however, is that the benefit of PIA over DF questions was robust across the two study populations.

Table C.3 University students versus MTurk workers.

	DF			PIA		
	Student	**MTurk**	p-value	**Student**	**MTurk**	p-value
N	56	68		56	55	
Response	152.6 (30.8)	154.0 (37.3)	0.819	12.1 (23.7)	20.2 (35.3)	0.160
Corr(X^I,response)	0.39 (0.20)	0.29 (0.22)	0.018	0.74 (0.26)	0.57 (0.33)	0.003
Corr(X^I,R)	0.62 (0.20)	0.49 (0.24)	0.003	0.76 (0.24)	0.61 (0.30)	0.004
RMSE of test set	22.4 (6.3)	26.0 (5.8)	0.001	17.8 (7.6)	22.5 (8.7)	0.003
Average response time (seconds)	10.2 (6.4)	10.7 (8.3)	0.692	12.5 (9.3)	13.2 (7.5)	0.692

Note. Mean, standard deviation (in parentheses), and p-value from independent-samples t-test are reported.

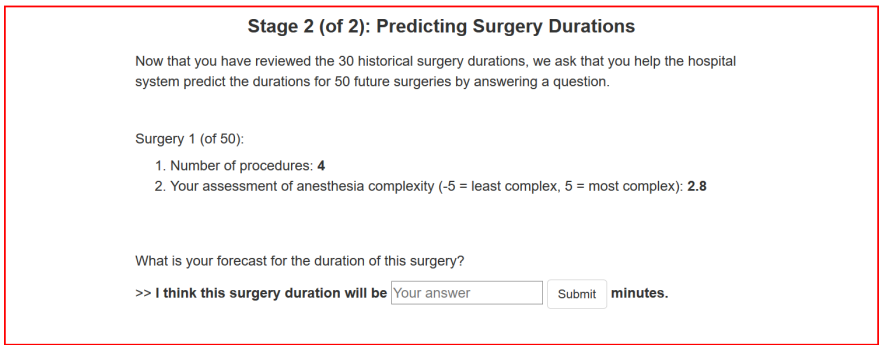
Table C.4 RMSE comparison of university students versus MTurk workers.

	(1)
	Root Mean Squared Error
Participant type (Base is Students)	
MTurk workers	3.67** (1.28)
Question type (Base is DF conditions)	
PIA conditions	-4.56*** (1.35)
Interaction effects (Base is Student × PIA condition)	
MTurk × PIA	1.04 (1.86)
Constant	22.36*** (0.95)
N	235
R^2	0.15

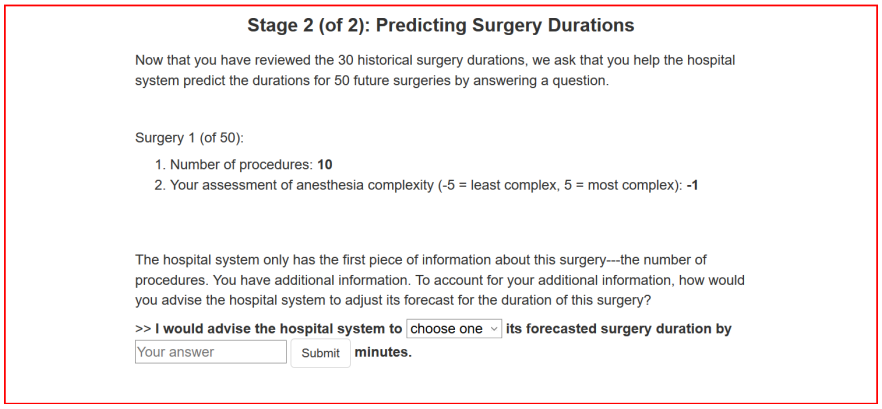
Note. Columns (1) is a linear regression model with RMSE as the dependent variable. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Appendix D: Additional Tables and Figures

Figure D.1 Experiment 1: Interface for the two conditions.



(a) Direct forecast (DF) condition



(b) Private information adjustment (PIA) condition

Figure D.2 Experiment 1 Instructions.

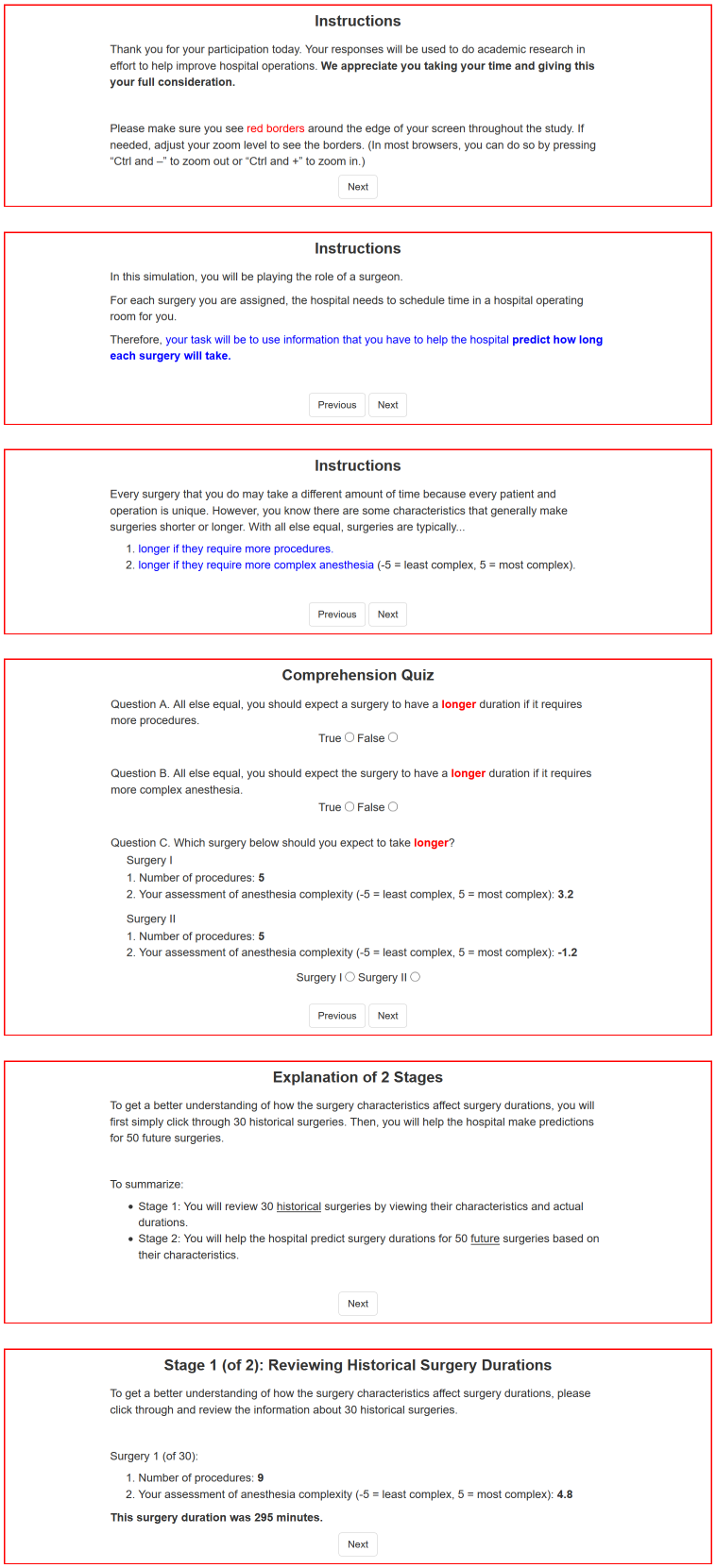


Table D.1 Experiment 1: Prediction algorithms for each condition.

Condition	Coefficient for participant dummies	Coefficient for X^P interacted w/ participant dummies	Coefficient for response interacted w/ participant dummies
Direct Forecast (DF)	34.5 (20.1)	10.8 (6.4)	0.5 (0.2)
Private Information Adjustment (PIA)	62.6 (10.5)	18.4 (2.8)	0.7 (0.4)

Table D.2 Experiment 2: Prediction algorithms for each condition.

Information type	Question type	Coefficient for participant dummies	Coefficient for X^P interacted w/ participant dummies	Coefficient for response interacted w/ participant dummies
Baseline	DF	110.6 (73.6)	10.9 (5.7)	0.6 (0.3)
	PIA	247.7 (5.8)	18.6 (6.4)	0.9 (0.7)
Public info complex	DF	77.9 (88.3)	3.0 (4.6), 4.7 (4.8)	0.7 (0.4)
	PIA	197.6 (9.6)	9.7 (1.8), 9.4 (1.9)	1.1 (0.7)
Private info complex	DF	108.1 (81.0)	14.5 (4.9)	0.6 (0.4)
	PIA	221.6 (20.2)	20.4 (4.4)	0.8 (0.6)

Table D.3 Experiment 3: Prediction algorithms for each condition.

Answer type	Question type	Coefficient for participant dummies	Coefficient for X^P interacted w/ participant dummies	Coefficient for response interacted w/ participant dummies
Numeric	DF	27.8 (20.4)	8.7 (10.0)	0.6 (0.4)
	PIA	67.2 (13.2)	16.9 (4.4)	0.7 (0.6)
Multiple choice	DF	20.0 (35.8)	15.2 (5.9)	19.5 (18.3)
	PIA	7.9 (32.8)	17.5 (5.4)	20.8 (15.9)

Table D.4 Experiment 2: Performance Comparison.

	(1)
	Root Mean Squared Error
Information type (Base is Baseline conditions)	
Public info complex conditions	5.96* (2.43)
Private info complex conditions	-4.64* (2.35)
Question type (Base is DF conditions)	
PIA conditions	-12.49*** (2.42)
Interaction effects (Base is Baseline × PIA condition)	
Public info complex × PIA	-7.42* (3.52)
Private info complex × PIA	10.77** (3.48)
Constant	35.25*** (1.66)
N	252
R^2	0.27

Note. Columns (1) is a linear regression model with RMSE as the dependent variable. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table D.5 Experiment 3: Performance comparison.

	(1)
	Root Mean Squared Error
Answer type (Base is Numeric conditions)	
Multiple choice conditions	-0.82 (1.97)
Question type (Base is DF conditions)	
PIA conditions	-5.01* (2.15)
Interaction effects (Base is Numeric × PIA condition)	
Multiple choice× PIA	-3.61 (2.90)
Constant	28.37*** (1.38)
N	170
R^2	0.15

Note. Columns (1) is a linear regression model with RMSE as the dependent variable. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.