**Supplementary Materials**

**MBA Forecasting Model**

This section describes the model we used in Studies 1, 2, and 4 to predict the success of MBA students. The model was an ordinary least squares regression built using data from the same pool of 115 students that participants were exposed to.

The dependent variable was an equal weighting of GPA, respect of fellow students (assessed via a survey), prestige of employer upon graduation (as measured in an annual poll of MBA students around the U.S.), and job success 2 years after graduation (measured by promotions and raises). We converted this variable into percentiles.

The model consisted of eight predictors, including students’: undergraduate degree (categorical: engineering, liberal arts, business, other), GMAT verbal score, GMAT quantitative score, years of work experience (categorical: less than 2, 2-5, more than 5), interview score (1 – below expectations to 5 – outstanding), essay score (1 – below expectations to 5 – outstanding), average salary since graduating from college, and average of parents’ education (1 – no college experience to 5 – multiple graduate degrees).

On average, the model’s predicted percentiles were 22.75 percentiles away from students’ realized percentiles. The model perfectly estimated the percentile of 3 of the 115 students and it was within 5 percentiles for 20 students. The model’s predictions were correlated at .53 with student’s actual percentiles, and the model’s R-squared was .28.

**Airline Passengers Model**

This section describes the model used in Studies 3a and 3b 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 perfectly predicted the rank of 6 states. The model’s predictions were correlated at .92 with states’ actual ranks, and the model’s pseudo R-squared was 0.26.

**Studies S1-S3**

In this section, we describe three studies that were not included in the paper. These studies were similar to Studies 3a and 3b, but with a few potentially important procedural differences (discussed below). These studies were conducted prior to Studies 3a and 3b. Indeed, Study 3b was conducted primarily to ensure that Study 3a’s results were replicable, especially in light of the mixed results observed in Studies S1-S3.

**Methods**

***Participants*.** We conducted Studies S1 and S2 using participants from amazon.com’s Mechanical Turk (MTurk) website.[[1]](#footnote-1) These participants received $0.50 for completing the study and they could earn up to an additional $2 for accurate forecasting performance. In Study S1 we decided in advance to recruit 400 participants (100 per condition) and in Study S2 we decided to recruit 600 participants (150 per condition). In both studies, participants who responded to the MTurk posting completed a question before they started the survey design to ensure that they were reading instructions. We programmed the survey to exclude any participants who failed this check (68 in Study S1 and 117 in Study S2), and a number of participants exited the survey before completing the key dependent measure (62 in Study S1 and 125 in Study S2). This left us with final samples of 414 in Study S1 and 610 in Study S2. These samples averaged 31 years of age and were 40-44% female.

We conducted Study S3 in the Wharton School’s Behavioral Lab. These participants received a $10 show-up fee for an hour-long session of experiments, of which ours was a 20-minute component, and they could earn up to an additional $8 for accurate forecasting performance. We decided to recruit as many participants as we could in two weeks. We could not screen out participants who had completed Studies 1, 2, and 4 (and similar studies) because it would have prevented the other researchers in the lab session from recruiting enough participants.[[2]](#footnote-2) As a result, our sample included 134 participants who had completed a study with the same general design. Six participants failed to complete the study’s key dependent measure, leaving us with a final sample of 417. The sample averaged 22 years of age and was 63% female.

**Procedure**. The procedures of Studies S1-S3 were identical, and similar to Study 3a. As in Study 3a, participants were randomly assigned to one of four conditions – model, human, model-and-human, and control – which determined whether, in the first stage of the experiment, they saw the model’s forecasts, made their own forecasts, both, or neither. The procedure was the same as Study 3a except for four changes. First, participants were not informed that the model was built by experts, and they were not told that the model was “a sophisticated model, put together by thoughtful analysts.” Second, unlike in Studies 3a and 3b, participants were given the state’s name on each trial. Third, we used a different payment rule to determine participants’ bonuses. Participants in Studies S1 and S2 were paid $2 if the average absolute error of their estimates was less than or equal to one. This bonus decreased by $0.25 for each additional unit of average error associated with their stage 2 estimates. Participants in Study S3 were paid $8 if the average absolute error of their estimates was less than or equal to one. This bonus decreased by $1 for each additional unit of average error associated with their stage 2 estimates. Participants were not given a bonus if their average absolute error exceeded eight. Fourth, participants made 10 incentivized stage 2 forecasts instead of one. Finally, participants did not complete the exploratory measure that asked how likely it is that the model predicts state's ranks almost perfectly, which was included in Study 3a.

**Results and Discussion**

***Forecasting Performance.*** As shown in Table S1, the model’s forecasts outperformed participants’ forecasts, and participants would have been better off tying their bonuses to the model in all studies.

Table S1

Forecasting Performance of Model vs. Human: Means (and Standard Deviations)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
| Bonus If Chose Model vs. Human |
|  | Model | Human | Difference | Paired t-test |
|  |  |  |  |  |
| Study S1 | $1.04 (0.30) | $0.53 (0.46) | $0.51 (0.49) | *t*(413) = 20.92, *p* < .001 |
|  |  |  |  |  |
| Study S2 | $1.06 (0.31) | $0.54 (0.45) | $0.52 (0.48) | *t*(607) = 26.32, *p* < .001 |
|  |  |  |  |  |
| Study S3 | $4.29 (1.15) | $2.19 (1.79) | $2.10 (1.92) | *t*(415) = 22.37, *p* < .001 |
| Average Absolute Error in Model-and-Human Condition (Stage 1’s Unincentivized Forecasts) |
|  | Model | Human | Difference | Paired t-test |
|  |  |  |  |  |
| Study S1 | 4.52 (1.30) | 7.40 (3.36) | -2.88 (3.39) | *t*(103) = -8.67, *p* < .001 |
|  |  |  |  |  |
| Study S2 | 4.25 (1.14) | 7.51 (2.63) | -3.26 (2.77) | *t*(153) = -14.60, *p* < .001 |
|  |  |  |  |  |
| Study S3 | 4.34 (1.21) | 7.58 (2.94) | -3.24 (3.03) | *t*(106) = -11.06, *p* < .001 |
| Average Absolute Error (Stage 2’s Incentivized Forecasts) |
|  | Model | Human | Difference | Paired t-test |
|  |  |  |  |  |
| Study S1 | 4.41 (1.15) | 7.04 (3.02) | -2.63 (3.02) | *t*(413) = -17.71, *p* < .001 |
|  |  |  |  |  |
| Study S2 | 4.30 (1.23) | 7.05 (3.02) | -2.75 (3.10) | *t*(607) = -21.85, *p* < .001 |
|  |  |  |  |  |
| Study S3 | 4.31 (1.14) | 6.95 (2.87) | -2.65 (2.94) | *t*(415) = -18.40, *p* < .001 |

***Main Analyses****.* Across studies, the results were mixed (see Table S2). In Study S1, participants who had not seen the model make forecasts with feedback were more likely to tie their bonuses to the model (50%) than were participants who had seen the model make forecasts with feedback (40%), *χ*2(1, *N* = 414) = 3.97, *p* = .046. This is consistent with our hypothesis. However, although Study S2 showed the same pattern — participants who saw the model perform (47%) were directionally less likely to bet on the model than were participants who did not see the model perform (51%) – this difference was not significant, *χ*2(1, *N* = 610) = 1.11, *p* = .292. Finally, Study S3, which used participants from the Wharton Behavioral Lab, showed a slight trend in the opposite direction: participants who saw the model perform were slightly more likely to choose the model than those who did not (45% vs. 42%), *χ*2(1, *N* = 417) = 0.33, *p* = .564. However, this trend reversed after excluding participants who had previously completed a similar study. In this case, participants who saw the model perform (40%) were directionally less likely to bet on the model than were participants who did not see the model perform (43%), *χ*2(1, *N* = 283) = 0.26, *p* = .609.

Table S2

Percentage of Participants Who Chose Model: Means (and Standard Deviations)

|  |
| --- |
|  |
|  | Control | Human | Model | Model-and-Human |
|  |  |  |  |  |
| Study S1 | 42%a (0.50) | 58%b (0.50) | 42%a (0.50) | 39%a (0.49) |
|  |  |  |  |  |
| Study S2 | 53%a (0.50) | 50%a (0.50) | 49%a (0.50) | 45%a (0.50) |
|  |  |  |  |  |
| Study S3 | 42%a (0.50) | 41%a (0.50) | 49%a (0.49) | 40%a (0.49) |
|  |

*Note*. Within each row, percentages with different subscripts differed at *p* < .05 using chi-square tests. Excluding participants from Study S3 who had completed a similar study, the percentages of participants selecting the model in Study S3 was: Control – 43%, Human – 43%, Model – 46%, Model-and-Human – 34%. None of these percentages significantly differed.

***Confidence***. Table S3 shows participants’ confidence ratings broken down by study and condition. Seeing the model perform marginally significantly decreased confidence in the model’s forecasts in Study S1, *t*(407) = 1.77, *p* = .078, and Study S3[[3]](#footnote-3), *t*(414) = 1.57, *p* = .118, and significantly decreased confidence in the model’s forecasts in Study S2, *t*(604) = 2.22, *p* = .027. Seeing the human perform did not decrease confidence in the human’s forecasts in any of the studies: S1, *t*(407) = 0.75, *p* = .454; S2, *t*(604) = -0.73, *p* = .469; S3[[4]](#footnote-4), *t*(414) = -0.37, *p* = .708.

Table S3

Confidence in Model’s and Human’s Forecasts: Means (and Standard Deviations)

|  |
| --- |
| Confidence in Model’s Forecasts |
|  | Control | Human | Model | Model-and-Human |
|  |  |  |  |  |
| Study S1 | 3.41a (0.82) | 3.49a (0.77) | 3.35a (0.70) | 3.28a (0.74) |
|  |  |  |  |  |
| Study S2 | 3.54a (0.78) | 3.53a (0.69) | 3.39a (0.73) | 3.42a (0.64) |
|  |  |  |  |  |
| Study S3 | 3.37a (0.84) | 3.56a (0.65) | 3.38a (0.71) | 3.33a (0.71) |
|  |
| Confidence in Human’s Forecasts |
|  | Control | Human | Model | Model-and-Human |
|  |  |  |  |  |
| Study S1 | 2.89a (0.78) | 2.90a (0.81) | 3.05a (0.78) | 2.91a (0.81) |
|  |  |  |  |  |
| Study S2 | 2.97a (0.81) | 3.03a (0.88) | 3.09a (0.79) | 3.12a (0.81) |
|  |  |  |  |  |
| Study S3 | 2.99a (0.82) | 3.06b (0.76) | 2.97a (0.82) | 2.96a (0.75) |

*Note*. Within each row, means with different subscripts differ at *p* < .05 using Tukey’s test. Excluding participants from Study S3 who had completed a similar study, average confidence in the model’s forecasts in Study S3 was: Control – 3.42ab (0.75), Human – 3.63a (0.67), Model – 3.29b (0.73), Model-and-Human – 3.33ab (0.75). Excluding participants from Study S3 who had completed a similar study, average confidence in the human’s forecasts in Study S3 was: Control – 2.98a (0.86), Human – 3.07a (0.68), Model – 2.94a (0.80), Model-and-Human – 3.04a (0.75).

***Discussion*.** There are many possible reasons for the conflicting choice results. We think that including participants who had previously been exposed to our experimental design in Study S3 is a likely explanation. Although data from Studies S1 and S2 supported our hypothesis, at least directionally, data from Study S3 did not. However, the data from Study S3 did directionally support our hypothesis when participants who had completed a similar study were excluded. It is possible that, after having completed a similar study in the Wharton Behavioral Lab, these participants had learned the financial advantages of choosing the model over the human. Additionally, these participants’ responses may have been tarnished because the majority of them were exposed to two conditions of our experimental paradigm. Whatever the cause, the consistent pattern of results across Studies 1, 3a, 3b, 4, S1, and S2 give us confidence that seeing a model perform usually makes people less likely to choose it.

1. Eligibility was restricted to U.S. residents with at least a 90% MTurk approval rating. [↑](#footnote-ref-1)
2. This “problem” was unique to Study S3. Study 1, 2, and 4’s participants had not previously participated in similar studies. [↑](#footnote-ref-2)
3. This difference was significant when participants who had taken a similar study were excluded, *t*(281) = 2.57, *p* = .011. [↑](#footnote-ref-3)
4. This difference remained nonsignificant when participants who had taken a similar study were excluded, *t*(281) =

-0.99, *p* = .321. [↑](#footnote-ref-4)