Predictive Training and Accuracy: Self-selection and Causal Factors

REGINA JOSEPH, Pytho LLC
PAVEL ATANASOV, Pytho LLC

1.INTRODUCTION

Longitudinal studies of crowd-based forecasting have sought to quantify the causal impact of training interventions on individual and collective forecasting accuracy. In forecasting tournaments, crowds of human forecasters supply probabilistic predictions [e.g. Mellers et al. 2014] which are aggregated statistically to produce collective estimates [Atanasov et al., 2017]. Superior collective intelligence in this context depends on the skill of individual contributors [Mellers et al. 2015]. Studies examining the effects of training have targeted structural knowledge that affects forecasting accuracy, like probability estimation and Bayesian reasoning [Chang et al, 2015; Fischbein and Gazit, 1984; Fong et al, 1986; Mellers and McGraw, 1999; Mellers et al, 2014; Mellers et al, 2015]; other studies have analyzed the basic concepts of heuristics and cognitive biases [Kahneman and Tversky, 1973, 1979], and more recent studies tested specific interventions to mitigate biases and improve accuracy [Barbey and Sloman, 2007; Benson and Onkal, 1992; Case et al, 1999; Chang et al, 2016; Flyvbjerg, 2008; Morewedge et al, 2017; Rhodes et al, 2017]. But can training improve predictive accuracy in settings where human forecasters are exposed to machine model predictions? How much of a training effect is causal vs. a result of self-selection? We show how targeted training improved the predictive accuracy of human subjects in a U.S. intelligence community-sponsored[1] forecasting tournament focused on hybridizing crowdsourced human judgment with forecasts supplied by machine models. Quantifying the training intervention's accuracy effect is critical for improving collective forecasts and investigating the human-computer interaction effect when forecasters work with machine models.

1.1 The H.A.B.I.T. Approach to Improving Accuracy in Crowdsourced Hybrid Forecasting

Hybrid forecasting—in which crowdsourced human predictions are combined with automated machine model projections—places different cognitive demands on forecasters than human-only tournaments. In a hybrid tournament, forecasters are confronted with assessing whether the probability estimates supplied by machine models are trustworthy enough to inform their own predictive efforts. To train hybrid forecasters, we reasoned that not only would we need to provide instruction in good forecasting technique, we would also need to familiarize forecasters with the machine models to which they would have access. Studies suggest that exposure to algorithms can induce excessive aversion [Dietvorst et al, 2015], but also over-confidence in algorithms [Logg et al, 2018]. Given the context of the forecasting process' "pain points" [Joseph, 2018], we hypothesized the cognitive burden of whether to integrate machine model data or reject it in pursuit of a more accurate forecast could be mitigated by briefly explaining how each model works and how to strike a balance between too little and too much trust in models. Thus we created a three-part training intervention to instruct forecasters in a hybrid tournament in 30 minutes or less on how to improve their forecasting accuracy. After a simple guide

[1] This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via 2017-17071900005. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. government. The U.S. government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

on how to use our research team's forecasting platform, the two-part H.A.B.I.T. training approach combines probabilistic reasoning with hybridization concepts, knitting them together using a character-based narrative device rendered in a cartoon format (both as static illustrations as well as short films; only the static format is tested here). The three parts are:

1) Training Module 1: How to use the platform—a video guide to forecasting platform features
2) Training Module 2: H,A.B.I.T.—a mnemonic of five core elements of elite forecasting technique
3) Training Module 3: Hybridizing Your Forecasting—a H.A.B.I.T. guide to machine models used

The H.A.B.I.T. approach was designed to deliver accuracy improvements to hybrid forecasters in approximately half the duration of prior art. H.A.B.I.T.'s development builds upon an hour-long non-hybrid forecasting training protocol, CHAMPS KNOW [Mellers et al, 2015; Chang et al, 2016].

2.1 Methods

In Study 1, approximately 40% of participants were recruited through public campaigns (public sample), while the rest were recruited via Amazon's Mechanical Turk (MTurk sample). Training completion was voluntary; forecasters were experimentally assigned to work independently or in teams; and given access to a) historical data and machine model projections, b) only historical data, or c) neither. The core sample included forecasters who attempted at least 10 scored questions (out of 187). Training completion was defined as viewing at least 90% of the content of at least one training module. To assess aggregate performance, forecasts from both public and MTurk samples were combined using a version of the aggregation algorithm described in Atanasov et al. [2017]. In Study 2, all participants were MTurkers; they were experimentally assigned to either training or a control condition, in which they read popular articles discussing forecasting as a practice, but offering no tips for boosting accuracy. Study 2 was completed in March 2019. The reported results are based on the sample of forecasters who attempted 5 or more out of 81 resolved questions. Individual accuracy was calculated as the Brier score per question (standardizing to distributions with M=0, SD=1), averaged across questions, then across forecasters.

2.2 Results

In Study 1, 19.4% of participants completed training. The rest were untrained, including those did not visit the training website (73.4%), and those who viewed but did not complete training (8.2%). Accuracy differences between trained and untrained forecasters were significant (t = -4.70, p < .001, Cohen's d = 0.43). Effect sizes were higher for MTurkers (d = .56) than in the public sample (d = 0.44).

Table 1. Standardized Brier scores for individuals by training condition in Study 1. Scores below zero denote better than average accuracy.

	All Forecasters		Public Sample		MTurk Sample	
	N	Mean (SD)	N	Mean (SD)	N	Mean (SD)
Training Completers	153	-0.09 (0.36)	48	-0.18 (0.57)	105	-0.06 (0.32)
Training Non-Completers	703	0.06 (0.35)	117	0.05 (0.51)	586	0.06 (0.19)
Cohen's d		0.43		0.44		0.56

Training completion was also associated with substantial and statistically significant improvement in aggregate accuracy. Aggregate forecasts from trained forecasters were more accurate than those made by forecasters who never visited the training website (M = 0.263, SD = 0.303 vs. M = 0.308, SD = 0.327, paired t(186) = 2.94, p = .009, paired d = 0.21). Aggregations of trained forecasters even outperformed the full sample of all forecasters, which included both trained and untrained ones (M = 0.283, SD = 0.311, paired t(186) = 1.92, p = .110, paired d = 0.14). The latter result is striking, because it shows that untrained forecasters, who made up approximately 80% of the cohort, provided no marginal benefit, and may have even contributed to a small reduction in aggregate accuracy.

Table 2. Comparison of aggregate accuracy by training status for N=187 questions.

Forecaster Set	N Forecasters	Aggregate Forecast Brier Score, Mean (SD)	Paired t, p-value vs. Completers**
All Forecasters	1443	0.283 (0.311)	$t = 1.92$, $p = .110$
Training Completers*	280	0.263 (0.303)	NA
Training Non-Visitors	1059	0.308 (0.327)	$t = 2.94$, $p = .009$

*Only includes forecasts on or after date of first visit to training site.
**p-values after Bonferroni adjustment for N=2 comparisons.

To what extent was the effect of training accuracy causal, versus an artifact of self-selection? To answer this, we estimated self-selection effects, namely, the extent to which forecasters that chose to complete training in Study 1 were: a) more talented or better positioned, b) harder working, or c) better from the start. This exercise of estimating causal effects from observational data parallels real-world challenges in organizations that provide training to their staff, but are unable to perform a gold-standard test for causal effects by assigning staff randomly to treatment and control groups.

First, to assess if training completers were better positioned or more talented at forecasting, we used regression models estimating the effect of training completion on accuracy, accounting for experimental conditions (teaming, machine model information access) and psychometric scores that have been associated with forecasting accuracy [Mellers et al. 2015]. Data availability constrained this analysis to the public sample. The training effect estimate was d = 0.42 in a univariate model.. The regression-estimated effect of training was $d = 0.42$ in a univariate model. This effect was reduced to $d = 0.36$ when we accounted for experimental conditions. Adding psychometric scores to the model reduced the estimated effect to $d = 0.34$. Second, to assess if trained forecasters worked harder than untrained ones, we compared overall activity, i.e. the total number of forecasts submitted, as well as the distribution of activity. Training instructed forecasters to focus efforts on a limited number of questions, and to revise forecasts often. Trained forecasters did submit more forecasts (M = 95, SD = 152) than untrained ones (M = 70, SD = 195). Consistent with training instructions, trained forecasters did not answer any more questions (M = 54, SD = 34 vs. M = 56, SD = 29) but submitted more forecasts per question (M = 1.7, SD = 1.4 vs. M = 1.2, SD = 1.2) than untrained ones. We used a regression model to estimate the impact of equalizing overall activity levels between trained and untrained forecasters. The activity adjustment reduced the accuracy differences in the full model from $d = 0.34$ to $d = 0.29$. After all adjustments, 69% of the training effect remained intact (d = 0.42 vs. d = 0.29). Third, to assess if trained forecasters were better from the start, we exploited the timing of activity, by comparing the forecast accuracy before versus after training exposure. We matched individual forecasts from trained and untrained forecasters on the same question on the same date, and compared Brier scores for each forecast pair. The forecasts submitted by trained forecasters *before* training exposure were non-significantly *less* accurate than those made by untrained ones (Mean difference = -0.01). In contrast, for forecasts submitted *at or after date of* training exposure, trained forecasters significantly outperformed (Mean difference = 0.12, $t > 4.00$, $p < .001$). This result implies that all (100%) of the accuracy difference could be attributed to training.

Study 2 allowed us to assess training effects experimentally: MTurk forecasters were randomly assigned to receive training or a control intervention in the third week they visited the platform. The standardized difference in forecaster accuracy after exposure to training vs. control intervention was $d = 0.54$. This was consistent with the unadjusted estimate among MTurk forecasters in Study 1.

In summary, results from both studies point to a sustained accuracy effect of this half-hour hybrid training intervention. This effect was roughly as large those of prior training programs for non-hybrid forecasting [Mellers et al., 2014, Chang et al. 2016], despite the current intervention taking roughly half the time to administer and being evaluated against a more stringent control condition.

REFERENCES

Atanasov, P., Rescober, P., Stone, E., Swift, S.A., Servan-Schreiber, E., Tetlock, P., Ungar, L. and Mellers, B., (2017). Distilling the wisdom of crowds: Prediction markets vs. prediction polls. *Management Science*, 63(3), pp.691-706.

Barbey, A. K., & Sloman, S. A. (2007). Base-rate respect: From ecological rationality to dual processes. *Behavioral and Brain Sciences,* 30(03), 241–254.

Benson, P. G., & Önkal, D. (1992). The effects of feedback and training on the performance of probability forecasters. *International Journal of Forecasting,* 8(4), 559–573.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review,* 78(1), 1– 3.

Case, D. A., Fantino, E. & Goodie, A. S. (1999). Base rate training without case cues reduces base-rate neglect. *Psychonomic Bulletin and Review,* 6, 319–327.

Chang, W., Chen, E., Mellers, B., & Tetlock, P. (2016). Developing expert political judgment: The impact of training and practice on judgmental accuracy in geopolitical forecasting tournaments. *Judgment and Decision Making,* 11(5), 509. https://www.sas.upenn.edu/~baron/journal/16/16511/jdm16511.pdf

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General,* 144(1), 114.

Fischbein, E., & Gazit, A. (1984). Does the teaching of probability improve probabilistic intuitions? *Educational Studies in Mathematics*, 15(1), 1–24.

Flyvbjerg, B. (2008). Curbing optimism bias and strategic misrepresentation in planning: Reference class forecasting in practice. *European Planning* Studies, 16(1), 3–21.

Fong, G. T., Krantz, D. H., & Nisbett, R. E. (1986). The effects of statistical training on thinking about everyday problems. *Cognitive Psychology*, 18(3), 253–292.

Graber, M. L. (2003). Metacognitive training to reduce diagnostic errors: ready for prime time? *Academic Medicine*, 78(8), 781.

Joseph, R. (2018). Pinpointing Pain Points: User Interface Strategies for Hybridizing Forecasting Performance. *Scenario Planning and Foresight 2018 Conference*, Operational Research Society, Coventry, UK Retrieved from doi: 10.13140/RG.2.2.25202.35520

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80(4), 237–251.

Kahneman, D., & Tversky, A. (1979). Intuitive prediction: Biases and corrective procedures. *Management Science*, 12, 313–327.

Logg, J. M., Minson, J.A., & Moore, D.A. (2019). Algorithm Appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151, 90-103. http://dx.doi.org/10.2139/ssrn.2941774

Mellers, B., & McGraw, A. P. (1999). How to improve Bayesian reasoning: Comment on Gigerenzer and Hoffrage. *Psychological Review*, 106, 417-424

Mellers, B., Stone, E., Atanasov, P., Rohrbaugh, N., Metz, E.,, Ungar, L., Bishop, M.M., Horowitz, M., Merkle, E., Tetlock, P.E. (2015). "The psychology of intelligence analysis: Drivers of prediction accuracy in world politics." Journal of experimental psychology: applied 21, no. 1 (2015): 1. https://repository.upenn.edu/cgi/viewcontent.cgi?article=1390&context=fnce_papers

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., ... & Murray, T. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological science*, 25(5), 1106-1115.

Morewedge, C. K., Yoon, H., Scopelliti, I., Symborski, C. W., Korris, J. H., & Kassam, K. S. (2015). Debiasing decisions: Improved decision making with a single training intervention. *Policy Insights from the Behavioral and Brain Sciences,* 2(1), 129-140. https://doi.org/10.1177/2372732215600886

Rhodes, R. E., Kopecky, J., Bos, N., McKneely, J., Gertner, A., Zaromb, F., Perrone, A., Spitaletta, J. (2017). Teaching Decision Making With Serious Games: An Independent Evaluation. *Games and Culture*, 12(3), 233–251. https://doi.org/10.1177/1555412016686642