Overcoming Algorithm Aversion: Supplementary Materials

High School Student Forecasting Model

This section describes the model used in Studies 1, 3, 4, S1, S2 and S3 to predict the percentile ranks of high school students on a standardized math test. The model was an ordinary least squares regression built using data from the High School Longitudinal Study of 2009 (HSLS:09). The model was built on all students except for a randomly selected subset of 1,000 students. We randomly selected 50 of the 1,000 students that were excluded when we built the model to use as targets for the prediction task.

The dependent variable was student's theta score on a standardized math that they took in their senior year (2011) converted to percentiles (0-100). The model consisted of nine predictors, including students': race (categorical: e.g. White, Asian), socioeconomic status quintile (categorical: First to Fifth), desired occupation at age 30 (categorical: e.g. Management Occupations, Architecture and Engineering Occupations), own prediction of their highest degree (categorical: e.g. Complete Associate's degree, Complete Master's degree), region of country (categorical: Northeast, Midwest, South, West), number of times they had taken the PSAT (categorical: Never, Once, Twice, 3 or more time, Don't know what this is), number of friends not going to college (categorical: None of them to All of them), favorite subject (categorical: e.g. English, Science, Art), and status having taken any AP test (categorical: Yes, No).

On average, the model's predicted percentiles were 17.49 percentiles away from students' realized percentiles for the 1,000 students who were excluded when we built the model, and 18.14 percentiles away from students' realized percentiles for the 50 students who were used for the prediction task. The model's predictions were correlated at .66 with student's actual percentiles, and the model's R-squared was 0.42.

Airline Passengers Model

The following text is copied verbatim from the supplemental materials of Dietvorst, Simmons, and Massey (2015). This section describes the model used in Study 2 to predict the rank of the 50 U.S. states in terms of the number of airline passengers who departed from those states in 2011. The model was an ordered logistic regression.

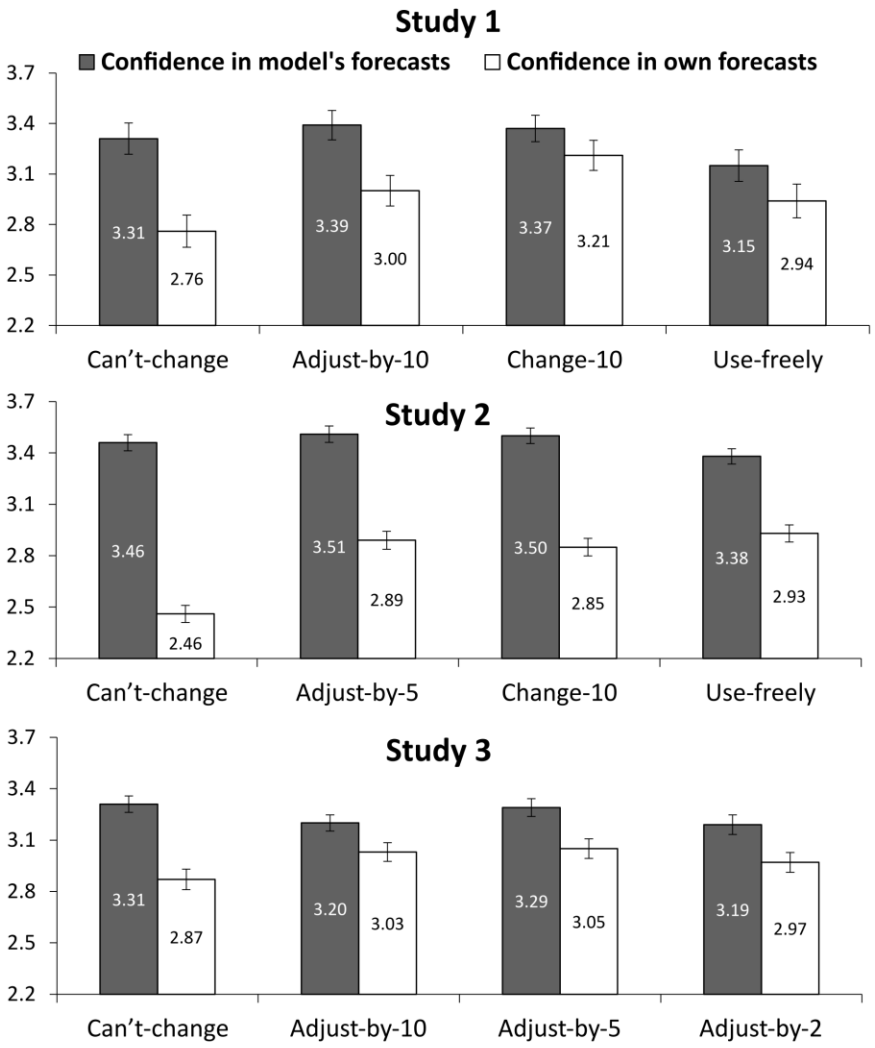
The dependent variable in the model was the ranks of the 50 U.S. states in 2006, 2007, 2008, 2009, and 2010 in terms of the number of airline passengers who had departed from those states. The predictors were each states': number of major airports, 2010 census population rank (1-50), rank in terms of number of counties (1-50), rank in terms of median household income in 2008 (1-50), and rank in terms of domestic travel expenditure in 2009 (1-50). On average, the model's predicted ranks in 2011 were 4.32 ranks away from states' actual ranks. The model's predictions were correlated at .92 with states' actual ranks, and the model's pseudo R-squared was 0.26.

Confidence Measures and Performance Estimates from Studies 1-3

In each experiment, participants rated their confidence in the model's forecasts and their own forecasts on 5-point scales (1=*none*; 5=*a lot*) towards the end of the survey. Also, participants estimated the average error of their own forecasts and the model's forecasts. The following two figures show this data for Studies 1, 2, and 3.

Figure S1.

Participants' confidence in the model's forecasts and their own forecasts

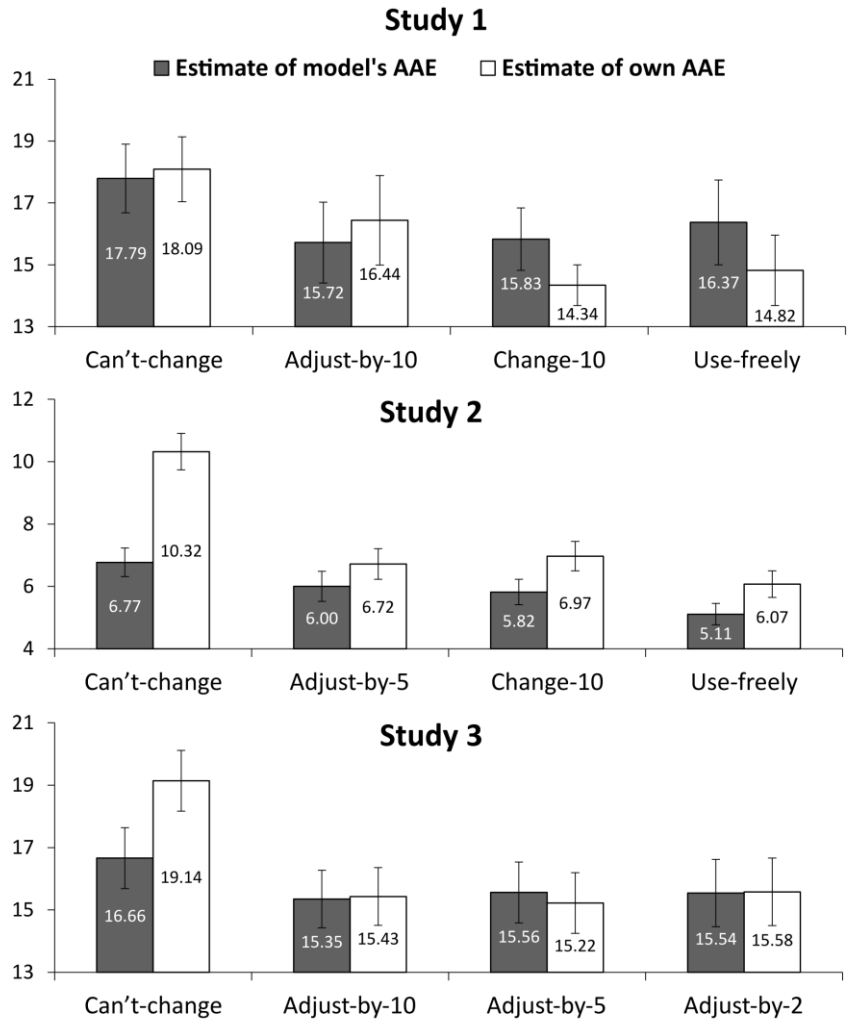


Note: Errors bars indicate ±1 standard error.

Confidence measures. Participants in Studies 1-3 did not have many consistent patterns in their confidence ratings of the model's forecasts and their own forecasts. The most consistent pattern is that participants in all conditions had at least marginally more confidence in the model's forecasts than their own forecasts, *t*'s >= 1.62, *p*'s <= .109.

Figure S2.

Participants' estimates of the model's average absolute error and their own average absolute error



Note: Errors bars indicate ±1 standard error.

Performance estimates. Participants in Studies 1-3 did not show many consistent patterns in their performance estimates. Participants in the can't-change conditions thought that their absolute error was higher than participants in the other three conditions, who saw the model's forecasts if they chose to use it, t's >= 2.26, p's <= .024. This could be because these participants never received the model's input when making their own forecasts. Also, participants in a few conditions thought that the model's forecasts were better than their own, and participants didn't believe that their own forecasts were significantly better than the model's in any condition.

Study S1

Methods

Overview. This study is a replication of Study 1 that only includes the use-freely and can't-change conditions.

Participants. We ran Study S1 in our university's behavioral lab. Participants who showed up to participate in the experiment received $10 for completing one hour of experiments, of which ours was a 20 minute portion. Participants could earn up to a $5 bonus from our study depending on their forecasting performance. The behavioral lab had the goal of recruiting at least 200 participants for our experiment. Ten participants exited the study before completing their forecasts, leaving us with a final sample of 240 participants who completed their forecasts. This sample averaged 24 years of age and was 59% female.

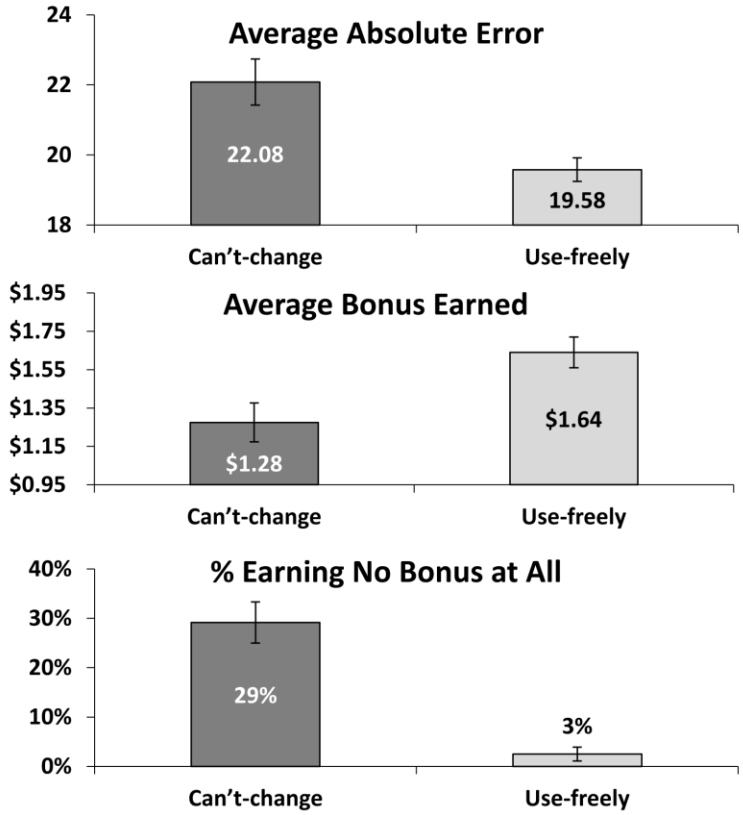
Procedures. This study's procedure was the same as study 1's except for three changes. First, we did not include the adjust-by-10 and change-10 conditions. Second, on the page after participants typed the sentence to ensure that they understood the payment rule, we did not have participants type another sentence to ensure that they understood the procedures for their condition. Third, participants did not learn their performance relative to other participants from their condition and answer questions based on this information as they did in the exploratory question included at the end of Study 1.

Results

Choosing to use the model. Participants in the can't-change condition chose to use the model 55.83% of the time. Participants in the use-freely condition gave the model's forecasts a significant amount of weight when making their own forecasts, as they did in Study 1. These participants provided forecasts that were 7.90 percentiles away from the model's on average while participants in the can't-change condition, who all made their own forecasts without seeing the model's, provided forecasts that were 18.21 away from the model's on average, $t(238) = 15.42$, $p < .001$.

Figure S4.

Study S1: Participants who could modify the model's forecasts performed better than participants who could not modify the model's forecasts.



Note: Errors bars indicate ±1 standard error.

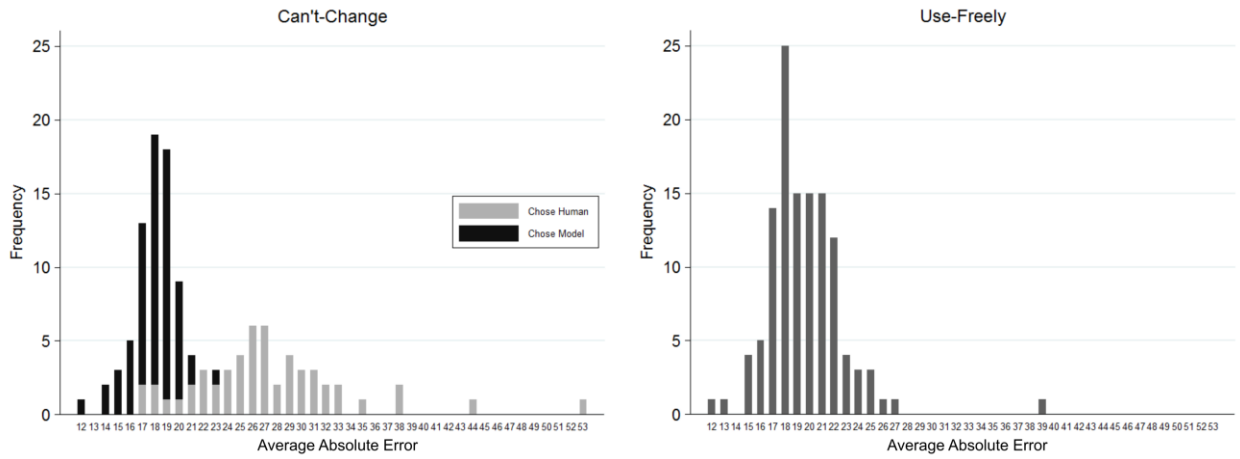
Forecasting performance. As shown in Figure S4, participants in the use-freely condition, who had the option to adjust the model's forecasts, outperformed those in the can't-change condition. Participants' forecasts in the can't-change condition were less accurate, $t(238) = -3.81$, $p < .001$, and earned them smaller bonuses, $t(238) = 3.63$, $p < .001$, than participants' forecasts in the use-freely condition.

Figure S5 displays the distribution of participants' performance by condition. Like Study 1, reliance on the model was strongly associated with better performance. Indeed, failing to choose to use the model was much more likely to result in very large average errors (and bonuses of $0). Also, participants in the

can't-change condition performed worse precisely because they were less likely to use the model, and not because their forecasting ability was worse. Additionally, participants' use of the model in the use-freely condition seems to have prevented them from making very many large errors.

Figure S5.

Study S1: The distribution of participants' average absolute errors by condition and whether or not they chose to use the model's forecasts.



Discussion. Similar to Study 1, participants who could modify the model's forecasts performed better and earned more money than participants who could not modify the model's forecasts. Also, once again, participants who could use the model's forecasts freely seemed to anchor on the model's forecasts, which improved their performance by reducing their chances of making large errors.

Study S2

Methods

Overview. This study is a conceptual replication of Study 3; however, instead of including adjust-by-10, adjust-by-5, and adjust-by-2 conditions, we included change-10, change-5, and change-2 conditions. The goal of this study was to see if participants would be sensitive to the number of times that they could overrule the model's forecasts. Like Study 3, we found that participants were relatively

insensitive to how much they could modify the model's forecasts; however, participants' increased likelihood of choosing the model did not consistently translate into better performance.

Participants. We ran Study S2 with participants from Amazon Mechanical Turk. Participants earned $1 for completing the study and could earn up to an additional $0.50 depending on their forecasting performance. We decided in advance to recruit 600 participants (150 per condition). We did not include a reading check in this study. Eighty-five participants quit the survey before completing their forecasts, leaving us with a final sample of 612 participants who completed their forecasts. This sample averaged 35 years of age and was 47% female.

Procedures. This study's procedure was the same as Study 1's except for three changes. First, we did not include the use-freely or adjust-by-10 conditions and included two new conditions instead (participants were still assigned to the can't-change and change-10 conditions). These two new conditions were the same as the change-10 condition, but participants could change the model's forecasts fewer times. In the change-5 condition participants decided to use either the model's forecasts or their own forecasts, but they could change five of the model's 20 forecasts by any amount if they chose to use the model. In the change-2 condition participants decided to use either the model's forecasts or their own forecasts, but they could change two of the model's 20 forecasts by any amount if they chose to use the model. Second, we recruited participants from the Amazon Mechanical Turk instead our university's behavioral lab. Third, we used the same bonus scheme as Study 3. Participants were paid a $0.50 bonus if their official forecasts were within five percentiles of students' actual percentiles on average and this bonus decreased by $0.10 for each additional five percentiles of average error in participants' forecasts.

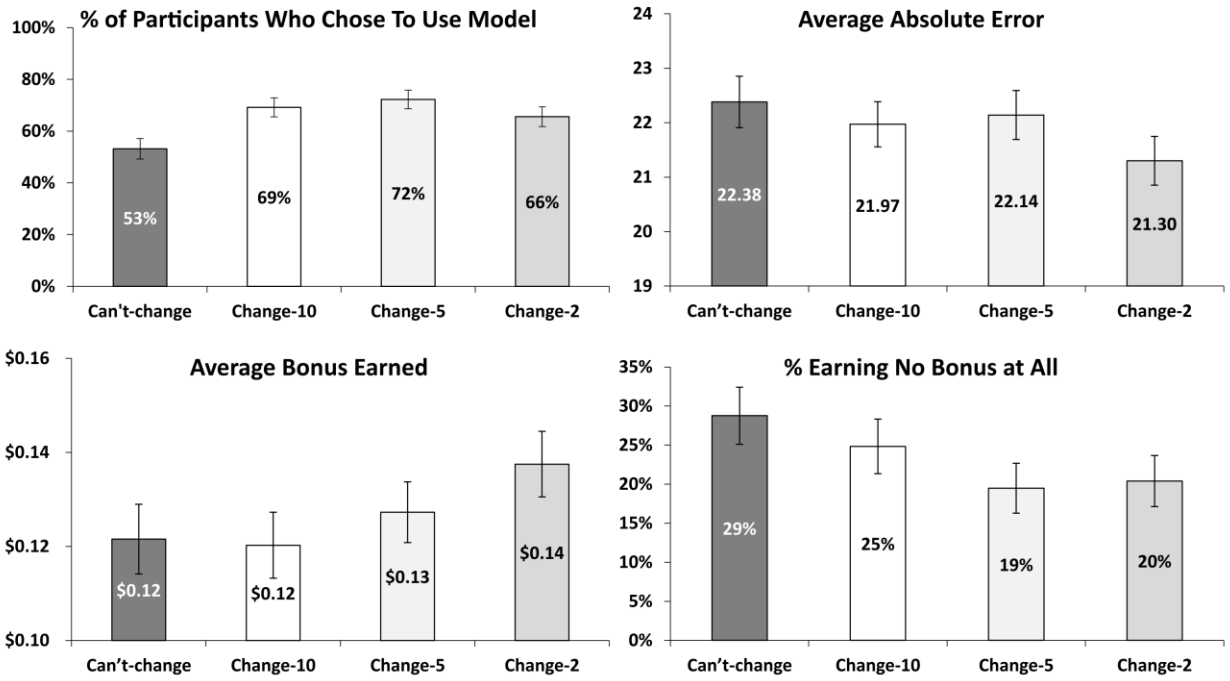
Results.

Choosing to use the model. Consistent with the results of Study 3, participants who had the option to modify the model's forecasts were significantly more likely to choose to use the model (see Figure S6). Whereas only 53% of participants in the can't-change condition chose to use the model's

forecasts, 69% of participants in the change-X conditions chose to the model, $\chi^2(1, N = 626) = 13.05$, $p < .001$. Also, similar to the results of Study 3, we found that participants' decision to use the model in the change-X conditions did not depend on how much they were able to adjust the model: 69%, 72%, and 66% chose to the model in the change-10, change-5, and change-2 conditions. These three conditions did not differ significantly, $\chi^2(2, N = 468) = 1.61$, $p = .447$. Once again, we cannot reject the possibility that participants may have been slightly sensitive to the number of changes that they could make to the model, but we can conclude that their willingness to use the model was not *detectably* altered by imposing a fivefold restriction on the number of adjustments that they could make.

Figure S6

Study S2: Participants who could restrictively modify the model's forecasts were more likely to choose to use the model.



Note: Errors bars indicate ±1 standard error.

Forecasting performance. Unlike Study 3, participants who were given the option to adjust the model's forecasts did not consistently perform better than those who were not (see Figure S6). Only

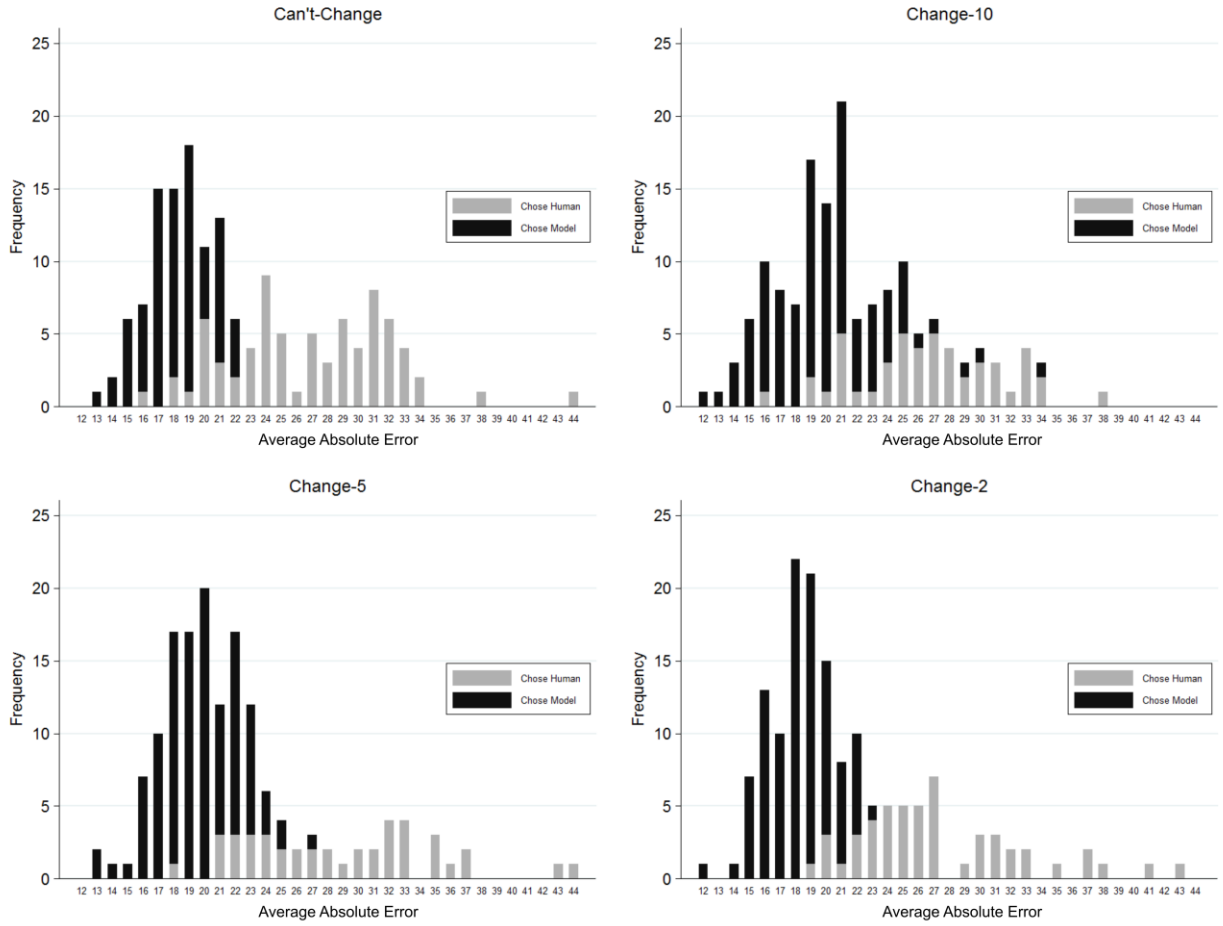
participants in the change-2 condition made marginally smaller errors, $t(303) = -1.66$, $p = .099$, and earned marginally more money, $t(303) = 1.56$, $p = .119$, than participants in the can't-change condition.[1] These results are consistent with the results of Study 2, in which participants in the change-10 condition did not outperform the can't-change condition.

Figure S7 displays the distribution of participants' performance by condition. Reliance on the model was associated with better performance; however, many participants in the change-10 and change-5 conditions performed poorly even though they chose to use the model. Participants in the can't-change condition who chose to use the model performed significantly better than participants in the change-10, $t(182) = 3.61$, $p < .001$, and change-5, $t(189) = 4.29$, $p < .001$, conditions who chose to use the model. In other words, participants who adjusted the model's forecasts in the change-10 and change-5 conditions made the model's forecasts significantly worse. These results are consistent with Study 2, where participants in the change-10 condition who chose to use the model made large adjustments to the model's forecasts, made the model's forecasts worse, and did not outperform the can't-change condition as a result.

[1] Participants in the can't-change condition had similar absolute average errors to participants in the change-10, $t(304) = -0.65$, $p = .517$, and change-5, $t(305) = -0.37$, $p = .710$, conditions. As a result, participants in the can't-change condition earned similar bonuses to participants in the change-10, $t(304) = 0.13$, $p = .898$, and change-5, $t(305) = 0.58$, $p = .580$, conditions.

Figure S7

Study S2: The distribution of participants' average absolute errors by condition and whether or not they chose to use the model's forecasts.



Discussion. We found that participants were more likely to use the model when they could modify its forecasts in this study. Also, consistent with Study 3, participants were relatively insensitive to how many of the model's forecasts they could overrule when making this choice. However, participants' increased propensity to use the model did not consistently translate into better performance because they made the model's forecasts significantly worse when adjusting them. For this reason, allowing participants to adjust all of the model's forecasts by a limited amount (i.e. the adjust-by-x process) seems like a more promising fix for algorithm aversion than allowing participants to adjust a limited number of the model's forecasts by an unlimited amount (i.e. the change-x process).

Study S3

Methods

Overview. Study S3 had the same design as Study 4, except we replaced the use-freely process with a forced-adjust-by-2 process. Unfortunately, there was a programming error in the survey after participants finished their first round of forecasts, rated their forecasting process, received performance feedback, and rated the three forecasting processes that they could choose from for stage 2. This programming error told all participants that one of their options for the stage 2 forecasts was adjusting the model by up to 10 percentiles, but half of participants actually had the option to adjust the model by up to 2 percentiles instead. Thus, participants who actually had the adjust-by-2 option were contaminated before choosing their Stage 2 forecasting method. For this reason, we only present the responses that participants entered before this programming error in the survey.

Participants. We ran Study S3 with participants from Amazon Mechanical Turk. Participants earned $1 for completing the study and could earn up to an additional $1 depending on their forecasting performance. We decided in advance to recruit 800 participants (200 per condition). Participants who began the study completed a question before they started the survey design to ensure that they were reading instructions. We programmed the survey to stop any participants who failed this check from taking the study (130 failed this check) and 151 participants quit the survey before completing all of their forecasts, leaving us with a final sample of 816 participants who completed their forecasts. This sample averaged 33 years of age and was 45% female.

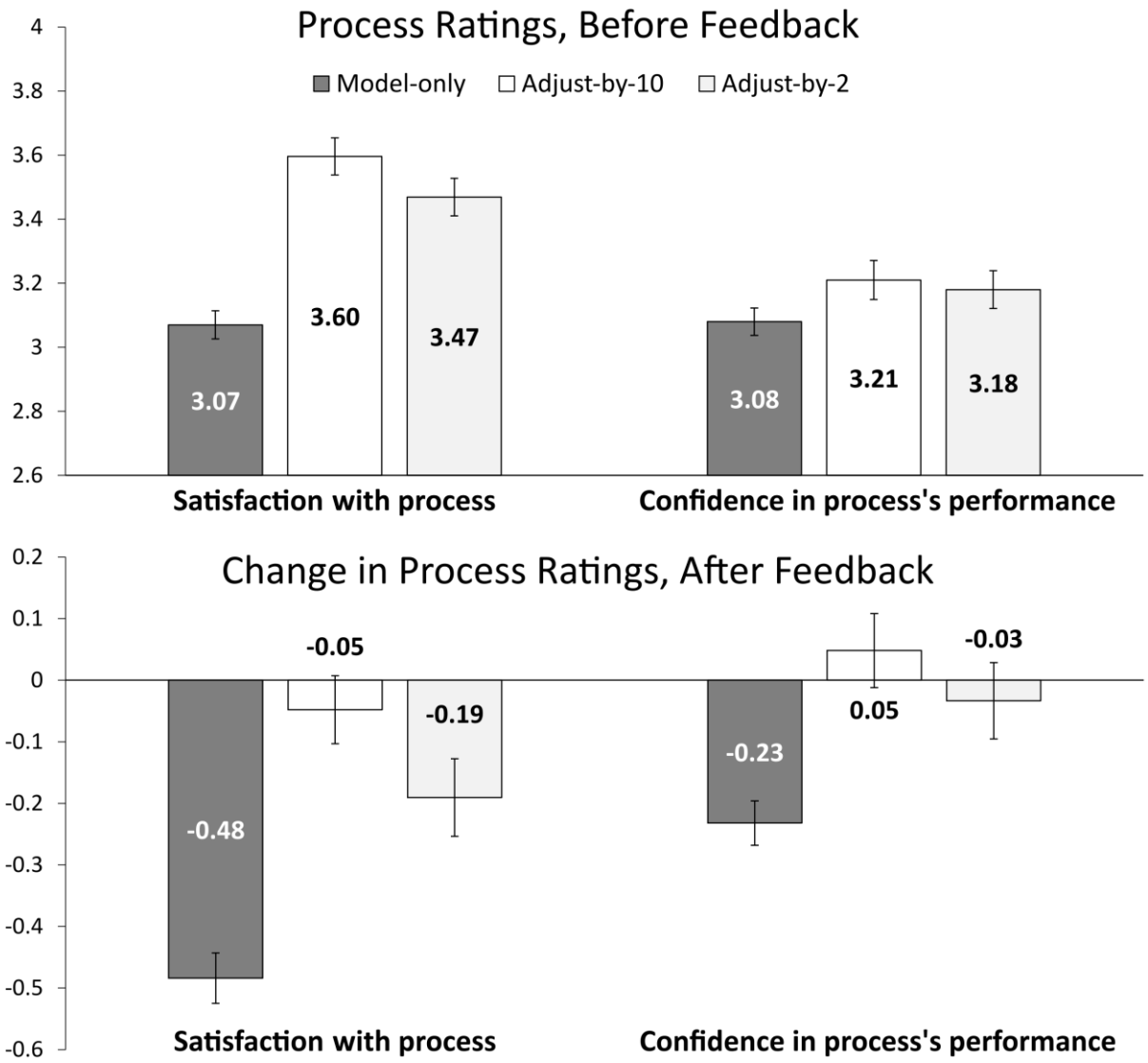
Procedures. This study's procedure was the same as Study 4's except for one change. We replaced the use-freely process with the adjust-by-2 forecasting process. As a result, one quarter of participants were assigned to use the adjust-by-2 process for their first round of forecasts (instead of the use-freely process), and these participants and half of participants in the model-only condition had the option to use the adjust-by-2 process for their Stage 2 forecasts (instead of the use-freely process).

Results.

Are people more satisfied with a forecasting process that allows them to modify a model's forecasts than with one that does not? As shown in Figure S8, participants in the adjust-by-10 and adjust-by-2 conditions, who were able to modify the model's forecasts, rated their assigned forecasting process more favorably than participants in the model-only condition, who could not modify the model's forecasts. Participants in the model-only condition, who could not modify the model's forecasts, were significantly less satisfied with their forecasting process than those assigned to the adjust-by-10 condition, $t(616) = 7.03$, $p < .001$, and adjust-by-2 condition, $t(619) = 5.32$, $p < .001$. Also, participants in the model-only condition were marginally less confident in the performance of their forecasting process than participants in the adjust-by-10 condition, $t(616) = 1.71$, $p = .088$, and directionally less confident in the performance of their forecasting process than participants in the adjust-by-2 condition, $t(619) = 1.37$, $p = .172$. Interestingly, participants in the adjust-by-2 condition were about equally confident in, $t(417) = 0.31$, $p = .753$, their assigned forecasting process compared to participants in the adjust-by-10 condition, even though they had less freedom to adjust the model's forecasts. However, participants in the adjust-by-10 condition were marginally more satisfied with, $t(417) = 1.54$, $p = .124$, their assigned forecasting process than participants in the adjust-by-2 condition.

Figure S8

Study S3: Participants who could modify the model's forecasts were more satisfied with their forecasting process (top panel) and reacted less harshly after learning that the process had erred (bottom panel).



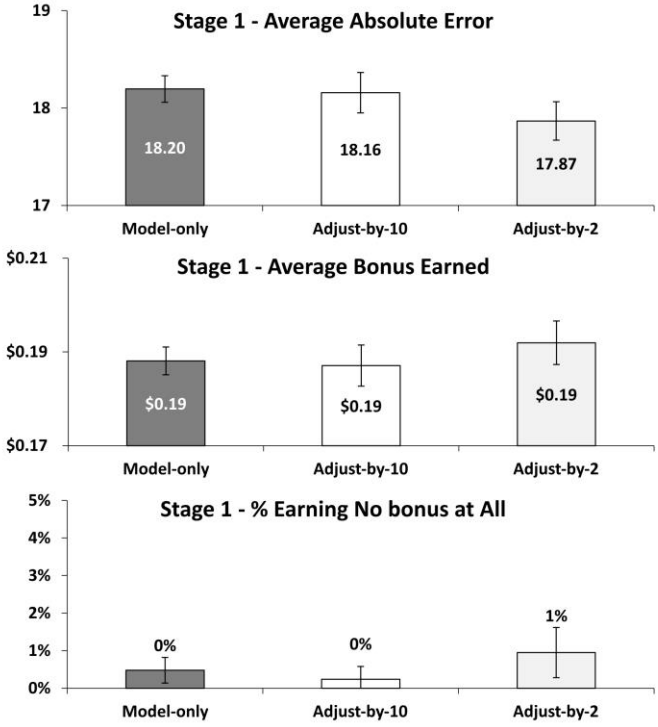
Note: Errors bars indicate ±1 standard error.

Are people more forgiving of forecasting errors when they were able to modify the model's forecasts than when they were not? To answer this question, we computed the change between participants' satisfaction/confidence with their Stage 1 process before vs. after receiving performance feedback. Positive values indicate that people's satisfaction/confidence increased after learning how well

they performed, whereas negative values indicate that people's satisfaction/confidence decreased after learning how well they performed. As shown in Figure S8, analyses of these measures revealed that participants in the adjust-by-10 and adjust-by-2 conditions, who were able to modify the model's forecasts, were less sensitive to performance feedback than participants in the model-only condition, who could not modify the model's forecasts.[2] As we found in Study 4, giving participants some control over the model's forecasts not only increased their satisfaction with their forecasting process; it also rendered that satisfaction more impervious to performance feedback.[3]

Figure S10.

Study S3: Participants' Stage 1 forecasting performance.



Note: Errors bars indicate ±1 standard error. Only the first round of forecasts are included because of the programming error described in the study overview.

[2] Participants in the model-only condition lost significantly more satisfaction with, $t(615) = 6.23$, $p < .001$, and confidence in, $t(615) = 4.22$, $p < .001$, their assigned forecasting process compared to participants in the adjust-by-10 condition. Also, participants in the model-only condition lost significantly more satisfaction with, $t(617) = 4.01$, $p < .001$, and confidence in, $t(617) = 2.96$, $p = .003$, their assigned forecasting process compared to participants in the adjust-by-2 condition.

[3] Every participant saw their forecasting process err. The best performing participant had an average absolute error of 10.3.

Stage 1 forecasting performance. These were not any significant performance differences between conditions in the Stage 1 forecasts. Participants in all conditions had similar average absolute errors (see Figure S10, all *p*'s >= .165) and earned similar bonuses (see Figure S10, all *p*'s >= .468). Additionally, only four participants earned a bonus of $0 because all participants were kept from straying more than 10 percentiles from the model's forecasts.

Discussion. This study replicated many of the findings from Study 4. Having participants modify the algorithm's forecasts had important benefits beyond increasing their use of the model. Allowing participants to modify an algorithm's forecasts both increased their satisfaction with their forecasting process and heightened their tolerance of errors.