informs
http://pubsonline.informs.org/journal/msom

MANUFACTURING & SERVICE OPERATIONS MANAGEMENT

Vol. 00, No. 0, Xxxxx 0000, pp. 000–000

Submitted to *Manufacturing & Service Operations Management*

ISSN 1523-4614, EISSN 1526-5498

Warnings and Endorsements: Improving Human-AI Collaboration Under Covariate Shift

Matthew DosSantos DiSorbo
Harvard Business School, mdisorbo@hbs.edu

Kris Johnson Ferreira
Harvard Business School, kferreira@hbs.edu

Abstract. 1. Problem definition: While artificial intelligence (AI) algorithms may perform well on data that are representative of the training set (inliers), they may err when extrapolating on non-representative data (outliers). These outliers often originate from covariate shift, where the joint distribution of input features changes from the training set to deployment. How can humans and algorithms work together to make better decisions when faced with outliers and inliers?

2. Methodology/results: We study a human-AI collaboration on prediction tasks using an anchor-and-adjust framework, and hypothesize that humans are biased towards naïve adjustment behavior: making adjustments to algorithmic predictions that are too similar across inliers and outliers, when ideally adjustments should be larger on outliers than inliers. In an online lab experiment, we demonstrate that participants are indeed unable to sufficiently differentiate absolute adjustments to an AI algorithm when faced with both inliers and outliers, leading to a 143-176% increase in their absolute deviation from the optimal prediction compared to participants who only face either all inliers or all outliers. We design a 'warning' that alerts participants when feature values constitute outliers and, in a second experiment, we show that this warning helps participants differentiate adjustments, ultimately reducing their absolute deviation from the optimal prediction by an average of 31% on outliers and 35% on inliers. We demonstrate that an additional intervention — 'endorsements' that alert participants when feature values constitute inliers — reduces participants' absolute deviation from the optimal prediction on inliers by an additional 34% on average.

3. Managerial implications: Our work uncovers a behavioral bias towards naïve adjustment behavior, and identifies a simple, educational intervention that mitigates this bias. Ultimately, we hope that this work will help managers best equip their employees with the knowledge they need to succeed in a human-AI collaboration.

Key words: human-AI collaboration, behavioral operations, experiments

1. Introduction

Organizations are making considerable investments in data-driven, artificial intelligence (AI) algorithms to support operational decision-making. Various algorithms have been leveraged across a number of settings, including in retail to predict demand and support price optimization (Ferreira et al. 2016), in advertising to improve efficiency (Qin and Jiang 2019) and in labor-scheduling decisions made by store managers (Kwon et al. 2022). NewVantagePartners (2023) finds that, according to executives with data leadership positions, 88% of the Fortune 1000 companies or organizations surveyed are increasing their investments in data and analytics. And yet, despite the rise of machine learning and other AI algorithms, obstacles remain in the path of seamless integration. For example, humans may be predisposed to avoid algorithms due to some underlying 'algorithm aversion' (Dietvorst et al. 2015). To tackle the challenge of employee resistance to AI algorithms, several recent studies have been conducted that aim to increase employees' adherence to algorithmic recommendations (e.g. Dietvorst et al. 2018, Sun et al. 2022, Caro and Saez de Tejada Cuenca 2023).

Although increasing employees' adherence may sometimes be beneficial, we note that in many contexts it is optimal for the employee to deviate from the algorithm's recommendation. To allow for such deviations, we consider the case of *human-AI collaboration* in a judge-advisor decision-making system (Sniezek and Buckley 1995). Specifically, an AI algorithm makes a recommendation in the form of a prediction, the employee views the algorithm's recommendation and then makes a final decision. Deviations from the algorithm's recommendation may be beneficial for many reasons. For example, the employee might have access to valuable predictive information that the algorithm does not take into account (Balakrishnan et al. 2022) or might be aware of operational constraints not considered by the algorithm (Sun et al. 2022). As another example — and our focus in this paper — employees' deviations might be beneficial in the presence of *outliers*.

We define an outlier as a vector of input features that is not representative of the training set and thus may be difficult for the algorithm to accurately predict. An important source of outlier generation is covariate shift, where the joint distribution of input features changes from training to testing set (Shimodaira 2000). Small changes, even in large datasets, can spell trouble for the deployment of AI algorithms. Unfortunately, it may be costly and time-consuming to frequently re-train the algorithm (Baier et al. 2019), and some algorithms require a substantial amount of new, shifted data before they can make reasonably accurate predictions. In periods between re-training, continuous monitoring and subjective human judgement can be critical in avoiding over-adherence to algorithmic predictions on outliers (Babic et al. 2021, Blyth 2018). Human domain

experts have contextual knowledge and intuition that can help them make beneficial adjustments to the algorithm's recommendations on outliers. However, allowing employees to deviate from the algorithm's recommendation does not necessarily guarantee that *all* deviations will be beneficial. Findings in Dietvorst et al. (2015) would suggest that if the employees observe an algorithm's poor performance on outliers, they may quickly also lose faith in the algorithm's competence on non-outliers (inliers), for which the algorithm performs well, leading to larger deviations and worse performance on inliers. How should we design the human-AI collaboration in such a way that helps the employee identify and correct for potentially poor algorithmic predictions on outliers without degrading performance on inliers?

We first introduce a mathematical model to describe how humans might use an AI algorithm's predictions in their decision-making. We employ an anchor-and-adjust model (Tversky and Kahneman 1974), which has been widely studied in the legal, forecasting, and negotiation literature (Furnham and Boo 2011). Put simply, humans first 'anchor' on a reference point — in our case, the algorithm's prediction — and then 'adjust' from this initial value to reach a final prediction. We hypothesize that, while humans may partially differentiate adjustments in the presence of both outliers and inliers, they are biased towards *naïve adjustment behavior*: making the same adjustment on every prediction. We show that this behavior is suboptimal, resulting in too-large absolute adjustments on inliers and too-small absolute adjustments on outliers.

We then conduct an online lab experiment to test our theory in the context of demand prediction, a common task across contexts like retail (Ren et al. 2020) and supply chain management (Zougagh et al. 2020). Further, many firms characterize their forecasts as based off of statistical software and then judgmentally adjusted by humans (Siemsen and Aloysius 2020). In our study, participants are given feature values that describe a product and an algorithm's demand prediction, and they are asked to predict product demand by directly adjusting, or accepting (not adjusting), the algorithm's predictions. The vector of feature values that describe a product can either constitute an inlier or an outlier, depending on whether the vector is representative of the training set used to develop the algorithm. As would be typical in practice, our algorithm performs well on inliers, but not on outliers. We directly manipulate the presence of outliers and inliers across participants: one condition contains only inliers, another only outliers, and another a mixture of inliers and outliers. We find that, when faced with the mixture of inliers and outliers, participants make larger (smaller) absolute adjustments to the algorithm's predictions for inliers (outliers) compared to the all-inlier (all-outlier) condition, confirming our theorized bias towards naïve adjustment behavior.

This bias results in significant performance degradation: participant absolute deviations from optimal predictions on inliers (outliers) are an average of 176% (143%) larger in the *Mixed* condition compared to the all-inlier (all-outlier) condition.

We develop design principles of the human-AI collaboration aimed at mitigating naïve adjustment behavior, and we test their effectiveness in a follow-on online lab experiment. Specifically, we introduce two interventions intended to improve performance: warnings (flagging outliers with a warning that a feature value is outside the range of the historical data and the algorithm may perform poorly), or warnings *and* endorsements (additionally flagging inliers with an endorsement that the feature values are within the ranges of the historical data and the algorithm is expected to perform well). 'Outlier focus' messages — similar in nature to our warnings-only intervention — have been studied in the context of nudging humans towards overriding algorithms on outliers (Poursabzi-Sangdeh et al. 2021) and preparing a human for the impact of covariate shift at the start of an experiment (Chiang and Yin 2021). Yet previous work on 'outlier focus' messages does not study the impact of encountering outliers on subsequent performance on inliers, an impact that is central to our work. Does calling attention to potential algorithm errors degrade overall trust of the algorithm, negatively impacting performance on inliers? The balance of this dual purpose — improving performance across a mixed exposure of inliers and outliers, which we believe best reflects practical applications — motivates us to additionally include an endorsement in our second intervention. It is unclear if the extra messages will have a beneficial effect: does the *lack* of warnings on inliers in our warnings-only intervention constitute a sufficient endorsement and/or avoid the potential of information overload? Note that endorsements are not a novel intervention: stating high levels of model accuracy, or stating that participants who closely adhere to the algorithm perform better, helps increase adherence (Yin et al. 2019, Snyder et al. 2023). However, once again, endorsements are not traditionally deployed alongside warnings. Participants in these studies are not also alerted about outliers, and it is unclear what impact the combination of warnings and endorsements will have on both outlier and inlier performance.

Replicating the mixture condition from our first study as a baseline, we find that outlier warnings do increase differentiation between adjustments on inliers and outliers, reducing the participants' absolute deviation from the optimal predictions by 35% for inliers and 31% for outliers, on average. Notably, we find that *also* endorsing inliers further reduces participants' absolute deviations from optimal predictions on inliers by an additional 34%, on average. This has consequences for the design of human-AI collaboration: while warnings can be useful, warnings *and* endorsements have an even more potent effect. We consider these as our main contributions:

• We define new theory that describes how humans use algorithmic recommendations in their decision-making using an anchor-and-adjust framework. We hypothesize that humans, while able to make partial differentiations between inliers and outliers, are ultimately — and significantly — biased towards *naïve adjustment behavior*, and we prove that this degrades task performance.

• We conduct an online lab experiment that supports our hypothesis that humans are biased toward naïve adjustment behavior: over- and under-adjusting on inliers and outliers, respectively. We demonstrate that this bias is costly, contributing to a statistically and economically significant increase in humans' absolute deviations from optimal predictions on both inliers and outliers.

• We introduce a warning on outliers (warning that a feature value is outside the range of the historical data and the algorithm may perform poorly) and show that it mitigates naïve adjustment behavior and improves human performance on both inliers and outliers. Furthermore, we demonstrate that an additional intervention — endorsements on inliers (endorsing that the feature values are within the ranges of the historical data and the algorithm is expected to perform well), in addition to warnings on outliers — improves performance on inliers even more.

2. Literature Review

Our work centers on judge-advisor systems in decision-making: an algorithm (the advisor) makes a recommendation in the form of a prediction, and then a human (the judge) considers the recommendation and freely chooses if and how to incorporate it into their final decision (Sniezek and Buckley 1995).

In many instances, human overrides of algorithmic recommendations have been found to hinder performance, and work has been done to study how to increase adherence to these recommendations in practice. For example, Kawaguchi (2021) found that vending machine managers did not, on average, follow a dynamic product assortment algorithm projected to increase revenue. Integrating worker forecasts into the algorithm increased adherence, even though the output of the algorithm was nearly the same as when worker forecasts were not integrated. In another example, when retail store managers did not adhere to a demand prediction and markdown algorithm that increased revenue, Caro and Saez de Tejada Cuenca (2023) found that adherence could be increased by displaying an interpretable metric so that managers could better understand and grow confidence in the tool's recommendations. Similar dynamics are explored in laboratory experiments. Seeing an algorithm err led participants to mistakenly avoid a superior algorithm (Dietvorst et al. 2015), and reliance rates on algorithmic recommendations were lower when a task was framed as more

subjective than objective (Castelo et al. 2019) or participants believed that the algorithm's recommendation process could not be understood as well as a human recommendation process (Yeomans et al. 2019). In such cases, allowing humans to modify algorithmic recommendations — even to a small degree — can increase adherence to the algorithm (Dietvorst et al. 2018).

In other instances, though, human overrides have been found to improve the performance of a faulty algorithm. Information unobservable to the algorithm can cause the algorithm to err. For example, Ibrahim et al. (2021) and Balakrishnan et al. (2022) found that humans can improve the algorithm's prediction using information that the algorithm does not have access to, but is valuable in predicting demand. Although overrides in Kesavan and Kushwaha (2020) reduced net profitability, they found that employee adjustments on the subset of growth-stage products — which traditionally have high levels of uncertainty — were beneficial. Another source of algorithm error includes technical mishaps. For example, De-Arteaga et al. (2020) studied an unfortunate scenario where a child maltreatment prediction algorithm glitched, and they found that the call workers who used the algorithm were able to mitigate the issue by correctly overriding the algorithm in some cases where it erred. We focus on a different source of algorithm error — outliers — and aim to identify if and how humans may be able to improve the algorithm's performance in the presence of outliers. We define outliers as instances where the vector of input features is not representative of the training set used to build the AI prediction algorithm and thus may be difficult for the algorithm to predict.

One major source of outlier generation — and the one we consider in this paper — is *covariate shift*, where the distribution of input feature vectors changes over time, but the conditional distribution of the outcomes given the inputs does not change (Shimodaira 2000). The machine learning literature proposes methods like importance weighted cross validation to enhance robustness to covariate shift (Sugiyama et al. 2007). Unfortunately, many methods require that the support of the new distribution of input feature vectors is contained within the support of the training distribution (Kouw and Loog 2019), which is often not satisfied. In addition, more effective methods are often computationally expensive (Ovadia et al. 2019). Thus, covariate shift remains a threat to algorithm usage and an area of active study (Bayram and Ahmed 2023). Rather than relying on algorithms alone to mitigate prediction errors on outliers, how might a human-AI collaboration be designed to better address the issue?

To the best of our knowledge, only two other papers — Poursabzi-Sangdeh et al. (2021) and Chiang and Yin (2021) — study human-AI collaboration in the presence of outliers and make

direct contributions to this important question. Each of these papers conduct online lab experiments where participants work with an algorithm to predict apartment prices. The main focus in Poursabzi-Sangdeh et al. (2021) is to study how varying levels of model size and transparency impact participants' use of the model. In a fourth and final experiment, the researchers found that participants who were given model transparency were less able to detect and correct for the model's poor predictions on outliers. 'Outlier focus' messages — where participants are told that the apartment has an unusual combination of bedrooms and bathrooms — helped to eliminate the negative impact of model transparency. Overall, these messages helped participants intuit how the unusual feature vectors would impact apartment price, leading them to deviate more on outliers.

Chiang and Yin (2021) ask participants to predict housing prices that have been beset by covariate shift. They test the impact of two interventions — explaining the possible impact of covariate shift at the start of the experiment, and visualizing input feature values at each prediction in relation to the training data — on algorithm reliance. Participants do not face a mixture of inliers and outliers; rather, for a given participant, data points are either all inliers or all outliers. Further, participants must decide at the beginning of each prediction task to either delegate the prediction to the algorithm or to make their own prediction. Once a participant chooses the latter, they do not have access to the algorithm for the remaining tasks. The resulting survival curves measure how long participants completely rely on the algorithm, which mimics how humans might initially authorize an algorithm to make decisions before later opting out. This type of collaboration is quite different from our work, where participants can make adjustments to, instead of completely relying on, the algorithm's prediction. Furthermore, in our work, and as is common in practice, participants do not lose future access to the algorithm if they decide to adjust its forecast. Chiang and Yin (2021) find that explaining the possible impact of covariate shift at the start of the experiment — but not visualizing input feature values at each prediction — helps decrease over-reliance on the algorithm for participants faced with outliers.

Both papers employ interventions that educate humans about outliers and successfully increase human deviation from the algorithm. We build on these initial promising findings by studying whether model mistakes on outliers erode trust on inliers, where the human should adhere to the model's accurate predictions. This is a salient question, as human-AI collaborations typically encounter a mixture of inliers and outliers. Further, we study not just the magnitude of the deviation from the algorithm's prediction but the magnitude of the deviation from the optimal prediction.

Building on Poursabzi-Sangdeh et al. (2021) and Chiang and Yin (2021), we develop an intervention that educates humans about the potential fallibility of the algorithm's outlier predictions. We then design an additional intervention that educates humans about the potential strength of the algorithm's inlier predictions. These interventions, at their core, convey the expected prediction accuracy of the algorithm to the human user. Communicating algorithm prediction accuracy is not novel: many researchers have shown that such displays can have potent effects on algorithm usage, and several ways of conveying algorithm prediction accuracy have been studied in the literature. These include stating the percent of correct predictions in the testing set for binary prediction algorithms (Lai and Tan 2019, Zhang et al. 2020, Yin et al. 2019), which increased adherence to the algorithm when accuracy was high; adherence was also increased when accuracy was conveyed by indicating other participants who followed the algorithm performed well (Snyder et al. 2023) or displaying the posterior distribution of the algorithm's prediction (McGrath et al. 2020). This increased adherence resulted in improved outcomes in Yin et al. (2019), Snyder et al. (2023) and McGrath et al. (2020), as the algorithm performed better than the vast majority of participants in these studies. However, Lai and Tan (2019) find that adherence also increased on *incorrect* AI predictions, and that the overall performance of the human-AI collaboration fell short of the AI acting alone. Zhang et al. (2020) find that stating a low algorithm accuracy did not meaningfully lower adherence in settings where the participant observes the algorithm's prediction; further, the interventions did not improve the overall prediction accuracy of the human-AI collaboration compared to the AI alone. Finally, these statements of accuracy do not include any mention or education about the impact of outliers, which Poursabzi-Sangdeh et al. (2021) and Chiang and Yin (2021) have shown to be effective in increasing human deviation from algorithmic predictions on outliers.

In our work, we consider the prevalent setting where humans encounter a mixture of inliers and outliers, and we aim to design interventions to differentially impact their adherence to the algorithm's recommendation in each case, i.e., both increasing adherence on inliers and decreasing adherence on outliers. Furthermore, we aim to design a human-AI collaboration that makes more accurate predictions than the AI alone in the presence of covariate shift.

3. Theory Development

In this section, we introduce a model that reflects how humans use an AI algorithm's recommendations to make their predictions, and we derive theory from this model to inform our hypotheses around the performance of human-AI collaboration in the presence of inliers and outliers.

3.1. Model Setting

Consider a setting where outcome Y_i is a function of feature vector $\boldsymbol{X}_i$ and i.i.d. random noise ϵ_i, where $\mathbb{E}[\epsilon_i] = 0$, for each instance i:

$$Y_i = f_{actual}(\boldsymbol{X}_i) + \epsilon_i. \tag{1}$$

We assume the $\boldsymbol{X}_i$ are independent, albeit not necessarily identically distributed, with domain $\mathcal{D}$.

A human decision-maker is asked to predict Y_i given a realization of the feature vector $\boldsymbol{x}_i$, but otherwise Equation 1 is unknown to the decision-maker. Note that the *optimal prediction* for Y_i is $f_{actual}(\boldsymbol{X}_i)$, but it is likely difficult for the human to perfectly recover f_{actual}. We consider a setting where the human has access to an algorithm, and they can freely choose if/how to use the algorithm to help make their prediction. The algorithm also predicts Y_i using feature vector $\boldsymbol{X}_i$:

$$\hat{Y_i}^{alg} = f_{alg}(\boldsymbol{X}_i). \tag{2}$$

Note that, given a realization of feature vector $\boldsymbol{x}_i$, the algorithm's prediction $\hat{y_i}^{alg}$ is deterministic.

The algorithm is developed ("fit") using a set of K historical "training" instances, $\mathcal{S} = \{(\boldsymbol{x}_k, y_k)\}_{k=1,\ldots,K}$, where $(\boldsymbol{x}_k, y_k)$ is a pair of the realized feature vector and outcome for instance k. Refer to Rokach et al. (2023) for an overview of how training data is used to fit an AI algorithm. Importantly, the set of $\{\boldsymbol{x}_k\}_{k=1,\ldots,K}$ included in $\mathcal{S}$ is often a strict subset of domain $\mathcal{D}$.

We can define a partition of $\mathcal{D}$ as the subsets $\mathcal{D}_I$ and $\mathcal{D}_O$ such that $\mathcal{D}_I$ denotes the set of all *inliers* relative to training set $\mathcal{S}$ and $\mathcal{D}_O$ denotes the set of all *outliers* relative to $\mathcal{S}$. We assume that $P(\boldsymbol{X}_i \in \mathcal{D}_I), P(\boldsymbol{X}_i \in \mathcal{D}_O) > 0$, which represents the interesting case where the domain consists of both inliers and outliers relative to training set $\mathcal{S}$. Numerous ways to classify $\boldsymbol{X}_i$ as an inlier vs. outlier have been proposed in the literature and vary in complexity; see Ben-Gal (2005) and Boukerche et al. (2020) for examples. Most definitions would characterize inliers as being data points that are representative of the training data set and outliers as being data points that are not representative of the training data set, e.g., due to covariate shift. We use the same characterization colloquially in our work. As we will later describe, our experimental conditions vary the extent to which we explicitly share information about partition $\{\{\mathcal{D}_I\}, \{\mathcal{D}_O\}\}$ with the human decision-maker.

Mathematically, we make the following assumption:

ASSUMPTION 1. $|\mathbb{E}[Y_i - \hat{Y}_i^{alg} | \boldsymbol{X}_i \in \mathcal{D}_I]| < |\mathbb{E}[Y_i - \hat{Y}_i^{alg} | \boldsymbol{X}_i \in \mathcal{D}_O]|$.

This assumption simply states that the algorithm's absolute expected error on inliers is less than the algorithm's absolute expected error on outliers, which is an implication of the commonly held assertion that outliers typically have larger errors than inliers (Aggarwal and Aggarwal 2017).

3.2. Anchor and Adjust Behavioral Model

To study how a human might use an AI algorithm's recommendation to inform their predictions, we follow a common heuristic in the decision-making literature known as 'anchor and adjust'. Tversky and Kahneman (1974) popularized this heuristic, suggesting that humans affix their decision to some initial value, perhaps from an external suggestion — the anchor — and then make an additive adjustment to arrive at a final decision.

In this vein, we hypothesize that humans anchor on the algorithm's prediction and then make an adjustment to arrive at a final prediction. We model human j's prediction for instance i as

$$\hat{Y}_{ij}^{final} = \hat{Y}_i^{alg} + f_{\text{adjust},j}(\boldsymbol{X}_i, \hat{Y}_i^{alg}) + \eta_{ij}, \tag{3}$$

where η_{ij} is an independent, zero-mean, bounded random noise reflecting the concept that humans are boundedly rational and make noisy predictions (Kahneman D 2022, Su 2008). For brevity, we omit the subscript j when the context is clear.

Note that how human j chooses to make an adjustment, $f_{\text{adjust},j}(\boldsymbol{X}_i, \hat{Y}_i^{alg})$, may be based on feature vector $\boldsymbol{X}_i$ and/or the AI's prediction $\hat{Y}_i^{alg}$. Clearly, human j's prediction accuracy depends on how they choose this adjustment function. We believe that humans tend to be biased towards *naïve adjustment behavior*, which we define as using the same constant adjustment regardless of the feature vector $\boldsymbol{X}_i$ and the AI's prediction $\hat{Y}_i^{alg}$; in other words, naïve adjustment behavior for human j is characterized by $f_{\text{adjust},j}(\boldsymbol{X}_i, \hat{Y}_i^{alg}) = \delta_j$ across each instance i. Naïve adjustment behavior is suboptimal because it does not differentiate the adjustments based on characteristics of a specific instance i or expected differences in the algorithm's prediction error across instances, e.g., for inliers vs. outliers. We formally study this suboptimality in Section 3.3. Our theoretical results will guide our experimental designs and hypotheses, which we preview in Section 3.4.

3.3. Suboptimality of Naïve Adjustment Behavior

Consider someone who makes naïve adjustments $f_{\text{adjust}}(\boldsymbol{X}_i, \hat{Y}_i^{alg}) = \delta \; \forall i$, but who chooses their constant adjustment δ in an optimal way. This is described by the following optimization problem:

$$NAdj: \min_{\delta \in \mathbb{R}} \mathbb{E}[\big(Y_i - (\hat{Y}_i^{alg} + \delta)\big)^2]. \tag{4}$$

$NAdj$ finds the constant adjustment δ^{NAdj} that minimizes the expected squared error of the human's prediction, $\hat{Y}_i^{final}$.

PROPOSITION 1. *Under naïve adjustment behavior, the optimal adjustment δ^{NAdj} is given by*

$$\delta^{NAdj} = \mathbb{E}[Y_i - \hat{Y}_i^{alg}].$$

All proofs are included in Appendix A.

Although we consider random variables here, in practice the human could approximate δ^{NAdj} given any $i = 1, ..., n$ data points including $(y_i, \hat{y}_i^{alg})$ with the estimator $\hat{\delta}^{NAdj} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i^{alg})$. The value of δ^{NAdj} is intuitive: it is the average difference between the outcome and the algorithm's prediction. However, although it is intuitive, naïve adjustment behavior can be suboptimal. The adjustment is constant across all instances, but the expected error of the algorithm is likely not. In particular, for our model setting and per Assumption 1, the absolute expected error of outliers is greater than that of inliers; thus, one would naturally expect a strategy where larger absolute adjustments are made on outlier predictions vs. inlier predictions to be superior to naïve adjustment behavior where an identical adjustment is applied for inliers and outliers.

We formalize this idea by defining *M-differential adjustment behavior* as partitioning instances into $m = 1, ..., M$ groups with different absolute expected algorithm error and using the same constant adjustment δ_m within each group. The parameterization of this definition by M makes the idea quite general. For example, $M = 1$ reduces to naïve adjustment behavior. On the other extreme, M could be arbitrarily large, leading to a different adjustment for each instance. Given our interest in studying the effect of outliers on adjusting behavior, we consider a 2-differential adjustment behavior, where the human chooses to use a different adjustment for the set of inliers, δ_I for $\mathcal{D}_I$, and the set of outliers, δ_O for $\mathcal{D}_O$. Although we consider only these 2 groups to build intuition for our hypotheses, our analyses could be extended to $M > 2$.

A person using 2-differential adjustment behavior with inlier vs. outlier partitions could use historical instances of y_i and $\hat{y}_i^{alg}$ to estimate the following optimization problem, which describes the optimal choice of adjustments on inliers and outliers (δ_I^{M2Adj} and δ_O^{M2Adj}, respectively):

$$M2Adj: \min_{\delta_I,\delta_O\in\mathbb{R}} \mathbb{E}[\big(Y_i - (\hat{Y}_i^{alg} + \delta_I)\big)^2 | \boldsymbol{X}_i \in \mathcal{D}_I] P(\boldsymbol{X}_i \in \mathcal{D}_I) + \\ \mathbb{E}[\big(Y_i - (\hat{Y}_i^{alg} + \delta_O)\big)^2 | \boldsymbol{X}_i \in \mathcal{D}_O] P(\boldsymbol{X}_i \in \mathcal{D}_O). \quad (5)$$

We assume here that the human can reliably and correctly differentiate between inliers and outliers, whether by experience with historical data or extensive domain expertise. This is to build theory only; we do not make this assumption in our experiments. The next proposition shows that 2-differential adjustment behavior with inlier vs. outlier partitions is superior to naïve adjustment behavior, and compares δ^{NAdj} to δ_O^{M2Adj} and δ_I^{M2Adj}.

PROPOSITION 2. *Under Assumption 1, $OPT^{M2Adj} \leq OPT^{NAdj}$, where OPT^{M2Adj} and OPT^{NAdj} are the optimal squared errors of $M2Adj$ and $NAdj$, respectively. Furthermore,* $\min(\delta_I^{M2Adj}, \delta_O^{M2Adj}) < \delta^{NAdj} < \max(\delta_I^{M2Adj}, \delta_O^{M2Adj})$.

The first part of Proposition 2 shows that the naïve adjuster — due to the inflexibility in their adjustments — suffers from worse performance. The second part of the proposition shows that adjustments made by the naïve adjuster fall between the adjustments made on inliers vs. outliers by the 2-differential adjuster. For ease of exposition, and in keeping with our experiments and many AI implementations in practice, we present a corollary to Proposition 2 for the special case where the algorithm's prediction is unbiased for inliers.

COROLLARY 1. *Under Assumption 1, and if $\mathbb{E}[Y_i - \hat{Y}_i^{alg} | \boldsymbol{X}_i \in \mathcal{D}_I] = 0$, $0 = \delta_I^{M2Adj} < |\delta^{NAdj}| < |\delta_O^{M2Adj}|$.*

In the special case of Corollary 1, we can see that someone biased towards naïve adjustment behavior makes adjustments that are too large in magnitude on inliers and too small in magnitude on outliers. For instance, if covariate shift was the source of the outliers, then this person would over-adjust on data points not impacted by covariate shift and under-adjust on data points impacted by covariate shift.

3.4. Hypotheses and Preview of Experiments

We first designed a controlled, online experiment, discussed in detail in Section 4, to test the implications of Proposition 2. Specifically, we implement a human-AI collaboration where participants are given ($\boldsymbol{x}_i$, $\hat{y}_i^{alg}$) sequentially for a set of instances and asked to make an adjustment to $\hat{y}_i^{alg}$ in order to make their final prediction. We employ three conditions that differ only as follows: relative

to the training set $\mathcal{S}$, the instances for which participants are asked to make adjustments to AI predictions are either (i) all inliers, (ii) all outliers, or (iii) a mixture of inliers and outliers. In the last two conditions, all outliers are a result of covariate shift. Participants are not explicitly informed whether $\boldsymbol{x}_i$ is an inlier or outlier, although they could infer this by comparing $\boldsymbol{x}_i$ to a representative sample of the training set $\mathcal{S}$ that they review at the beginning of the experiment. In addition, after studying a sample of the training set in a practice phase, participants in each condition observe the algorithm's performance and true outcome for many different realizations of $\boldsymbol{X}_i$. These realizations are representative of the instances in that condition: inliers, outliers, or both. This experience provides participants with the "domain expertise" they may have in practice that could influence their adjustment decisions.

We are most interested in the third condition — where participants face both inliers and outliers — as this is a key motivation in practice behind deploying AI as part of a human-AI collaboration. Thus, we aim to empirically study whether participants in the third condition are biased *towards* naïve adjustment behavior and suffer an increase in prediction error as a result. To do this, we use conditions (i) and (ii) to represent what people *would* do if they were to make adjustments by separately considering inliers vs. outliers since, by construction, people in each of these conditions only face one type of data point (either all inliers or all outliers).

Proposition 2 and Corollary 1 lead us directly to the following two hypotheses:

HYPOTHESIS 1. *Relative to people who are only tasked with making predictions on inliers or who are only tasked with making predictions on outliers, people who are tasked with making predictions on both inliers and outliers will over-adjust (under-adjust) the algorithm's predictions on inliers (outliers).*

HYPOTHESIS 2. *Relative to people who are only tasked with making predictions on inliers or who are only tasked with making predictions on outliers, people who are tasked with making predictions on both inliers and outliers will have higher prediction error on both inliers and outliers.*

After identifying a bias towards naïve adjustment behavior in our first experiment, our next goal is to mitigate this bias. To do so, we educate humans about whether data points are inliers vs. outliers, along with warning them that the algorithm may perform poorly on outliers vs. endorsing the algorithm for likely performing well on inliers. We hypothesize that these interventions will help mitigate naïve adjustment behavior by explicitly giving humans a simple partition of the data and reason to apply different adjustments in each subset. In our second experiment, we test the

effectiveness of these warnings and endorsements in the setting where participants face both inliers and outliers (just like condition (iii) in our first experiment).

Proposition 2 and Corollary 1 lead us to the following two hypotheses related to how warnings and endorsements impact adjustment behavior:

HYPOTHESIS 3. *Relative to people who receive no warnings or endorsements, people who receive only warnings on outliers will under-adjust less on outliers, and people who receive both warnings on outliers and endorsements on inliers will under-adjust (over-adjust) less on outliers (inliers).*

HYPOTHESIS 4. *Relative to people who receive only warnings on outliers, people who receive both warnings on outliers and endorsements on inliers will over-adjust less on inliers.*

Furthermore, Proposition 2 leads us to the following two hypotheses related to how warnings and endorsements impact predictive performance:

HYPOTHESIS 5. *Relative to people who receive no warnings or endorsements, people who receive only warnings on outliers will have lower prediction error on outliers and overall, and people who receive both warnings on outliers and endorsements on inliers will have lower prediction error on outliers, inliers, and overall.*

HYPOTHESIS 6. *Relative to people who receive only warnings on outliers, people who receive both warnings on outliers and endorsements on inliers will have lower prediction error on inliers and overall.*

4. Experiment I: Humans Exhibit Bias Towards Naïve Adjustment Behavior

This study tests Hypotheses 1 and 2 in a controlled online lab experiment. We pre-registered our experiment, including sample size, treatment conditions, exclusion criteria and analysis, here. All statistical tests reported in these results are, unless otherwise indicated, pre-registered.

4.1. Design

4.1.1. Participant Experience Participants are asked to predict daily demand of a product. Each new day i is characterized by two features ("Feature A" and "Feature B") that constitute the data $\boldsymbol{X}_i = (A_i, B_i)$. The outcome Y_i is the actual demand for day i, and participants make predictions by adjusting an algorithm's prediction $\hat{y}_i^{alg}$. Their adjustment plus the algorithm's prediction constitutes their final prediction, $\hat{y}_i^{final}$. The following steps provide more details about the participant journey. Screenshots depicting these steps are provided in Appendix C.

1. *Instructions & Comprehension Checks*. Participants are introduced to the concept of predicting daily demand — via adjustments — using realizations of Features A and B and the algorithm's prediction $\hat{y}_i^{alg}$. They are told that the algorithm was designed to help predict demand, and are tested for comprehension of their objective: minimizing absolute prediction error. Participants are told they will eventually make predictions in a Practice Phase and an incentivized Final Phase (in addition to a base compensation of $\$7$, participants receive a bonus of $\$7 - \$0.20 \times$ (Root Mean Squared Error)) and are tested for comprehension on what they will have access to during each prediction (Features A and B and the algorithm's prediction).

2. *Review Historical Data*. Participants are given realized historical data for 15 days. Each day provides the historical realization of features $\boldsymbol{x}_i = (a_i, b_i)$, the true demand y_i, the algorithm's prediction $\hat{y}_i^{alg}$, and the algorithm's absolute error $|y_i - \hat{y}_i^{alg}|$. Participants may continue to review additional historical data as desired.

3. *Practice Phase*. Participants are tested on their understanding of how their adjustments affect predictions by practicing on a single day; they may continue to practice by making adjustments and observing performance data on additional practice days as desired. Then, participants make predictions (via adjustments) for $i = 1, ..., 15$ new days. After each prediction, they are told the true outcome y_i, their absolute error $|y_i - \hat{y}_i^{final}|$, and the algorithm's absolute error $|y_i - \hat{y}_i^{alg}|$. At the end of the Practice Phase, participants are shown a summary table with a_i, b_i, y_i, $\hat{y}_i^{alg}$, $\hat{y}_i^{final}$, $|y_i - \hat{y}_i^{alg}|$, and $|y_i - \hat{y}_i^{final}|$ for each of the 15 days.

4. *Final Phase*. Just as in the Practice Phase, participants make predictions (via adjustments), this time for $i = 1, ..., 20$ new days. After each prediction, they are told the true outcome y_i, their absolute error $|y_i - \hat{y}_i^{final}|$, and the algorithm's absolute error $|y_i - \hat{y}_i^{alg}|$. Unlike the Practice Phase, their predictions in the Final Phase are incentivized.

Notably, participants are not given $f_{actual}(\boldsymbol{X}_i)$ or $f_{alg}(\boldsymbol{X}_i)$, nor are they explicitly told which data points lie in $\mathcal{D}_I$ or $\mathcal{D}_O$, although they could infer this from their experience in Steps 2-3.

4.1.2. Data Generation The feature vectors $\boldsymbol{X}_i = (A_i, B_i)$ reviewed by the participants in Step 2 are generated by drawing A_i and B_i from independent discrete uniform random variables with supports {1, 2, ..., 21} and {1, 2, ..., 25}, respectively. We consider an algorithm whose set of historical training instances $\mathcal{S}$ have features generated from the same distributions. Thus, we naturally define a realization $\boldsymbol{x}_i = (a_i, b_i)$ as an inlier if it falls within the support of $\mathcal{S}$, i.e., if $a_i \in \{1, 2, ..., 21\}$ and $b_i \in \{1, 2, ..., 25\}$. If either a_i or b_i fall outside of this support — i.e., as a result of covariate shift — $\boldsymbol{x}_i$ would be defined as an outlier. We generate outliers in our experiment

by drawing A_i and B_i from independent discrete uniform random variables with supports {25, 26, ..., 45} and {1, 2, ..., 25}, respectively; note that Feature A's support is different for outliers vs. inliers, whereas Feature B's is the same. Demand is generated by the following function:

$$Y_i = f_{actual}(\boldsymbol{X}_i) + \epsilon_i = \begin{cases} 21 + A_i + 2B_i + \epsilon_i & \text{if } \boldsymbol{X}_i \in \mathcal{D}_I \\ 21 + 2A_i + 2B_i + \epsilon_i & \text{if } \boldsymbol{X}_i \in \mathcal{D}_O, \end{cases} \tag{6}$$

where noise ϵ_i is distributed as a discrete uniform random variable with support {-10, -9, ..., 10}. Notice that for outliers, Feature A has a larger impact on demand.

We assume that $\mathcal{S}$ contains enough historical instances such that the algorithm has been trained to recover the "optimal" predictive function on inliers. However, since the algorithm did not train on any outliers, it incorrectly uses this same predictive function on outliers. This leads to the following algorithm predictions for *all* feature vectors $\boldsymbol{X}_i$:

$$\hat{Y}_i^{alg} = \mathbb{E}_\epsilon[Y_i | \boldsymbol{X}_i \in \mathcal{D}_I] = 21 + A_i + 2B_i. \tag{7}$$

Note that the algorithm's absolute expected error on inliers with respect to the random noise ϵ_i is $|\mathbb{E}_\epsilon[Y_i - \hat{Y}_i^{alg} | \boldsymbol{X}_i \in \mathcal{D}_I]| = 0$, whereas the algorithm's absolute expected error on outliers with respect to the random noise ϵ_i is $|\mathbb{E}_\epsilon[Y_i - \hat{Y}_i^{alg} | \boldsymbol{X}_i \in \mathcal{D}_O]| = A_i > 0$; thus, the special case of Corollary 1 is satisfied. Specifically, for inlier data points, $Y_i - \hat{Y}_i^{alg} = \epsilon_i$, or the algorithm correctly predicts demand other than zero-mean noise. For outlier data points, $Y_i - \hat{Y}_i^{alg} = A_i + \epsilon_i$, or the algorithm underestimates the demand by Feature A plus zero-mean noise.

4.1.3. Conditions Participants are randomly assigned to one of three conditions. The only difference across conditions is $P(\boldsymbol{X}_i \in \mathcal{D}_I)$ – and therefore also $P(\boldsymbol{X}_i \in \mathcal{D}_O)$ – in Steps 3 and 4.

1. **All-Inliers**: $P(\boldsymbol{X}_i \in \mathcal{D}_I) = 1$ and $P(\boldsymbol{X}_i \in \mathcal{D}_O) = 0$, i.e., there are no outliers in Steps 3 and 4. Note that in this case the algorithm always makes optimal predictions.

2. **All-Outliers**: $P(\boldsymbol{X}_i \in \mathcal{D}_I) = 0$ and $P(\boldsymbol{X}_i \in \mathcal{D}_O) = 1$, i.e., there are no inliers in Steps 3 and 4. Note that in this case the algorithm is always expected to underestimate demand by A_i.

3. **Mixed**: $P(\boldsymbol{X}_i \in \mathcal{D}_I) = P(\boldsymbol{X}_i \in \mathcal{D}_O) = \frac{1}{2}$, i.e., there are both inliers and outliers in Steps 3 and 4 with equal probability. Note that in this case the algorithm sometimes makes optimal predictions (on inliers) and sometimes underestimates demand (on outliers).

These three conditions can be used to construct a 2×2 mixed design, with one dimension indicating whether data points are representative of the training set (inliers vs. outliers) and the

other dimension indicating whether humans are exposed to a 'single' type of data (either all inliers or all outliers) or 'mixed' data (both inliers and outliers) in Steps 3 and 4. See Table 1 for a depiction of how our experimental conditions map to this 2×2 mixed design.

Table 1 Mapping experimental conditions to 2×2 mixed design.

		Types of Data in Steps 3 & 4	
		Single	Mixed
Representativeness in Training Set	Inliers	*All-Inliers*	*Mixed*
	Outliers	*All-Outliers*	*Mixed*

We note that although both conditions (2) and (3) reflect covariate shift, the third condition is our condition of interest, as its mixed set of data points reflects practical applications where algorithms *sometimes* encounter outliers not represented in the algorithm's training set. The *All-Outliers* condition alone is less interesting, since if in practice covariate shift caused a complete shift from inliers to outliers, an algorithm trained on a now irrelevant data set would likely be discarded. Instead, we use the first two conditions together as controls. They allow us to observe the adjustments and performance of people who separately consider inliers vs. outliers since, by construction, participants in these two conditions are only shown one of these two types of data in Steps 3 and 4. Making comparisons across each row of Table 1 allows us to determine if humans facing a mix of both inliers and outliers can sufficiently distinguish between the two, or if there is evidence that humans are biased towards naïve adjustment behavior.

4.1.4. Dependent Variables To test the behavior predicted in Hypothesis 1, we use *median absolute adjustment (MedA)*. This is calculated for participant j across instances $i = 1, ..., 20$ in the Final Phase (Step 4) separately for inliers and outliers as

$$MedA_j^I = \text{median}(|\hat{y}_i^{final} - \hat{y}_i^{alg}| \; \forall i \text{ s.t. } \boldsymbol{x}_i \in \mathcal{D}_I); \tag{8}$$

$$MedA_j^O = \text{median}(|\hat{y}_i^{final} - \hat{y}_i^{alg}| \; \forall i \text{ s.t. } \boldsymbol{x}_i \in \mathcal{D}_O). \tag{9}$$

Note that $MedA_j^I$ is calculated only in conditions (1) and (3), and $MedA_j^O$ is calculated only in conditions (2) and (3).

To test the performance predicted in Hypothesis 2, we use *median absolute deviation from the optimal prediction (MedDOP)*. Once again, this is calculated for participant j across instances $i = 1, ..., 20$ in the Final Phase (Step 4) separately for inliers and outliers:

$$MedDOP_j^I = \text{median}(|\hat{y}_i^{final} - \hat{y}_i^{alg}| \; \forall i \text{ s.t. } \boldsymbol{x}_i \in \mathcal{D}_I); \tag{10}$$

$$MedDOP_j^O = \text{median}(|\hat{y}_i^{final} - \hat{y}_i^{alg} - a_i| \; \forall i \text{ s.t. } \boldsymbol{x}_i \in \mathcal{D}_O). \tag{11}$$

In a sense, ***MedDOP*** is a more precise version of median absolute error – a more common performance metric – that we can use because we are controlling the data generation process and thus know the optimal prediction $\mathbb{E}[Y_i]$. For inliers, since $\mathbb{E}[Y_i] = \hat{y}_i^{alg}$, a participant's absolute deviation from the optimal prediction will be identical to their absolute adjustment to the algorithm's prediction; notice that $MedA_j^I = MedDOP_j^I$. However, for outliers, the optimal prediction for instance i is $\hat{y}_i^{alg} + a_i$ and thus a participant's absolute deviation from the optimal prediction will not generally equal their absolute adjustment. Just like median absolute error, a lower ***MedDOP*** indicates a better performing human-AI collaboration, and optimal predictions result in a ***MedDOP*** of zero.

4.2. Results

We ran a study on Prolific and recruited 300 participants who were located in the United States, had completed at least a High School diploma, had an approval rating of $99\% - 100\%$ on Prolific and had at least 25 previous submissions on the platform. Additional participant details, including treatment assignment, exclusion criteria, and payment data, are available in Appendix B.1.

4.2.1. Adjustment Results The key dependent variable — median absolute adjustment to the algorithm ($MedA$) — is depicted in Figure 1. We can use participant adjustments in the *All-Inliers* and *All-Outliers* conditions to infer the behavior of someone considering inliers and outliers separately since, by construction, participants in each of these two conditions experience just a single type of data in Steps 3 and 4. Unsurprisingly, when people separately make adjustments on inliers and outliers, they correctly make smaller absolute adjustments to the algorithm on inliers than on outliers ($t = -30.545$, $p < .0001$). Similarly, participants in the *Mixed* condition who make adjustments on both inliers and outliers in Steps 3 and 4 also make smaller absolute adjustments to the algorithm on inliers than on outliers ($t = -7.316$, $p < .0001$); that is, there is evidence of at least some differentiation in adjustments as opposed to purely naïve adjustment behavior.

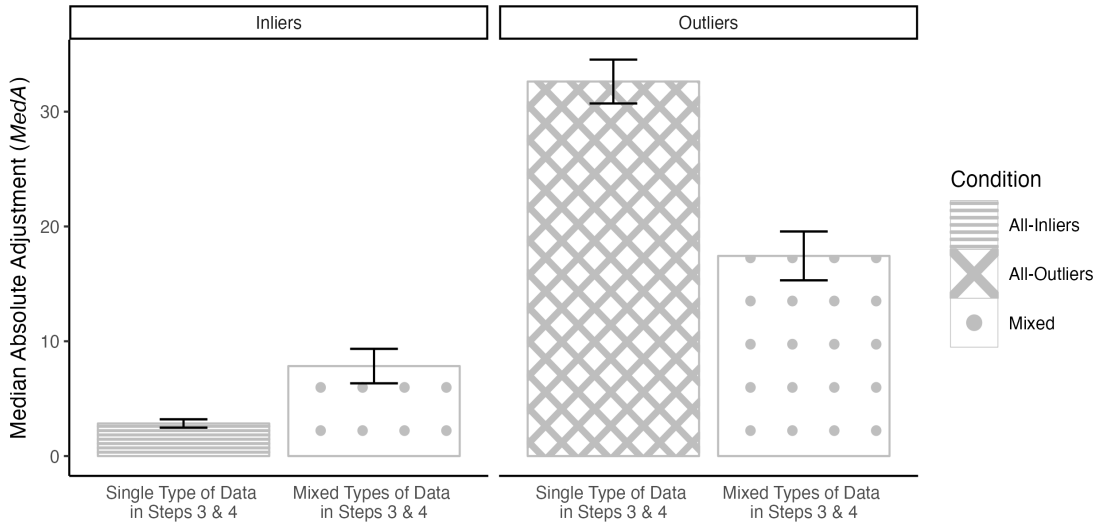


Figure 1 **Median absolute adjustment results are averaged (mean) by the types of data the participant observes in Steps 3 and 4 (single or mixed), separately for inliers and outliers. Standard error bars included.**

However, as stated in Hypothesis 1, we hypothesize that participants in the *Mixed* condition do not differentiate adjustments *enough*, i.e., they have a bias *towards* naïve adjustment behavior.

To evaluate this hypothesis, we perform two one-sided t-tests that compare mean values of *MedA* across each row of Table 1. Let $\mathcal{C}_k$ be the set of participants assigned to condition k, for $k \in \{1, 2, 3\}$ corresponding to the *All-Inliers*, *All-Outliers* and *Mixed* conditions, respectively. Our first t-test compares the mean *MedA* for inliers ($\frac{\sum_{j\in\mathcal{C}_3} MedA_j^I}{|\mathcal{C}_3|} > \frac{\sum_{j\in\mathcal{C}_1} MedA_j^I}{|\mathcal{C}_1|}$), i.e., whether people who experience a mixture of both inliers and outliers in Steps 3 and 4 make larger absolute adjustments on inliers compared to people who experience only inliers. As indicated by the left two bars of Figure 1 and confirmed in our t-test, absolute adjustments on inliers in the *Mixed* condition were significantly larger than absolute adjustments on inliers in the *All-Inliers* condition ($t = 6.4284,\ p < .0001$).

Similarly, our second t-test compares the mean *MedA* for outliers ($\frac{\sum_{j\in\mathcal{C}_3} MedA_j^O}{|\mathcal{C}_3|} < \frac{\sum_{j\in\mathcal{C}_2} MedA_j^O}{|\mathcal{C}_2|}$), i.e., whether people who experience a mixture of both inliers and outliers in Steps 3 and 4 make smaller absolute adjustments on outliers compared to people who experience only outliers. As indicated by the right two bars of Figure 1 and confirmed in our t-test, absolute adjustments on outliers in the *Mixed* condition were significantly smaller than absolute adjustments on outliers in the *All-Outliers* condition ($t = -10.567,\ p < .0001$).

These results support Hypothesis 1. Humans facing covariate shift who are exposed to both inliers and outliers over-adjust (under-adjust) the algorithm on inliers (outliers) compared to what they would have done had they only been exposed to inliers (outliers). In general, this experiment indicates that humans are biased towards naïve adjustment behavior in that they do not sufficiently distinguish between inliers and outliers when faced with both types of instances.

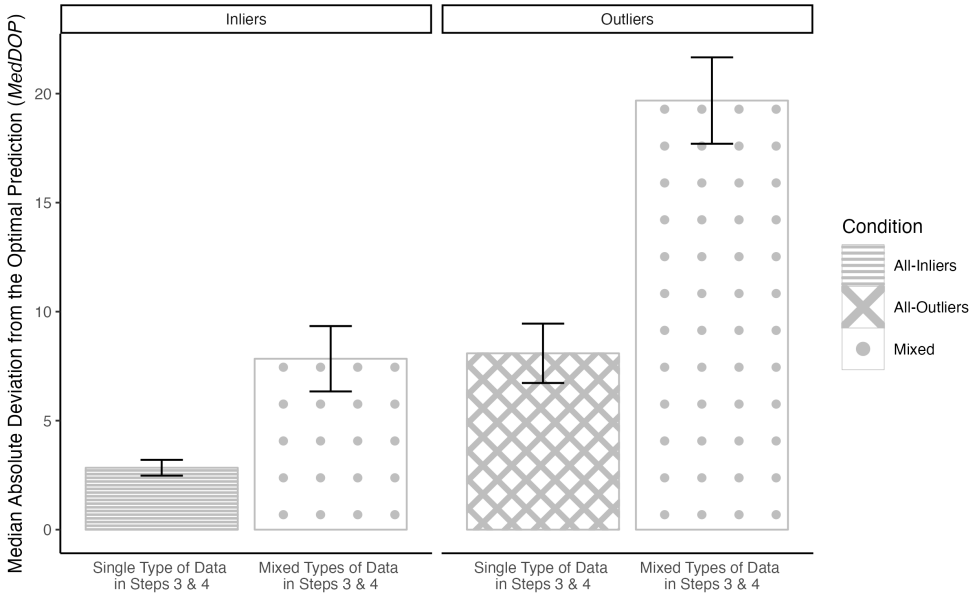


Figure 2 **Median absolute deviation from the optimal prediction results are averaged (mean) by the types of data the participant observes in Steps 3 and 4 (single or mixed), separately for inliers and outliers. Standard error bars included.**

4.2.2. Performance Results In this section, we explore the consequences of a bias towards naïve adjustment behavior on predictive performance. Our key dependent variable, median absolute deviation from the optimal prediction ($MedDOP$), is depicted in Figure 2.

As detailed in Hypothesis 2, we predict that the bias towards naïve adjustment behavior confirmed in Section 4.2.1 will lead to worse prediction errors on both inliers and outliers for people in the *Mixed* condition compared to people in the *All-Inliers* and *All-Outliers* conditions. To evaluate this hypothesis, we perform two one-sided t-tests that compare mean values of *MedDOP* across each row of Table 1. Our first t-test compares the mean *MedDOP* for inliers ($\frac{\sum_{j\in\mathcal{C}_3} MedDOP_j^I}{|\mathcal{C}_3|} > \frac{\sum_{j\in\mathcal{C}_1} MedDOP_j^I}{|\mathcal{C}_1|}$), i.e., whether people who experience a mixture of both inliers and outliers in Steps 3 and 4 make larger absolute deviations from optimal predictions on inliers compared to people

who experience only inliers. As indicated by the left two bars of Figure 2 and confirmed in our t-test, absolute deviations from optimal predictions on inliers in the *Mixed* condition were significantly larger than those on inliers in the *All-Inliers* condition ($t = 6.4284,\ p < .0001$). As described in Section 4.1.4, note that this is an equivalent test and result to that of $MedA$ for inliers, since $MedA_j^I = MedDOP_j^I$ for each participant j. Performance degradation is substantial: the mean *MedDOP* on inliers is 176% larger in the *Mixed* condition compared to the *All-Inliers* condition.

Similarly, our second t-test compares the mean *MedDOP* for outliers ($\frac{\sum_{j\in\mathcal{C}_3} MedDOP_j^O}{|\mathcal{C}_3|} > \frac{\sum_{j\in\mathcal{C}_2} MedDOP_j^O}{|\mathcal{C}_2|}$), i.e., whether people who experience a mixture of both inliers and outliers in Steps 3 and 4 make larger absolute deviations from optimal predictions on outliers compared to people who experience only outliers. As indicated by the right two bars of Figure 2 and confirmed in our t-test, absolute deviations from optimal predictions on outliers in the *Mixed* condition were significantly larger than those on outliers in the *All-Outliers* condition ($t = 9.592,\ p < .0001$). Performance degradation is substantial: the mean *MedDOP* on outliers is 143% larger in the *Mixed* condition compared to the *All-Outliers* condition.

Ultimately, we find results in support of Hypothesis 2. Participants who are exposed to both inliers and outliers perform worse — *much* worse — on both inliers and outliers, compared to the participants only exposed to a single type. This is consequential, since many human-AI collaborations face some mixture of inliers and outliers.

4.2.3. Summary of Additional Analyses We carry out robustness checks in Appendix E to ensure that our results persist when controlling for a variety of reported demographic data. We find that the effects in Sections 4.2.1 and 4.2.2 that support Hypotheses 1 and 2 are still significant when controlling for age, level of education, gender and if the participant has taken a statistics class.

4.3. Discussion

Experiment I provides evidence supporting a bias towards *naïve adjustment behavior*: participants were unable to sufficiently differentiate their adjustments between inliers and outliers, leading to a significant degradation in prediction accuracy. Notably, in the recent paper Balakrishnan et al. (2022), they also find evidence that humans are unable to sufficiently differentiate their reliance on algorithms, albeit in a different setting that considers the presence of private information as opposed to covariate shift. In both papers, participants were found to over-adhere to the algorithm when it made sizable errors, and under-adhere to the algorithm when it made more accurate predictions. Importantly, this general finding from both our work and Balakrishnan et al. (2022) suggests

that interventions targeted at helping humans better differentiate when they should adhere more vs. deviate more from an algorithm could be quite valuable to improve human-AI collaboration. In Section 5, we develop and test the effectiveness of such interventions for covariate shift.

As with any lab experiment, it is important to consider how the findings translate to the field, and where limitations may exist. For algorithms that, like ours, only consider two features, it may be reasonable to expect that trained domain experts would be more adept at identifying outliers than the participants in our experiment. However, even the most experienced domain experts will likely find it challenging to identify outliers when there are many — potentially hundreds of — features, as is often the case in applied settings, or when the process of feature engineering transforms raw data points. Further, one could imagine a case where outliers are defined conditionally: perhaps the i^{th} value of Feature A is unremarkable compared to the *marginal* distribution of Feature A within the training set, but *conditional* on the i^{th} value of Feature B it may be regarded as an outlier. One could imagine the challenge this data point might pose an algorithm, and the difficulty a human — even a domain expert — would face in identifying it as an outlier. Although this is not a direction we explore in the lab, our theory presented in Section 3 leads us to the same set of hypotheses.

5. Experiment II: Mitigating Bias via Warnings and Endorsements

After identifying a bias towards naïve adjustment behavior in our first experiment, our next goal is to mitigate this bias and improve the human-AI collaboration in the presence of covariate shift. What operational design choices might be helpful in this scenario?

One possibility would be to restrict human access by only allowing humans to make adjustments to the algorithm on outliers that are the result of covariate shift; on inliers, the algorithm's recommendations could be automatically deployed. While this design would likely work well in our experiment, it would be unlikely to generalize in practice: there are many additional factors beyond covariate shift that would lead to beneficial human adjustments on inliers. For example, private information, algorithm 'glitches', operational constraints, and high levels of external uncertainty all create environments where a human adjustment to the algorithm on inliers not impacted by covariate shift may provide significant value (Balakrishnan et al. 2022, De-Arteaga et al. 2020, Sun et al. 2022, Kesavan and Kushwaha 2020). We therefore seek to implement an intervention that helps humans appropriately adjust the algorithm on outliers while maintaining critical adjustment authorization on both outliers and inliers.

At the core of our intervention, we educate humans about whether data points are inliers vs. outliers, along with warning them that the algorithm may perform poorly on outliers vs. endorsing the algorithm for likely performing well on inliers. We hypothesize that these warnings and endorsements will help mitigate the bias towards naïve adjustment behavior by explicitly giving humans a simple partition of the data and reason to apply different adjustments in each subset; our theory in Section 3 suggests this is a superior strategy.

We also consider a 'warnings only' intervention – flagging data points that are outliers and warning humans that the algorithm may perform poorly – that does not flag inliers or endorse the algorithm's accuracy on them. Similar warnings-only interventions have been studied in recent papers (Poursabzi-Sangdeh et al. 2021, Chiang and Yin 2021), but we extend these studies by considering how these warnings impact adjustment behavior and predictive performance on *both* outliers and inliers. It is a priori unclear whether a warnings-only intervention will (i) help humans implicitly understand that non-flagged instances are inliers and should have better algorithm performance, or (ii) erode human trust in the algorithm altogether and thus increase absolute adjustments on both outliers and inliers.

We conduct a controlled online lab experiment to study these interventions and test Hypotheses 3–6. We pre-registered our experiment, including sample size, treatment conditions, exclusion criteria and analysis, here. All statistical tests reported are, unless otherwise indicated, pre-registered.

5.1. Design

The format of this experiment is nearly identical to the *Mixed* condition from Experiment I. Specifically, the participant experience, data generation, and dependent variables are identical to what is described for the *Mixed* condition in Sections 4.1.1, 4.1.2, and 4.1.4, respectively, except for the additions detailed below in the description of each condition. We also include an additional version of our $MedDOP$ dependent variable, defined for the overall set of instances participant j encounters, since we are interested in not only the performance on inliers vs. outliers, but also on the overall performance. Specifically, we define

$$MedDOP_j = \text{median}(|\hat{y}_i^{final} - \hat{y}_i^{alg}| \,\forall i \text{ s.t. } \boldsymbol{x}_i \in \mathcal{D}_I, |\hat{y}_i^{final} - \hat{y}_i^{alg} - a_i| \,\forall i \text{ s.t. } \boldsymbol{x}_i \in \mathcal{D}_O). \quad (12)$$

The only difference in the study design is the set of experimental conditions, where we implement our interventions intended to mitigate the bias towards naïve adjustment behavior. Specifically, participants are randomly assigned to one of the following three conditions:

1. **No Warnings or Endorsements**: Identical to the *Mixed* condition in Experiment I.

2. **Warnings Only**: At the beginning of the Practice Phase and Final Phase (Steps 3 and 4), the following message is displayed to participants to warn them about the possibility that the algorithm may perform poorly on outliers: "The historical data had values of Feature A ranging from 1 to 21 and values of Feature B ranging from 1 to 25. The algorithm may perform poorly on predictions with feature values outside of those ranges." A similar message was given above the summary table at the end of Step 3. Further, in Steps 3 and 4, for each instance i such that $\boldsymbol{x}_i$ is an outlier, the participant is warned that a feature value is outside the range of the training set and thus the algorithm may perform poorly. Messages in the *Warnings Only* and *Warnings and Endorsements* conditions are depicted in Appendix D.

3. **Warnings and Endorsements**: At the beginning of the Practice Phase and Final Phase (Steps 3 and 4), the following message is displayed to participants to warn them about the possibility that the algorithm may perform poorly on outliers and endorse the algorithm's likely strong performance on inliers: "The historical data had values of Feature A ranging from 1 to 21 and values of Feature B ranging from 1 to 25. The algorithm is expected to perform well on predictions with feature values within those ranges, but may perform poorly on predictions with feature values outside of those ranges." A similar message was given above the summary table at the end of Step 3. Furthermore, in Steps 3 and 4, for each instance i such that $\boldsymbol{x}_i$ is an outlier, the participant is warned that a feature value is outside the range of the training set and thus the algorithm may perform poorly (identical to the *Warnings-Only* condition). In addition, for each instance i such that $\boldsymbol{x}_i$ is an inlier, the participant is told that the feature values are within the range of the training set and given an endorsement that the algorithm is expected to perform well.

5.2. Results

We ran a study on Prolific and recruited 450 participants. The participants were located in the United States, had completed at least a High School diploma, had an approval rating of $99\% - 100\%$ on Prolific and had at least 25 previous submissions on the platform. Additional participant details are available in Appendix B.2. To help describe our results, we define $\mathcal{C}_N$, $\mathcal{C}_W$, and $\mathcal{C}_{WE}$ as the set of participants assigned to condition *No Warnings or Endorsements*, *Warnings Only*, and *Warnings and Endorsements*, respectively.

5.2.1. Adjustment Results The key dependent variable — median absolute adjustment to the algorithm ($MedA$) — is depicted in Figure 3.

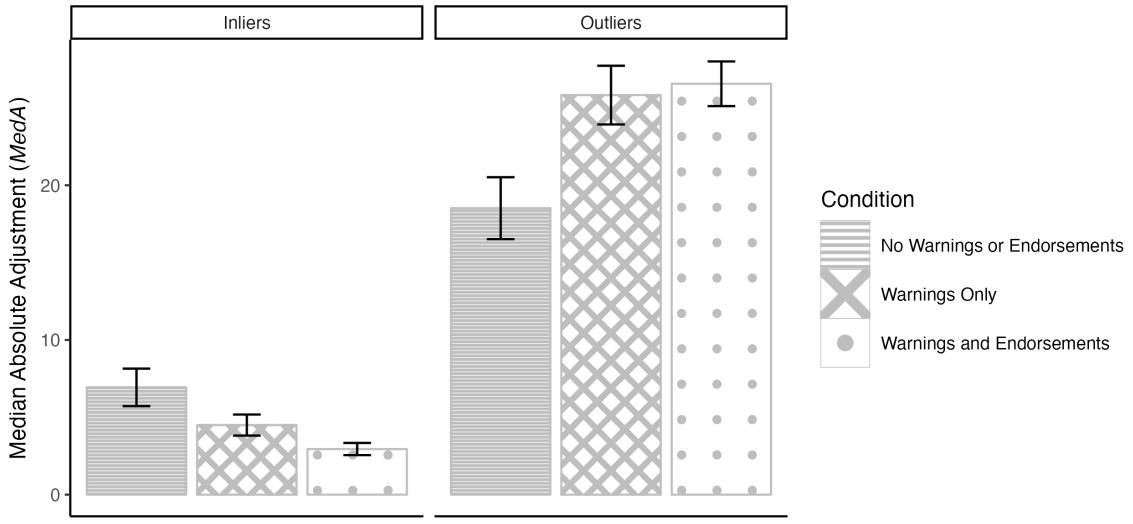


Figure 3 **Median absolute adjustment results are averaged (mean) by condition, separately for inliers and outliers. Standard error bars included.**

As detailed in part of Hypothesis 3, we hypothesize that participants in the *Warnings Only* and *Warnings and Endorsements* conditions will make larger absolute adjustments on outliers than participants in the *No Warnings or Endorsements* condition. To test this hypothesis, we perform two one-sided t-tests comparing the mean *MedA* for outliers ($\frac{\sum_{j\in\mathcal{C}_W} MedA_j^O}{|\mathcal{C}_W|} > \frac{\sum_{j\in\mathcal{C}_N} MedA_j^O}{|\mathcal{C}_N|}$ and $\frac{\sum_{j\in\mathcal{C}_{WE}} MedA_j^O}{|\mathcal{C}_{WE}|} > \frac{\sum_{j\in\mathcal{C}_N} MedA_j^O}{|\mathcal{C}_N|}$).

As indicated by the three right-hand bars of Figure 3 and confirmed in our t-tests, absolute adjustments on outliers in the *Warnings Only* and *Warnings and Endorsements* conditions were significantly larger than absolute adjustments on outliers in the *No Warnings or Endorsements* condition ($t = 5.2447,\ p < .0001$ and $t = 6.4427,\ p < .0001$, respectively). These results provide evidence that warning humans about covariate shift and flagging outliers – regardless of whether additional endorsements are provided on inliers representative of the training data – is an effective intervention if the aim is solely to increase absolute adjustments on outliers.

However, we are also concerned about the impact of interventions on inliers. As detailed in the remainder of Hypothesis 3, we hypothesize that participants in the *Warnings and Endorsements* condition will make smaller absolute adjustments on inliers compared to participants in the *No Warnings or Endorsements* condition. To test this hypothesis, we perform a one-sided t-test comparing the mean *MedA* for inliers ($\frac{\sum_{j\in\mathcal{C}_{WE}} MedA_j^I}{|\mathcal{C}_{WE}|} < \frac{\sum_{j\in\mathcal{C}_N} MedA_j^I}{|\mathcal{C}_N|}$). Furthermore, as detailed in Hypothesis 4, we hypothesize that participants in the *Warnings and Endorsements* condition will

make smaller absolute adjustments on inliers compared to participants in the *Warnings Only* condition. To test this hypothesis, we perform a one-sided t-test comparing the mean *MedA* for inliers ($\frac{\sum_{j\in\mathcal{C}_{WE}} MedA_j^I}{|\mathcal{C}_{WE}|} < \frac{\sum_{j\in\mathcal{C}_W} MedA_j^I}{|\mathcal{C}_W|}$). As indicated by the three left-hand bars of Figure 3 and confirmed in our t-tests, absolute adjustments on inliers in the *Warnings and Endorsements* condition were significantly smaller than absolute adjustments on inliers in the *No Warnings or Endorsements* condition ($t = -6.1919,\ p < .0001$) and in the *Warnings Only* condition ($t = -3.8937,\ p < .0001$).

These results provide strong evidence for Hypotheses 3-4. Humans who are given warnings on outliers make larger absolute adjustments on outliers, regardless of whether they are also given endorsements on inliers. Put differently, endorsing inliers does not seem to lead to over-trust in an algorithm that negates the gains of outlier warnings. Furthermore, humans who also receive endorsements on inliers make smaller absolute adjustments on inliers than those who receive no warnings or endorsements, and importantly, than those who only receive an outlier warning. These results suggest that it is important to explicitly provide humans with information about *both* subsets – inliers and outliers – and educate them as to why they might want to make different adjustments for each subset. In general, this experiment indicates that both warnings *and* endorsements combat the bias towards naïve adjustment behavior in the presence of covariate shift; namely, participants are able to differentiate adjustments between inliers and outliers to a higher degree.

5.2.2. Performance Results In this section, we discuss performance implications of behavior changes induced by our interventions and evaluate Hypotheses 5-6. Our key dependent variable, median absolute deviation from the optimal prediction ($MedDOP$), is depicted in Figure 4.

We hypothesize that our *Warnings Only* and *Warnings and Endorsements* interventions will lead to improved prediction errors on outliers, compared to prediction errors of people who receive neither intervention. To evaluate these hypotheses, we perform two one-sided t-tests that compare mean values of $MedDOP$ ($\frac{\sum_{j\in\mathcal{C}_W} MedDOP_j^O}{|\mathcal{C}_W|} < \frac{\sum_{j\in\mathcal{C}_N} MedDOP_j^O}{|\mathcal{C}_N|}$ and $\frac{\sum_{j\in\mathcal{C}_{WE}} MedDOP_j^O}{|\mathcal{C}_{WE}|} < \frac{\sum_{j\in\mathcal{C}_N} MedDOP_j^O}{|\mathcal{C}_N|}$). As indicated by the three right-hand bars of Figure 4 and confirmed in our t-tests, absolute deviations from the optimal predictions on outliers in both the *Warnings Only* and *Warnings and Endorsements* conditions were significantly smaller than those in the *No Warnings or Endorsements* condition ($t = -5.3399,\ p < .0001$ and $t = -6.6564,\ p < .0001$, respectively). Improvement is substantial: relative to no warnings or endorsements, warnings led to a 31% reduction in mean $MedDOP$ on outliers, and warnings *and* endorsements led to a 37% reduction.

We next evaluate how our interventions impact prediction errors on inliers. We hypothesize that our *Warnings and Endorsements* intervention will lead to improved prediction errors on inliers

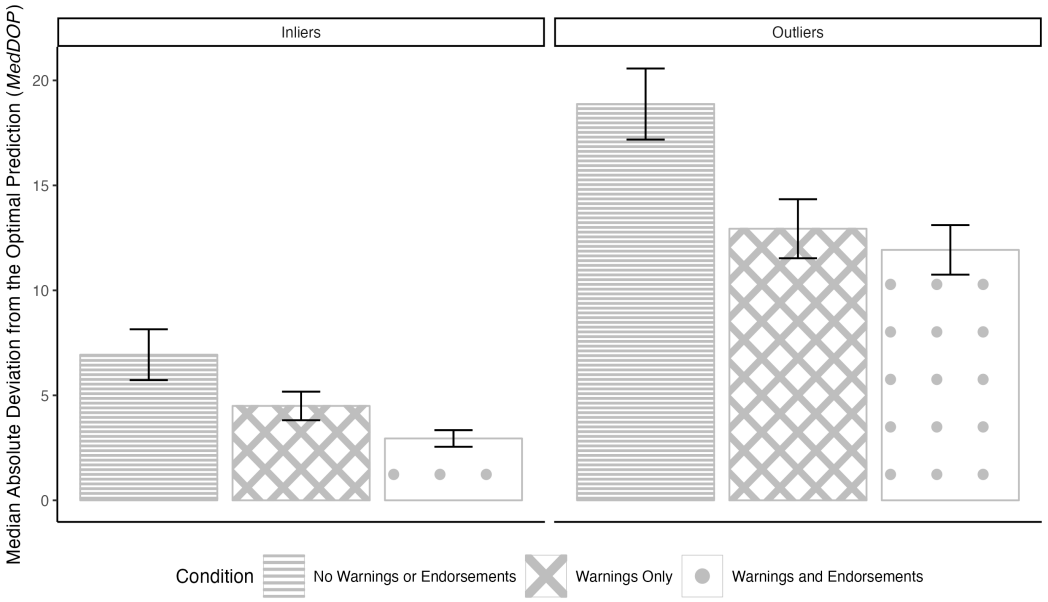


Figure 4 **Median absolute deviation from the optimal prediction results are averaged (mean) by condition, separately for inliers and outliers. Standard error bars included.**

compared to both the *No Warnings or Endorsements* baseline and the *Warnings Only* intervention. To test these hypotheses, we perform two one-sided t-tests that compare mean values of $MedDOP$ ($\frac{\sum_{j\in\mathcal{C}_{WE}} MedDOP_j^I}{|\mathcal{C}_{WE}|} < \frac{\sum_{j\in\mathcal{C}_N} MedDOP_j^I}{|\mathcal{C}_N|}$ and $\frac{\sum_{j\in\mathcal{C}_{WE}} MedDOP_j^I}{|\mathcal{C}_{WE}|} < \frac{\sum_{j\in\mathcal{C}_W} MedDOP_j^I}{|\mathcal{C}_W|}$). As indicated by the three left-hand bars of Figure 4 and confirmed in our t-tests, absolute deviations from the optimal predictions on inliers in the *Warnings and Endorsements* condition were significantly smaller than those in the *No Warnings or Endorsements* condition ($t = -6.1919,\ p < .0001$) and the *Warnings Only* condition ($t = -3.8937,\ p < .0001$). As described in Section 4.1.4, note that this is an equivalent test and result to that of $MedA$ for inliers, since $MedA_j^I = MedDOP_j^I$ for each participant j. Performance improvement is substantial: warnings and endorsements led to a 58% reduction in mean $MedDOP$ of inliers compared to no warnings or endorsements, and a 34% reduction in mean $MedDOP$ of inliers compared to warnings only.

We also found evidence that absolute adjustments on inliers — and, equivalently, absolute deviations from the optimal predictions on inliers — in the *Warnings Only* condition were significantly and substantially smaller than those in the *No Warnings or Endorsements* condition ($t = -3.4735, p = 0.0003$ and a 35% reduction in mean $MedDOP$ of inliers). We did not pre-register this hypothesis and, once again, the absolute deviations on inliers were *even lower* in the *Warnings and Endorsements* condition. However, this is reassuring evidence that warnings on outliers do not 'spill over' to inliers in the sense of participants making harmful, larger absolute adjustments.

Finally, we test our hypotheses that the overall absolute deviations from the optimal predictions (across both inliers and outliers) were the largest when no warnings or endorsements were provided; we hypothesized that warnings would improve overall performance, and the combination of warnings and endorsements would improve it even further. To evaluate these hypotheses, we perform three one-sided t-tests that compare mean values of $MedDOP$ ($\frac{\sum_{j\in\mathcal{C}_W} MedDOP_j}{|\mathcal{C}_W|} < \frac{\sum_{j\in\mathcal{C}_N} MedDOP_j}{|\mathcal{C}_N|}$, $\frac{\sum_{j\in\mathcal{C}_{WE}} MedDOP_j}{|\mathcal{C}_{WE}|} < \frac{\sum_{j\in\mathcal{C}_N} MedDOP_j}{|\mathcal{C}_N|}$ and $\frac{\sum_{j\in\mathcal{C}_{WE}} MedDOP_j}{|\mathcal{C}_{WE}|} < \frac{\sum_{j\in\mathcal{C}_W} MedDOP_j}{|\mathcal{C}_W|}$). The results confirm our hypotheses: the overall absolute deviation from the optimal prediction was smaller in the *Warnings Only* condition compared to the *No Warnings or Endorsements* condition ($t = -4.2677,\ p < .0001$) and in the *Warnings and Endorsements* condition compared to both the *No Warnings or Endorsements* condition ($t = -8.1905,\ p < .0001$) and the *Warnings Only* condition ($t = -3.5516,\ p = .0002$). Again, overall performance improvement is substantial: warnings led to a 29% reduction in mean $MedDOP$ compared to no warnings or endorsements, and both warnings and endorsements further decreased mean $MedDOP$ by 27%.

Ultimately, we find support for Hypotheses 5 -6. Participants who receive warnings on outliers perform better on outliers, and overall, compared to receiving no warning. But participants who receive warnings on outliers *and* endorsements on inliers have an additional performance improvement on inliers, and overall, compared to receiving no messages or only receiving warnings on outliers. There is even suggestive evidence that endorsements improve performance on *outliers*. Participants in the *Warnings and Endorsements* condition have lower absolute deviations from optimal predictions on outliers than participants in the *Warnings Only* condition; however, this difference is not significant ($t = -1.0852,\ p = 0.1394$) and not pre-registered.

5.2.3. Summary of Additional Analyses Similar to Experiment I, we carry out robustness checks in Appendix E to ensure our results persist when controlling for demographic data. Once again, we find that the results in Sections 5.2.1 and 5.2.2 remain significant even when controlling for age, level of education, gender and whether the participant has taken a statistics class.

5.3. Discussion

Experiment II provides evidence that warnings and endorsements are effective interventions for mitigating naïve adjustment behavior. The combination of warnings and endorsements explicitly gives humans a simple partition of the data and educates them as to why they should consider applying different adjustments in each subset. In turn, humans are able to more fully differentiate adjustments between inliers and outliers when facing covariate shift. The upshot of this improved

differentiation is a significant improvement in performance. Participants who received warnings and endorsements made substantially more accurate predictions on outliers, inliers, and overall compared to participants who didn't receive any interventions, and substantially more accurate predictions on inliers and overall compared to participants who received only warnings.

As before, it is important to consider how our findings in the lab translate to the field. In our experiment, Step 3 is used to provide the participant with intuition and "domain expertise". Studying the algorithm's performance along with the true outcomes, on realizations that are representative of their assigned condition, helps participants understand how to adjust the algorithm's predictions for inliers vs. outliers. We note that, in this experiment, a meta-algorithm could be constructed using the data in Step 3 to learn how to differentially adjust the algorithm's predictions, potentially circumventing the need for human intervention. While this approach might serve in our experiment, it is unlikely to be sufficient in practice for two main reasons. First, there can be many features — potentially hundreds — that can generate outliers, and outliers may even be defined conditionally (i.e., the value of Feature A is large with respect to the value of Feature B). Retraining an algorithm, or fitting a meta-algorithm, to try to learn the impact of these outliers could be time-consuming and would likely require several instances of similar feature vectors. Second, human decision-makers in practice may have broader contextual understanding regarding how to adjust the algorithm's prediction when a warning alerts them of an outlier. This intuition may lead to even better performance, since "domain expertise" in the field likely begets a deeper understanding than studying historical data instances (as we had to do in our more restrictive lab setting).

Finally, in a recent paper by Balakrishnan et al. (2022), they similarly develop a successful intervention targeted at educating humans as to why and when they should consider deviating more or less from algorithms (albeit in a different setting that considers the presence of private information as opposed to covariate shift). Together, our work and Balakrishnan et al. (2022) give strong support of a promising class of interventions to improve human-AI collaborations: helping the human understand how and when to differentially rely on the AI's recommendations.

6. Conclusion

We propose an anchor-and-adjust model to describe how humans might use an AI's recommendation in a human-AI collaboration. We consider the common setting where covariate shift occurs, i.e., where the AI algorithm is trained on historical data that is no longer fully representative of the prediction tasks at hand. This setting leads to a partition of feature vectors into inliers (for which

the AI's training set is representative) and outliers (for which the AI's training set is not representative). We hypothesize that humans are biased towards naïve adjustment behavior: insufficiently differentiating adjustments across inliers and outliers. When the algorithm has better performance on inliers compared to outliers – a dynamic that is commonplace in practice – we show via a mathematical model that naïve adjustment behavior leads to over-adjustments on inliers, under-adjustments on outliers and, ultimately, suboptimal predictive performance. Our results in Section 4 confirm the findings of our mathematical model.

To mitigate this bias towards naïve adjustment behavior, we consider interventions targeted at educating humans about two types of data for which they should consider making adjustments differently (inliers vs. outliers), while allowing humans to maintain override authority on all prediction tasks (including inliers). In practice, override authority is a critical component of human-AI collaboration, since there are numerous reasons beyond covariate shift for which the human should override the AI's recommendation. The most successful intervention — *Warnings and Endorsements* — explicitly gives humans information about the partition of outliers vs. inliers along with reasons to apply different adjustments in each subset via warnings and endorsements, respectively.

Importantly, our intervention is very easy to implement in practice. The training data set for which the algorithm is developed, $\mathcal{S}$, can be used at the time of (re-)training by the system designer to specify a set of rules defining which feature vector realizations would be classified as inliers; all others would be classified as outliers. Notice that this could be done regardless of knowledge of the underlying AI algorithm; in other words, our intervention could even be applied if the AI algorithm itself was a "black box" (i.e., developed by an external provider). Finally, minimal changes to the user interface and training would be required, and no prior knowledge of covariate shift or AI would be necessary for the human users.

We hope that our work inspires other researchers to consider how and when humans can make improvements on AI recommendations, and to design interventions to capitalize on these improvement opportunities while at the same time benefiting from AI's advantages. Ultimately, we hope that such a body of academic work will serve as a guide to practitioners on how to best equip their employees who are tasked with human-AI collaboration.

References

Aggarwal CC, Aggarwal CC (2017) *An introduction to outlier analysis* (Springer).

Babic B, Cohen IG, Evgeniou T, Gerke S (2021) When machine learning goes off the rails. *Harvard Business Review* 21–32.

Baier L, Jöhren F, Seebacher S (2019) Challenges in the deployment and operation of machine learning in practice. *ECIS*, volume 1.

Balakrishnan M, Ferreira K, Tong J (2022) Improving human-algorithm collaboration: Causes and mitigation of over- and under-adherence. *Available at SSRN 4298669* .

Bayram F, Ahmed BS (2023) A domain-region based evaluation of ml performance robustness to covariate shift. *Neural Computing and Applications* 1–23.

Ben-Gal I (2005) Outlier detection. *Data mining and knowledge discovery handbook* 131–146.

Blyth S (2018) Big data and machine learning won't save us from another financial crisis. *Harvard Business Review* .

Boukerche A, Zheng L, Alfandi O (2020) Outlier detection: Methods, models, and classification. *ACM Computing Surveys (CSUR)* 53(3):1–37.

Caro F, Saez de Tejada Cuenca A (2023) Believing in analytics: Managers' adherence to price recommendations from a dss. *Manufacturing & Service Operations Management* 25(2):524–542.

Castelo N, Bos MW, Lehmann DR (2019) Task-dependent algorithm aversion. *Journal of Marketing Research* 56(5):809–825.

Chiang CW, Yin M (2021) You'd better stop! understanding human reliance on machine learning models under covariate shift. *13th ACM web science conference 2021*, 120–129.

De-Arteaga M, Fogliato R, Chouldechova A (2020) A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–12.

Dietvorst BJ, Simmons JP, Massey C (2015) Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144(1):114.

Dietvorst BJ, Simmons JP, Massey C (2018) Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science* 64(3):1155–1170.

Ferreira KJ, Lee BHA, Simchi-Levi D (2016) Analytics for an online retailer: Demand forecasting and price optimization. *Manufacturing & Service Operations Management* 18(1):69–88.

Furnham A, Boo HC (2011) A literature review of the anchoring effect. *The journal of socio-economics* 40(1):35–42.

Ibrahim R, Kim SH, Tong J (2021) Eliciting human judgment for prediction algorithms. *Management Science* 67(4):2314–2325.

Kahneman D SC Sibony O (2022) *Noise* (UK: HarperCollins).

Kawaguchi K (2021) When will workers follow an algorithm? a field experiment with a retail business. *Management Science* 67(3):1670–1695.

Kesavan S, Kushwaha T (2020) Field experiment on the profit implications of merchants' discretionary power to override data-driven decision-making tools. *Management Science* 66(11):5182–5190.

Kouw WM, Loog M (2019) A review of domain adaptation without target labels. *IEEE transactions on pattern analysis and machine intelligence* 43(3):766–785.

Kwon C, Raman A, Tamayo J (2022) Human-computer interactions in demand forecasting and labor scheduling decisions. *Available at SSRN 4296344* .

Lai V, Tan C (2019) On human predictions with explanations and predictions of machine learning models: A case study on deception detection. *Proceedings of the conference on fairness, accountability, and transparency*, 29–38.

McGrath S, Mehta P, Zytek A, Lage I, Lakkaraju H (2020) When does uncertainty matter?: Understanding the impact of predictive uncertainty in ML assisted decision making. *arXiv preprint arXiv:2011.06167* .

NewVantagePartners (2023) Data and analytics leadership annual executive survey. Technical report, NewVantage Partners.

Ovadia Y, Fertig E, Ren J, Nado Z, Sculley D, Nowozin S, Dillon J, Lakshminarayanan B, Snoek J (2019) Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems* 32.

Poursabzi-Sangdeh F, Goldstein DG, Hofman JM, Wortman Vaughan JW, Wallach H (2021) Manipulating and measuring model interpretability. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21 (New York, NY, USA: Association for Computing Machinery).

Qin X, Jiang Z (2019) The impact of ai on the advertising process: The chinese experience. *Journal of Advertising* 48(4):338–346.

Ren S, Chan HL, Siqin T (2020) Demand forecasting in retail operations for fashionable products: methods, practices, and real case study. *Annals of Operations Research* 291:761–777.

Rokach L, Maimon O, Shmueli E (2023) *Machine Learning for Data Science Handbook: Data Mining and Knowledge Discovery Handbook* (Springer Nature).

Shimodaira H (2000) Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference* 90(2):227–244.

Siemsen E, Aloysius J (2020) Supply chain analytics and the evolving work of supply chain managers. *Association for Supply Chain Management White Paper* .

Sniezek JA, Buckley T (1995) Cueing and cognitive conflict in judge-advisor decision making. *Organizational behavior and human decision processes* 62(2):159–174.

Snyder C, Keppler S, Leider S (2023) Algorithmic understanding under pressure: A behavioral study of algorithm reliance in large-scale personalized services. *Available at SSRN 4066823* .

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

Sugiyama M, Krauledat M, Müller KR (2007) Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research* 8(5).

Sun J, Zhang DJ, Hu H, Van Mieghem JA (2022) Predicting human discretion to adjust algorithmic prescription: A large-scale field experiment in warehouse operations. *Management Science* 68(2):846–865.

Tversky A, Kahneman D (1974) Judgment under uncertainty: Heuristics and biases. *Science* 185(4157):1124–1131.

Yeomans M, Shah A, Mullainathan S, Kleinberg J (2019) Making sense of recommendations. *Journal of Behavioral Decision Making* 32(4):403–414.

Yin M, Wortman Vaughan J, Wallach H (2019) Understanding the effect of accuracy on trust in machine learning models. *Proceedings of the 2019 CHI conference on human factors in computing systems*, 1–12.

Zhang Y, Liao QV, Bellamy RK (2020) Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 295–305.

Zougagh N, Charkaoui A, Echchatbi A (2020) Prediction models of demand in supply chain. *Procedia Computer Science* 177:462–467.

Appendix. E-Companion.

A. Proofs

Proof of Proposition 1 Expanding the square in the objective function of $NAdj$:

$$\mathbb{E}[\big(Y_i - (\hat{Y}_i^{alg} + \delta)\big)^2],$$

$$= \mathbb{E}[Y_i^2 - 2Y_i\hat{Y}_i^{alg} - 2Y_i\delta + (\hat{Y}_i^{alg})^2 + 2\hat{Y}_i^{alg}\delta + \delta^2].$$

By linearity:

$$= \mathbb{E}[Y_i^2 - 2Y_i\hat{Y}_i^{alg} + (\hat{Y}_i^{alg})^2] + 2\delta\,\mathbb{E}[\hat{Y}_i^{alg} - Y_i] + \delta^2.$$

The first order condition is

$$2\,\mathbb{E}[\hat{Y}_i^{alg} - Y_i] + 2\delta = 0,$$

and solving for δ gives $\delta^{NAdj} = \mathbb{E}[Y_i - \hat{Y}_i^{alg}]$. The second order condition confirms convexity. □

Proof of Proposition 2 To show the first part of the proposition, consider that $NAdj$ can be rewritten as

$$NAdj: \min_{\delta_I, \delta_O \in \mathbb{R}, \delta_I = \delta_O} \mathbb{E}[\big(Y_i - (\hat{Y}_i^{alg} + \delta_I)\big)^2 | \boldsymbol{X}_i \in \mathcal{D}_I] P(\boldsymbol{X}_i \in \mathcal{D}_I) + \mathbb{E}[\big(Y_i - (\hat{Y}_i^{alg} + \delta_O)\big)^2 | \boldsymbol{X}_i \in \mathcal{D}_O] P(\boldsymbol{X}_i \in \mathcal{D}_O).$$

This is identical to $M2Adj$ except with the additional constraint that $\delta_I = \delta_O$. Therefore, $NAdj$ and $M2Adj$ have the same decision variables, minimize the same objective, and the feasible region of $NAdj$ is a subset of the feasible region of $M2Adj$, which means that any feasible solution of $NAdj$ must be a feasible solution of $M2Adj$. This implies that the optimal solution of $NAdj$ is a feasible solution of $M2Adj$ with the same objective value; in turn, OPT^{NAdj} is an upper bound for OPT^{M2Adj}.

To show the comparison of δ^{NAdj} to δ_O^{M2Adj} and δ_I^{M2Adj} in the second part of the proposition, we separate the $M2Adj$ problem into two different problems, $M2Adj = M2Adj_I + M2Adj_O$, where

$$M2Adj_I: \min_{\delta_I \in \mathbb{R}} \mathbb{E}[\big(Y_i - (\hat{Y}_i^{alg} + \delta_I)\big)^2 | \boldsymbol{X}_i \in \mathcal{D}_I] P(\boldsymbol{X}_i \in \mathcal{D}_I), \tag{13}$$

$$M2Adj_O: \min_{\delta_O \in \mathbb{R}} \mathbb{E}[\big(Y_i - (\hat{Y}_i^{alg} + \delta_O)\big)^2 | \boldsymbol{X}_i \in \mathcal{D}_O] P(\boldsymbol{X}_i \in \mathcal{D}_O). \tag{14}$$

Since $P(\boldsymbol{X}_i \in \mathcal{D}_I)$ does not depend on δ_I, we can write

$$M2Adj_I: \; P(\boldsymbol{X}_i \in \mathcal{D}_I) \min_{\delta_I \in \mathbb{R}} \mathbb{E}[\big(Y_i - (\hat{Y}_i^{alg} + \delta_I)\big)^2 | \boldsymbol{X}_i \in \mathcal{D}_I]. \tag{15}$$

Note that the minimization in (15) is nearly identical to that in *NAdj*, except that it conditions on inlier instances. Following the same steps as in the proof of Proposition 1 yields

$$\delta_I^{M2Adj} = \mathbb{E}[Y_i - \hat{Y}_i^{alg} | \boldsymbol{X}_i \in \mathcal{D}_I].$$

Similarly, we find that

$$\delta_O^{M2Adj} = \mathbb{E}[Y_i - \hat{Y}_i^{alg} | \boldsymbol{X}_i \in \mathcal{D}_O].$$

By Assumption 1, we have

$$|\delta_I^{M2Adj}| < |\delta_O^{M2Adj}|.$$

Next, to compare δ^{NAdj} to δ_I^{M2Adj} and δ_O^{M2Adj}, we can write $\mathbb{E}[Y_i - \hat{Y}_i^{alg}]$ using the law of total expectation:

$$\delta^{NAdj} = \mathbb{E}[Y_i - \hat{Y}_i^{alg}] = \mathbb{E}[Y_i - \hat{Y}_i^{alg} | \boldsymbol{X}_i \in \mathcal{D}_O] P(\boldsymbol{X}_i \in \mathcal{D}_O) + \mathbb{E}[Y_i - \hat{Y}_i^{alg} | \boldsymbol{X}_i \in \mathcal{D}_I] P(\boldsymbol{X}_i \in \mathcal{D}_I)$$

$$\delta^{NAdj} = \delta_O^{M2Adj} P(\boldsymbol{X}_i \in \mathcal{D}_O) + \delta_I^{M2Adj} P(\boldsymbol{X}_i \in \mathcal{D}_I) \tag{16}$$

Notice that since $P(\boldsymbol{X}_i \in \mathcal{D}_O) + P(\boldsymbol{X}_i \in \mathcal{D}_I) = 1$, δ^{NAdj} is a convex combination of δ_O^{M2Adj} and δ_I^{M2Adj}. Since $P(\boldsymbol{X}_i \in \mathcal{D}_O)$ and $P(\boldsymbol{X}_i \in \mathcal{D}_I)$ are both positive, we have $\min(\delta_I^{M2Adj}, \delta_O^{M2Adj}) < \delta^{NAdj} < \max(\delta_I^{M2Adj}, \delta_O^{M2Adj})$.
□

B. Exclusion Criteria

B.1. Experiment I

One participant began the study while recruitment was ongoing but finished after recruitment had ended; we kept this participant per our pre-registered guidelines and thus retained a sample of $n = 301$. Participants were randomly assigned to conditions, and we had 124 participants in the *All-Inliers* condition, 71 in the *All-Outliers* condition and 106 in the *Mixed* condition. We acknowledge that the number of participants in each condition varied more than expected, given that we randomly assigned participants to each condition. First, we account for these different sample sizes across every statistical test in Sections 4.2.1 and 4.2.2. Second, we investigated to ensure that there wasn't a significant difference in dropout rate across conditions. Prolific reported that 31 participants started but did not complete the experiment, for which 10 had been assigned to the *All-Inliers* condition, 11 to the *Mixed* condition, 7 to the *All-Outliers* condition, and 3 unassigned likely due to returning their submissions very early.

We excluded 9 participants from our analyses who had an average absolute error larger than 35 in the Final Phase, because such large errors suggest that the participant did not understand they were entering adjustments instead of predictions or did not take the study seriously. We excluded 1 additional participant who made an adjustment that resulted in a negative demand prediction and 0 additional participants who made a prediction over 300 (unusually large). Finally, we excluded 5 participants who failed 2 or more comprehension checks. Specifically, we asked each participant to complete three separate comprehension checks. Two were multiple choice questions that tested their understanding of the prediction task, and the other was a numerical input question, testing if the participant understood how to make a final prediction via an adjustment to the algorithm's prediction. For the multiple choice questions, we

considered any incorrect answer as a failure; for the numerical input question, we considered it a failure if participants couldn't calculate the correct prediction in two separate tries. Altogether, we excluded 15 participants, or 5.0% of the dataset. All of these exclusion criteria were pre-registered. Of the 15 excluded participants, 5 had been assigned to the *All-Inliers* condition, 7 to the *Mixed* condition and 3 to the *All-Outliers* condition. These participants were still paid, just not included in our analyses.

The median time to complete the experiment among all non-excluded participants was 19.9 minutes. The average bonus payment among all participants was \$4.26 (SD = \$1.63).

B.2. Experiment 2

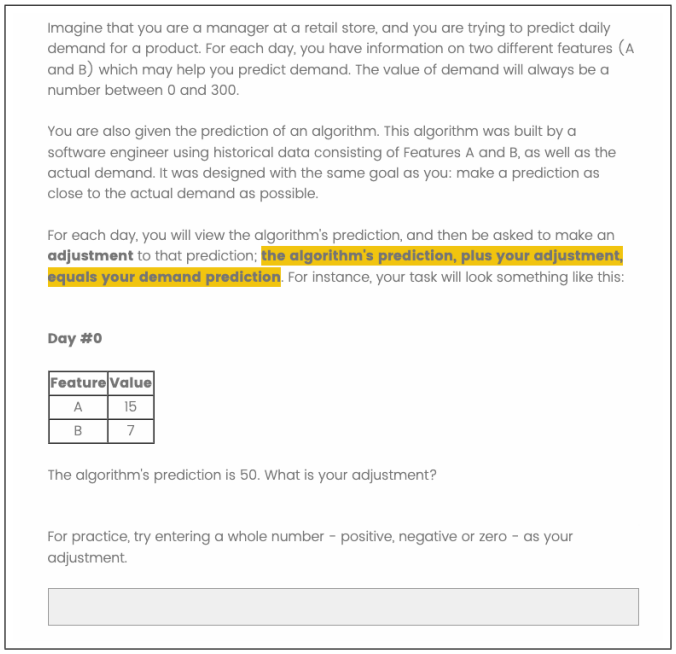
Participants were randomly assigned to conditions, and we had 129 in the *No Warnings or Endorsements* condition, 136 in the *Warnings Only* condition, and 185 in the *Warnings and Endorsements* condition. We acknowledge that the number of participants in each condition varied more than expected, given we randomly assigned participants to each condition. First, we account for these different sample sizes across every statistical test in Sections 5.2.1 and 5.2.2. Second, we investigated to ensure that there wasn't a significant difference in drop-out rates across conditions. Prolific reported that 35 participants started but did not complete the experiment, for which 12 had been assigned to the *No Warnings or Endorsements* condition, 9 to the *Warnings Only* condition, 6 to the *Warnings and Endorsements* condition, and 8 unassigned likely due to returning their submissions very early.

Overall, we excluded 3 participants who had average errors above 35, 1 participant who made a prediction over 300, and 13 additional participants who failed 2 or more of the comprehension checks. Altogether, we excluded 17 participants, or 3.8% of the sample size. All of these exclusion criteria were pre-registered. Of the 17 excluded participants, 5 had been assigned to the *No Warnings or Endorsements* condition, 5 to the *Warnings Only* condition, and 7 to the *Warnings and Endorsements* condition. Excluded participants were still paid, just not included in our analyses.

The median time to complete the experiment among all non-excluded participants was 20.6 minutes. The average bonus payment among all participants was \$3.88 (SD = \$1.30).

C. Experiment I: Participant Experience

C.1. Step 1: Instructions & Comprehension Checks



Imagine that you are a manager at a retail store, and you are trying to predict daily demand for a product. For each day, you have information on two different features (A and B) which may help you predict demand. The value of demand will always be a number between 0 and 300.

You are also given the prediction of an algorithm. This algorithm was built by a software engineer using historical data consisting of Features A and B, as well as the actual demand. It was designed with the same goal as you: make a prediction as close to the actual demand as possible.

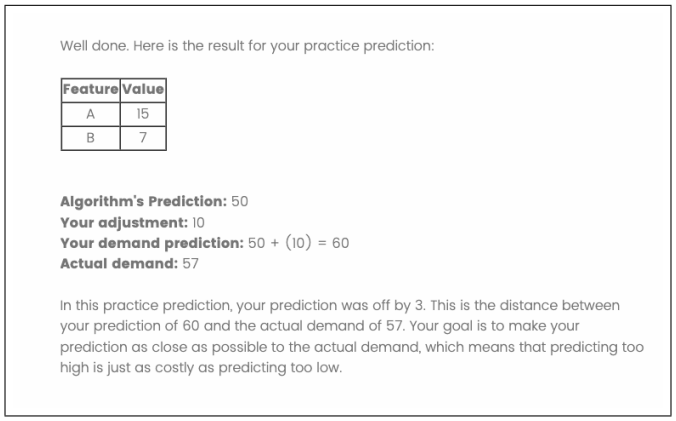
For each day, you will view the algorithm's prediction, and then be asked to make an **adjustment** to that prediction; **the algorithm's prediction, plus your adjustment, equals your demand prediction**. For instance, your task will look something like this:

Day #0

Feature	Value
A	15
B	7

The algorithm's prediction is 50. What is your adjustment?

For practice, try entering a whole number - positive, negative or zero - as your adjustment.

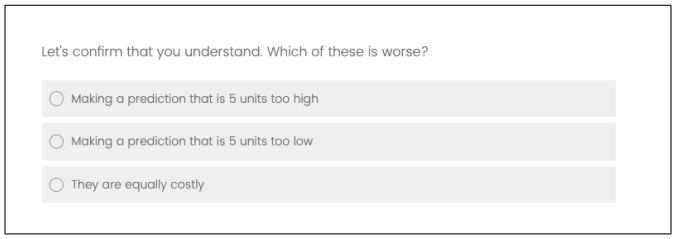


Well done. Here is the result for your practice prediction:

Feature	Value
A	15
B	7

Algorithm's Prediction: 50
Your adjustment: 10
Your demand prediction: 50 + (10) = 60
Actual demand: 57

In this practice prediction, your prediction was off by 3. This is the distance between your prediction of 60 and the actual demand of 57. Your goal is to make your prediction as close as possible to the actual demand, which means that predicting too high is just as costly as predicting too low.



Let's confirm that you understand. Which of these is worse?

- ○ Making a prediction that is 5 units too high
- ○ Making a prediction that is 5 units too low
- ○ They are equally costly

Of course, for this practice example, you did not have much useful information when making your adjustment! Don't worry: you will be able to view feature and demand data for previous days to help understand how to predict future demand.

After familiarizing yourself with the historical data, you will make daily demand adjustments in two stages: 15 demand adjustments in the **Initial Prediction Stage** and 20 demand adjustments in the **Final Prediction Stage**. For each day, you will be given the values of Features A and B, as well as the algorithm's prediction, and will be asked to make an adjustment.

Please make your adjustments to reflect your best guess about what the demand will be. You will receive a bonus between $0 and $7 in the Final Prediction Stage: the more accurate your predictions, the larger your bonus will be. To see the full formula for your bonus calculation in the Final Prediction Stage, click below.

Bonus Formula

For the Final Prediction Stage, we will calculate your squared error as: (your prediction - the actual demand)^2. We will take the average over your 20 predictions to get your average prediction squared error. Your final bonus is $7.00 - 0.20*sqrt(your average prediction squared error). If this number is negative, you will receive a bonus of $0.

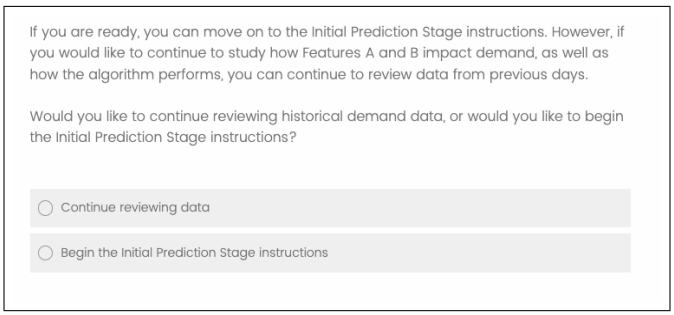
Let's confirm that you understand. What will you have access to when making adjustments in the Initial Prediction Stage?

- ○ Nothing
- ○ Only Feature A
- ○ Feature B and the algorithm's prediction
- ○ Feature A, Feature B, and the algorithm's prediction

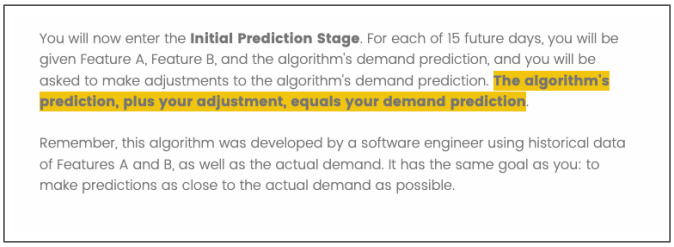
C.2. Step 2: Review Historical Data

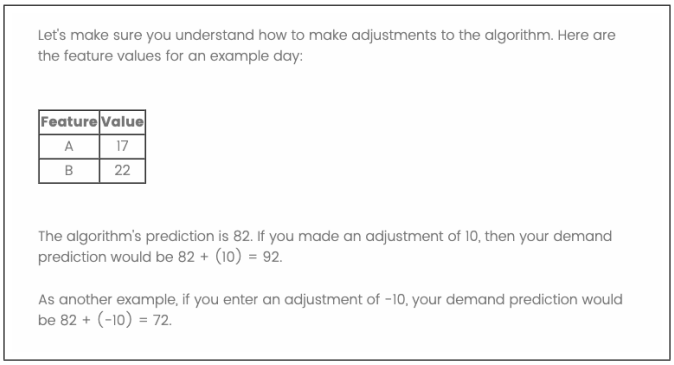
Please review the following demand data for 15 previous days. For each day, you can see the values of Features A and B, as well as the actual demand. You can also see the predictions of the algorithm, as well as the algorithm's error. Spend some time familiarizing yourself with this information, as it should help you make adjustments for future days.

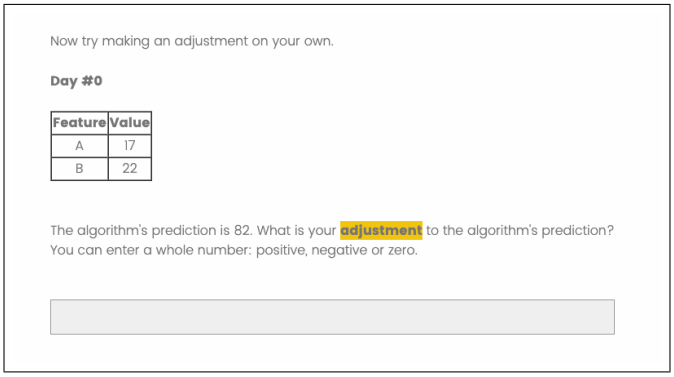
Feature A	Feature B	Actual Demand	Algorithm's Demand Prediction	Algorithm's Prediction Error
3	14	48	52	4
12	15	60	63	3
35	17	133	90	43
17	3	35	44	9
44	21	160	107	53
2	17	66	57	9
3	20	60	64	4
29	18	119	86	33
14	12	57	59	2
36	14	127	85	42
4	4	23	33	10
20	3	53	47	6
33	13	103	80	23
9	22	71	74	3
14	6	55	47	8

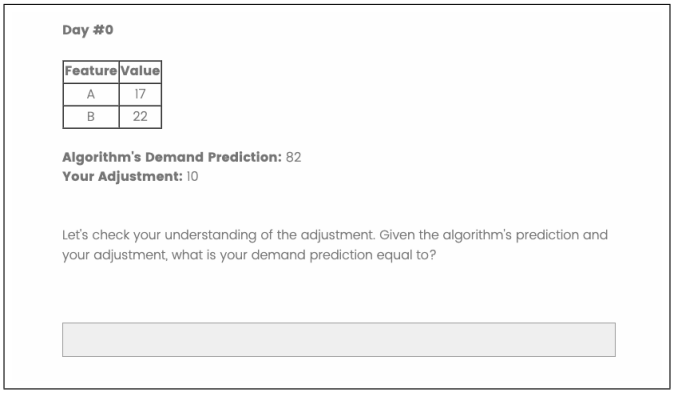


C.3. Step 3: Practice Phase







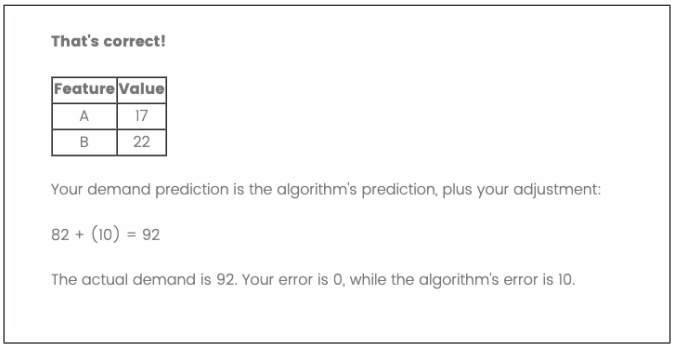


Day #0

Feature	Value
A	17
B	22

Algorithm's Demand Prediction: 82
Your Adjustment: 10

Let's check your understanding of the adjustment. Given the algorithm's prediction and your adjustment, what is your demand prediction equal to?



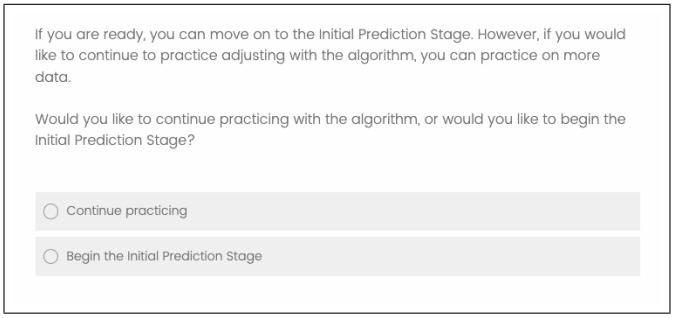
That's correct!

Feature	Value
A	17
B	22

Your demand prediction is the algorithm's prediction, plus your adjustment:

82 + (10) = 92

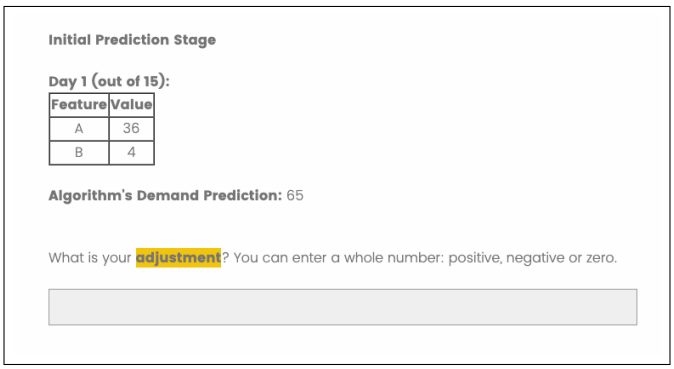
The actual demand is 92. Your error is 0, while the algorithm's error is 10.



If you are ready, you can move on to the Initial Prediction Stage. However, if you would like to continue to practice adjusting with the algorithm, you can practice on more data.

Would you like to continue practicing with the algorithm, or would you like to begin the Initial Prediction Stage?

- ◯ Continue practicing
- ◯ Begin the Initial Prediction Stage



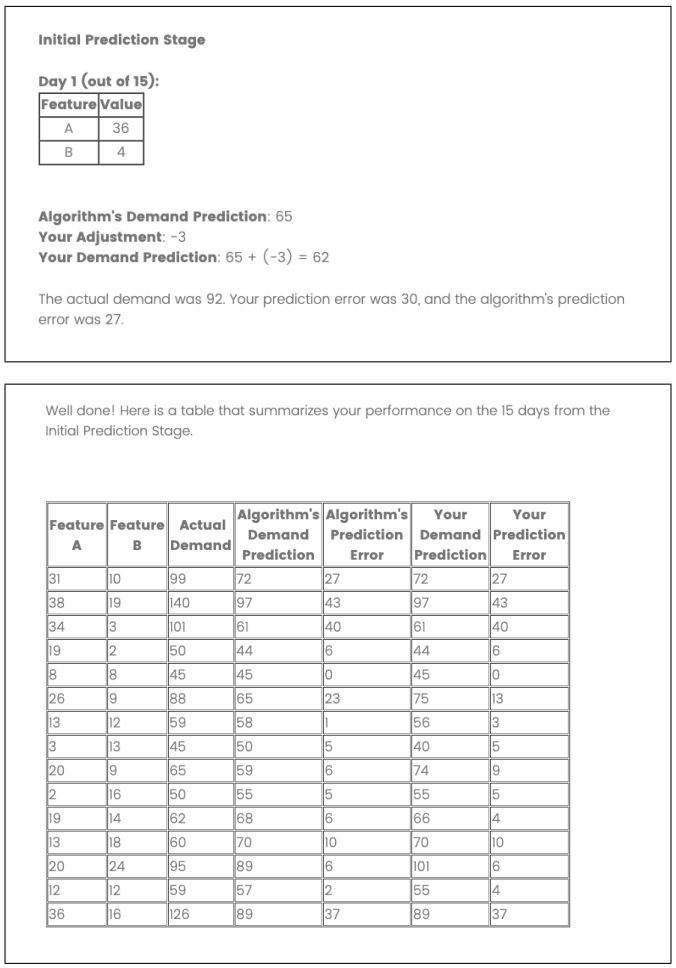
Initial Prediction Stage

Day 1 (out of 15):

Feature	Value
A	36
B	4

Algorithm's Demand Prediction: 65

What is your **adjustment**? You can enter a whole number: positive, negative or zero.



Initial Prediction Stage

Day 1 (out of 15):

Feature	Value
A	36
B	4

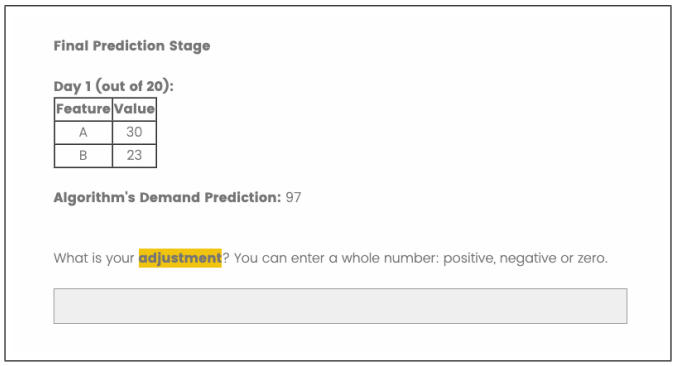
Algorithm's Demand Prediction: 65
Your Adjustment: -3
Your Demand Prediction: 65 + (-3) = 62

The actual demand was 92. Your prediction error was 30, and the algorithm's prediction error was 27.

Well done! Here is a table that summarizes your performance on the 15 days from the Initial Prediction Stage.

Feature A	Feature B	Actual Demand	Algorithm's Demand Prediction	Algorithm's Prediction Error	Your Demand Prediction	Your Prediction Error
31	10	99	72	27	72	27
38	19	140	97	43	97	43
34	3	101	61	40	61	40
19	2	50	44	6	44	6
8	8	45	45	0	45	0
26	9	88	65	23	75	13
13	12	59	58	1	56	3
3	13	45	50	5	40	5
20	9	65	59	6	74	9
2	16	50	55	5	55	5
19	14	62	68	6	66	4
13	18	60	70	10	70	10
20	24	95	89	6	101	6
12	12	59	57	2	55	4
36	16	126	89	37	89	37

C.4. Step 4: Final Phase



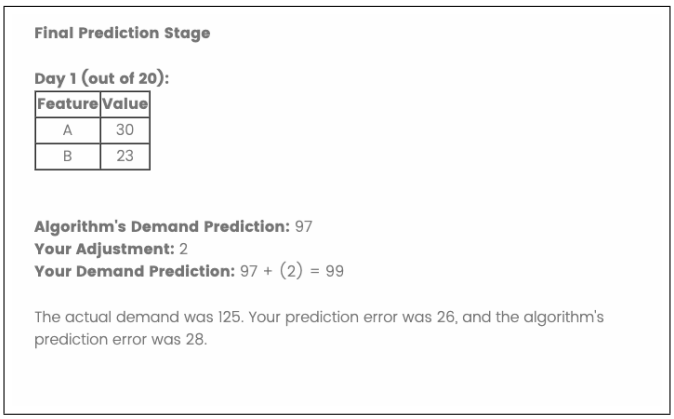
Final Prediction Stage

Day 1 (out of 20):

Feature	Value
A	30
B	23

Algorithm's Demand Prediction: 97

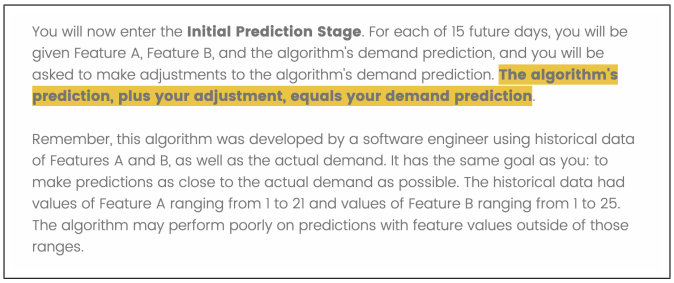
What is your **adjustment**? You can enter a whole number: positive, negative or zero.

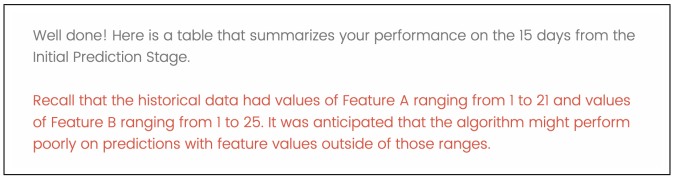


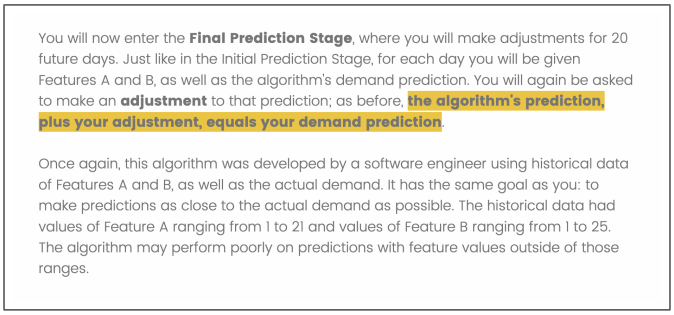
D. Experiment II: Participant Experience

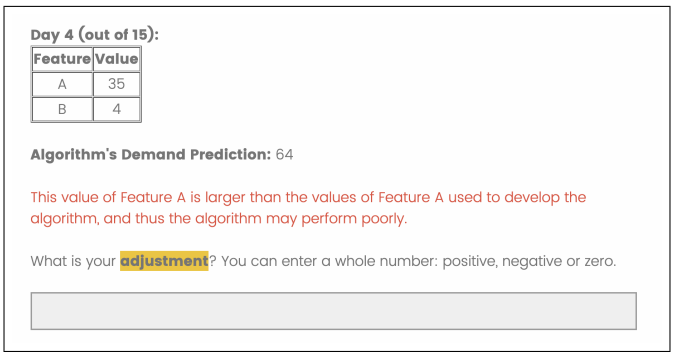
D.1. *Warnings Only* Condition

The *Warnings Only* condition is identical to the *Mixed* condition except for some important differences: extra text in the Practice and Final Phase introductions, and warnings during adjustments on outliers in both the Practice and Final Phases.



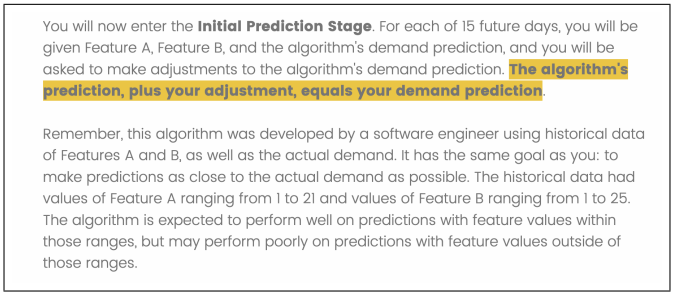


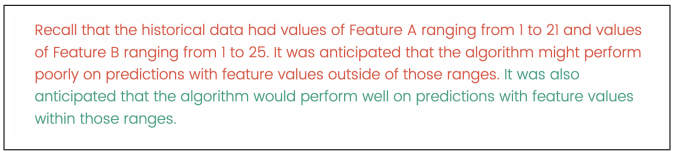


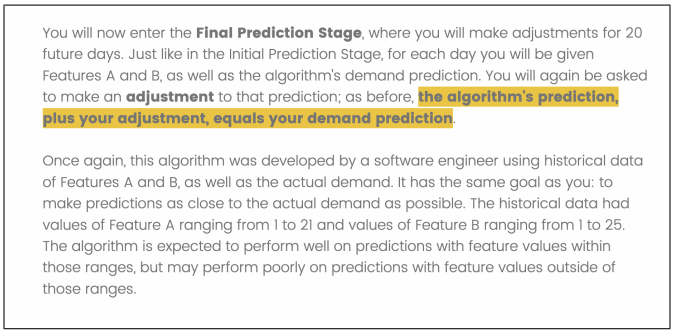


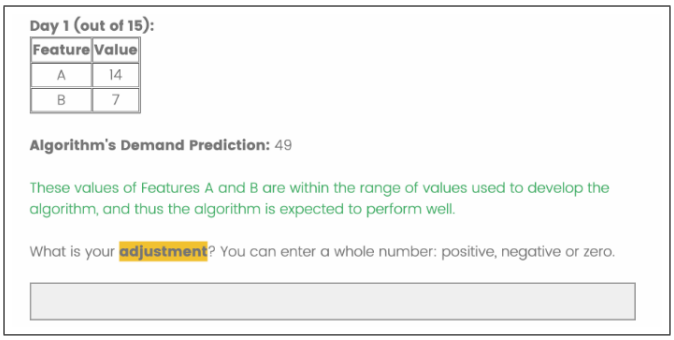
D.2. *Warnings and Endorsements* Condition

The *Warnings and Endorsements* condition adds additional text — including endorsements on inliers — to the changes made in the *Warnings Only* condition. Again, compared to the *Mixed* condition, there is extra text in the Practice and Final Phase introductions, warnings on outliers and endorsements on inliers during adjustments in both the Practice and Final Phases.









E. Robustness Checks

To test the robustness of our results in Experiment I, we run regressions on our critical outcomes while controlling for a variety of demographic variables. Tables 2 and 3 map the median absolute adjustment on inliers and outliers, respectively, as outcome variables. We see that the *Mixed* condition has significantly larger (smaller) absolute adjustments on inliers (outliers), even when controlling for covariates like age, gender and education level. Table 4 maps $MedDOP_j^O$ as the outcome variable (the inlier table is excluded because $MedA_j^I = MedDOP_j^I$). Once again, we see that the *Mixed* condition has a significantly larger median absolute deviation from the optimal prediction in all of the regressions.

	Education	Gender	Age	Statistics Class
Intercept	8.46^{***}	8.19^{***}	7.33^{***}	8.61^{***}
	(0.93)	(0.65)	(0.90)	(0.60)
All-Inliers	-4.89^{***}	-5.05^{***}	-4.90^{***}	-4.87^{***}
	(0.71)	(0.72)	(0.72)	(0.71)
Bachelor's degree	-1.59			
	(1.02)			
Doctoral degree	-1.37			
	(3.15)			
High school graduate	0.46			
	(1.07)			
Master's degree	-1.63			
	(1.50)			
Male		-0.67		
		(0.73)		
Other		-0.77		
		(2.42)		
26-35			0.20	
			(1.04)	
36-45			0.19	
			(1.11)	
46-55			2.64^{*}	
			(1.31)	
56-65			-0.17	
			(1.50)	
66-75			1.14	
			(2.31)	
Have taken a Statistics class				-1.81^{*}
				(0.71)
R^2	0.21	0.19	0.20	0.21
Adj. R^2	0.19	0.18	0.18	0.20
Num. obs.	218	218	218	218

$MedA_j^I$ (Experiment I) $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 2 Default categories are: *Mixed* (control), Associate's Degree, Female, Age 18 - 25, and have not taken a statistics class.

	Education	Gender	Age	Statistics Class
Intercept	18.04^{***}	16.09^{***}	18.33^{***}	16.46^{***}
	(1.77)	(1.24)	(1.78)	(1.16)
All-Outliers	15.30^{***}	15.29^{***}	14.96^{***}	14.91^{***}
	(1.51)	(1.52)	(1.53)	(1.52)
Bachelor's degree	1.56			
	(2.04)			
Doctoral degree	0.81			
	(6.94)			
High school graduate	-2.92			
	(2.12)			
Master's degree	-3.01			
	(2.69)			
Male		2.58		
		(1.51)		
Other		2.09		
		(4.43)		
26-35			1.03	
			(2.17)	
36-45			-2.85	
			(2.30)	
46-55			-0.50	
			(2.50)	
56-65			-4.71	
			(3.49)	
66-75			-2.40	
			(4.28)	
Have taken a Statistics class				2.31
				(1.50)
R^2	0.41	0.39	0.40	0.39
Adj. R^2	0.39	0.38	0.38	0.38
Num. obs.	167	167	167	167

$MedA_j^O$ (Experiment I) $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 3 Default categories are: *Mixed* (control), Associate's Degree, Female, Age 18 - 25, and have not taken a statistics class.

	Education	Gender	Age	Statistics Class
Intercept	19.56^{***}	21.15^{***}	17.91^{***}	21.02^{***}
	(1.56)	(1.08)	(1.56)	(1.00)
All-Outliers	-11.66^{***}	-11.71^{***}	-11.50^{***}	-11.21^{***}
	(1.34)	(1.32)	(1.34)	(1.32)
Bachelor's degree	-1.43			
	(1.81)			
Doctoral degree	-2.98			
	(6.13)			
High school graduate	1.71			
	(1.88)			
Master's degree	1.66			
	(2.38)			
Male		-2.81^{*}		
		(1.32)		
Other		-2.57		
		(3.86)		
26-35			1.11	
			(1.90)	
36-45			2.99	
			(2.02)	
46-55			1.56	
			(2.20)	
56-65			5.79	
			(3.06)	
66-75			2.34	
			(3.76)	
Have taken a Statistics class				-3.16^{*}
				(1.30)
R^2	0.33	0.33	0.34	0.34
Adj. R^2	0.31	0.32	0.31	0.33
Num. obs.	167	167	167	167

$MedDOP_j^O$ (Experiment I) $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 4 Default categories are: *Mixed* (control), Associate's Degree, Female, Age 18 - 25, and have not taken a statistics class.

We run similar regressions for Experiment II. Tables 5 and 6 map the median absolute adjustments on inliers and outliers as outcomes, while Table 7 maps the $MedDOP_j^O$ as the outcome. Once again, our results persist even when controlling for a variety of demographic covariates. The *No Warnings or Endorsements* condition has a significantly larger median absolute adjustment and median absolute deviation from the optimal prediction on inliers compared to the *Warnings Only* condition, which in turn has a significantly larger median absolute adjustment and median absolute deviation from the optimal prediction on inliers compared to the *Warnings and Endorsements* condition. Further, the *No Warnings or Endorsements* condition has a significantly smaller (larger) median absolute adjustment (median absolute deviation from the optimal prediction) on outliers compared to the *Warnings Only* condition, which in turn has a non-significantly smaller (larger) median absolute adjustment (median absolute deviation from the optimal prediction) on outliers compared to the *Warnings and Endorsements* condition.

	Education	Gender	Age	Statistics Class
Intercept	4.12^{***}	4.63^{***}	4.24^{***}	4.94^{***}
	(0.67)	(0.48)	(0.65)	(0.45)
No Warnings or Endorsements	2.51^{***}	2.45^{***}	2.27^{***}	2.43^{***}
	(0.57)	(0.58)	(0.57)	(0.57)
Warnings and Endorsements	-1.54^{**}	-1.53^{**}	-1.59^{**}	-1.53^{**}
	(0.53)	(0.53)	(0.53)	(0.52)
Bachelor's degree	0.33			
	(0.68)			
Doctoral degree	-1.78			
	(1.44)			
High school graduate	0.92			
	(0.69)			
Master's degree	-0.32			
	(0.88)			
Male		-0.24		
		(0.45)		
Other		-0.12		
		(1.66)		
Prefer not to say		-1.76		
		(2.68)		
26-35			-0.05	
			(0.66)	
36-45			0.22	
			(0.70)	
46-55			0.06	
			(0.82)	
56-65			2.57^{**}	
			(0.84)	
66-75			-0.76	
			(1.38)	
Have taken a Statistics class				-0.90^{*}
				(0.44)
R^2	0.13	0.12	0.14	0.12
Adj. R^2	0.12	0.11	0.13	0.12
Num. obs.	433	433	433	433

$MedA_j^I$ (Experiment II) $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 5 Default categories are: *Warnings Only*, Associate's Degree, Female, Age 18 - 25, and have not taken a statistics class.

	Education	Gender	Age	Statistics Class
Intercept	24.88^{***}	23.98^{***}	28.60^{***}	25.82^{***}
	(1.56)	(1.09)	(1.50)	(1.05)
No Warnings or Endorsements	-7.31^{***}	-7.15^{***}	-7.18^{***}	-7.31^{***}
	(1.33)	(1.32)	(1.31)	(1.33)
Warnings and Endorsements	0.74	0.59	0.50	0.73
	(1.23)	(1.21)	(1.22)	(1.22)
Bachelor's degree	0.96			
	(1.58)			
Doctoral degree	−0.10			
	(3.36)			
High school graduate	1.42			
	(1.61)			
Master's degree	0.95			
	(2.05)			
Male		3.44^{***}		
		(1.03)		
Other		−0.52		
		(3.80)		
Prefer not to say		−0.08		
		(6.11)		
26-35			−1.42	
			(1.53)	
36-45			-3.26^{*}	
			(1.63)	
46-55			-4.47^{*}	
			(1.89)	
56-65			-6.41^{***}	
			(1.93)	
66-75			−5.53	
			(3.19)	
Have taken a Statistics class				0.02
				(1.02)
R^2	0.10	0.12	0.13	0.10
Adj. R^2	0.09	0.11	0.12	0.09
Num. obs.	433	433	433	433

$MedA_j^O$ (Experiment II) $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 6 Default categories are: *Warnings Only*, Associate's Degree, Female, Age 18 - 25, and have not taken a statistics class.

	Education	Gender	Age	Statistics Class
Intercept	13.57^{***}	14.27^{***}	10.82^{***}	13.10^{***}
	(1.25)	(0.88)	(1.20)	(0.85)
No Warnings or Endorsements	5.90^{***}	5.87^{***}	5.73^{***}	5.94^{***}
	(1.07)	(1.06)	(1.04)	(1.07)
Warnings and Endorsements	-1.03	-0.83	-0.75	-1.00
	(0.98)	(0.97)	(0.97)	(0.98)
Bachelor's degree	-1.06			
	(1.27)			
Doctoral degree	1.48			
	(2.69)			
High school graduate	-0.71			
	(1.29)			
Master's degree	-0.10			
	(1.64)			
Male		-2.64^{**}		
		(0.83)		
Other		2.36		
		(3.05)		
Prefer not to say		-0.74		
		(4.91)		
26-35			0.76	
			(1.22)	
36-45			2.36	
			(1.29)	
46-55			2.54	
			(1.50)	
56-65			6.51^{***}	
			(1.54)	
66-75			5.84^{*}	
			(2.53)	
Have taken a Statistics class				-0.32
				(0.82)
R^2	0.11	0.13	0.16	0.11
Adj. R^2	0.10	0.12	0.15	0.10
Num. obs.	433	433	433	433

$MedDOP_j^O$ (Experiment II) $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 7 Default categories are: *Warnings Only*, Associate's Degree, Female, Age 18 - 25, and have not taken a statistics class.