Creating a user-friendly Bayesian Network tool for Forecasting

Ross Gruetzemacher[1,2,3,*], Toby D. Pilditch[2,4,5,*], David Paradice[6], Christy Manning[1,2], Amory Bennett[7], and Coralie Consigny[2,8]

[1]Wichita State University, [2]Transformative Futures Institute, [3]Centre for the Study of Existential Risk, University of Cambridge, [4]University of Oxford, [5]University College London, [6]Auburn University, [7]Quorum Research, [8]Forecasting Research Institute

Forecasting is crucial across many domains, prompting efforts to enhance accuracy through various methods, including aggregation techniques, expert sampling, and Bayesian Networks (BNs). While BN models have shown effectiveness in specific domains like meteorology, their broader application has been limited by the statistical expertise required to effectively use them. This study demonstrates that BNs can easily be utilized by a statistically-naive population to enhance forecasting reasoning. To overcome previous usability barriers, we developed a user-friendly BN tool accompanied by training modules for its use. In a randomized controlled trial during the 2023 NCAA Division I football season, 85 participants predicted a series of game outcomes over three weeks per game, across a five week period. Our treatment group was provided with a specialized training in conditional probability and BN utilization, along with access to pre-structured BN models to parameterize to generate forecasts. In contrast, our control group received basic training on probabilistic reasoning and calibration. While our analysis revealed that forecasts generated by the treatment group were comparable in accuracy to those of the control group, our primary focus centered on evaluating user feedback dimensions. Notably, participants reported positive perceptions of the BN tool and associated training on its usability and helpfulness in reasoning. These findings underscore the potential for broader adoption of such methods in forecasting applications, as user-friendliness is essential for using long-term widespread uptake of such complex methodologies. Moreover, in assessing accuracy enhancements, we found that, at this early stage, all accuracy measures approached ceiling levels. While this outcome is unsurprising given the simplicity of the forecasting domain intended for a tool pilot, it highlights a promising path for future investigations into the efficacy and scalability of BN methodologies for forecasting.

Keywords: Bayesian methods, Bayesian networks, forecasting, reasoning under uncertainty

1. Introduction

Forecasting is highly valued for its potential to inform decisions. In response to its recognized importance, there has been a concerted effort to refine and enhance forecasting accuracy through various approaches. These efforts have encompassed both informal and formal strategies, including aggregation techniques like Wisdom of the Crowd and prediction markets (Hastie & Kameda 2005), as well as sampling to uncover and leverage individuals identified for their forecasting prowess, referred

to as Superforecasters (Tetlock & Gardner 2016). Additionally, Bayesian methods, including the use of BNs (Pearl 1988), have been explored for their potential in improving forecast precision in applied domains (see e.g., Boneh et al. 2015). However, the complexity and technical nature of BNs often pose significant barriers to entry, limiting their usability amongst statistically-naive populations (including experts from other domains). In this work, we seek to overcome these shortcomings by developing a user-friendly BN tool and training designed to make BN methodologies and complex forecasting accessible and useful to those without formal statistical or BN expertise. We conducted an experiment within the context of a football tournament to allow individuals to use and engage with this tool, enabling us to gauge its reception and effectiveness in real-world application.

1.1. Bayesian Networks

Just as we leverage tools like notepads and cameras to bolster our memory, we can employ methodologies to assist in navigating the intricacies of reasoning tasks that involve making deductions from uncertain data. Probability theory enables the definition of precise relationships between varying degrees of belief (Gilio and Over 2012; Politzer 2016), while BNs elucidate the exact effects of changes in the likelihood of certain information on the probability of other, connected information pieces (Fenton and Neil 2018; Korb and Nicholson 2011; Pearl 1988, 2000). BN models are composed of two components, described below:

- First, a directed acyclic graph (DAG) structure, consisting of nodes (symbolizing variables) and arrows (or “edges”, symbolizing relationships between variables, and thus how one variable may influence another). These structures are represented visually, and are labeled with both the variables, and the possible “states” those variables can occupy (e.g., we might name a variable for a forecasted football game as “Arkansas v Missouri”, and two states that game can resolve as “Arkansas Wins” or “Missouri Wins”).
- Second, once the labeled structure is completed, probability tables are created for each node, which detail the likelihoods relevant to those variables. Nodes without incoming arrows, termed "parent" nodes, have unconditional probability tables that represent the prior probability of each variable state. Conversely, "child" nodes, which receive arrows from other nodes, feature conditional probability tables outlining the probability of each state of the child node given the various states (in combination) of its parent nodes.

With these two components completed, the parameterized BN can be used to make reasoning inferences automatically (i.e., the probability calculus required to conditionalize/update probabilities of variable states in light of surrounding variable probabilities is handled automatically by the model). In this way, when a variable state has been observed (e.g., a previous game outcome), that “observation” can correspondingly be made in the BN (i.e., the specific state is “set” to 100% / confirmed), such that outstanding probabilities of interest will update automatically given this change in knowledge state.

BNs’ powerful inference capability makes them increasingly valuable in fields requiring detailed predictions based on uncertain and interconnected variables, such as forensic science (Smit et al. 2016), healthcare (Constantinou et al. 2016; Fenton and Neil 2010), and meteorology (Boneh et al. 2015; Abramson et al. 1996). However, despite their growing application, the use of these techniques has typically been restricted to direct implementation by those with significant expertise in Bayesian probability or extensive specialized training (Nicholson et al. 2011; Smit et al. 2016).

Recent lab studies have indicated that with appropriate training and tool availability, BNs can facilitate reasoning-under-uncertainty (e.g., diagnosis and prediction) among lay persons under lab conditions (see Cruz et al. 2020). Although the properties of BNs are known to minimize inaccuracy (Pettigrew 2016) including in experts domains (Cofino et al. 2002), open questions persist as to whether they can be adapted to enhance forecasting abilities among laypersons and BN-naive subject matter experts without the direct assistance of BN specialists, in a manner that is both timely and cognitively manageable. This adaptability is essential to remove both a barrier to entry and a bottleneck to enhancing forecasting more broadly, as necessary involvement of a BN expert curtails applicability and scalability, whilst insufficient tool usability will prohibit long-term use. Correcting this would be particularly relevant for applications like forecasting tournaments or decision-making scenarios involving BN-naive subject matter experts, where ease of use and cognitive ease are important considerations.

While it is normatively reasonable to assume that BNs properties should extend to forecasting contexts, an obvious, yet persistent challenge remains: the very features that make BNs powerful – complex probabilistic modeling and understanding interconnected variable relationships – are aspects that the average person finds difficult to naturally comprehend (Juslin et al. 2009). The complexity of learning BN reasoning and models has been presented as a barrier to their use in forecasting (FRI conditional trees paper that will be published by then) as well as in broader context (see Rehder et al. 2017; Rottman et al. 2016).

1.2. Current Research

The work we present here seeks to address two interrelated questions regarding the viability of BNs for improving forecasting elicitation:

1. Can BN-naive participants who engage with the BN tool and training be left with positive perceptions regarding its usability, helpfulness, and intention to use in future?
2. Can BN-naive participants be trained in the use of BNs, such that the forecasts generated by their models yield an accuracy advantage (over reasonable controls)?

To address these questions, we conducted a randomized control trial (RCT) in which our treatment groups were trained on using conditional probabilities on a pre-structured, labeled BN model specifically designed for this study with user experience in mind. We would like to emphasize that this RCT was intended as the first phase of a more substantive research stream. Consequently, at this point, we seek to build an empirical base that a) provides a proof of principle that BN-naive forecasters can engage with BNs to enhance their reasoning abilities within a forecasting context, and b) assesses the BN tool's user-friendliness to confirm its intuitiveness and accessibility. This foundational stage (focusing on parameterization tasks) paves the way for future studies that will progressively increase the modeling complexity and incorporate more advanced tasks such as selecting or building BN structures.

2. Methods

2.1. Participants

A total of 85 participants were recruited from a large university in the southern U.S. with a strong American football culture and a member of the National Collegiate Athletic Association's (NCAA's) highly competitive Southeastern Conference (SEC). Of this initial 85, 47 were randomly assigned to the treatment condition, and 38 to the control condition.[1] Participants were recruited via email and word-of-mouth solicitations with details including the duration of the study (five weeks), expected amount of weekly engagement, and compensation. Of these participants, 56 completed the study (32 from the treatment condition, and 24 from the control condition). The treatment group consisted of 22 participants who identified as male, eight as female, and two preferred not to disclose, with ages between 19 and 28 years old (*Mean* = 21.97; *SD* = 1.95 years). The control group consisted of 17 participants who identified as male, five as female, and two preferred not to disclose, with ages between 20 and 27 years old (*Mean* = 21.64; *SD* = 1.4 years).

For the follow-up survey, conducted to collect feedback on the BN tool and training, all 56 of the completing participants were contacted, of whom 38 participants agreed to participate – with 25 from the treatment group, and 13 from the control group. Participants were recruited via email, with details of the survey requirements and compensation. From the treatment group, 18 participants identified as male, seven as female, with ages between 20 and 28 years old (*Mean* = 21.56; *SD* = 1.79 years). From the treatment group, 10 participants identified as male, three as female, with ages between 21 and 24 years old (*Mean* = 21.62; *SD* = 0.84 years).

Across the five weeks of the study, the median total time for completing forecasts across both conditions was 92 minutes,[2] resulting in an effective hourly wage of $67/hour (before bonuses). The median total time for completing forecasts in the treatment condition was 165 minutes. The median total time for completing forecasts in the control condition was 29 minutes.

2.2. Design

Participants assigned to the control group received a basic 60-minute training module (the "first module") on probabilistic reasoning and calibration. Participants in the treatment group received the same training as the control group, plus an additional 60-minute session divided in two modules designed to simplify the understanding and use of BNs for forecasting, with an intuitive interface and certain aspects of BN parameterization automated. The second module offered a step-to-step guidance on how to formulate and input conditional probabilities using an example BN model, and the third module focused on adding "observations" – the real-world results they observe on the outcome of the games – to their BN models. The total training duration for participants in the treatment group was 120 minutes. Like the control group's training, the BN training material could be accessed again at any point during the study. A detailed description of the training material provided to both control and treatment group can be found in section 2.4.

---

[1] A slightly greater proportion of participants was assigned to the treatment group in anticipation of attrition based on severe attrition in previous studies utilizing BN elicitation tools (Nyberg et al. 2020).

[2] Time spent data is provided by the forecasting platform. The platform measures time spent by participants *with the browser window open and in focus*, so the time spent data reported here is really a subset of total time on task. The platform does not measure e.g. time spent by participants reviewing team records on other websites.

The forecasting phase, following training, was designed to test the usability and effectiveness of the training and BN tool. Over five weeks, participants provided weekly forecasts of the 2023 NCAA Division I football season (American football), focusing on in-conference games for SEC teams during November 2023. Specifically, we focused on games in weeks 11 to 13 of the season. This resulted in five target forecasts in week 11, two for week 12, and five for week 13[3] – in sum, participants forecasted 12 games. Participants began providing forecasts three weeks out from the target game (i.e., forecasts began in week 9, during October, resulting in a five week study), and updated those forecasts across subsequent weeks leading up to the game (40-minute sessions per week). Forecasts were of binary game outcomes (i.e., the predicted winner of the game).

Participants in the treatment group were provided with BN model structures for their target game forecasts each week within a custom designed tool[4] identical to the one used during training, which they could then parameterize with conditional probabilities to inform their game forecasts and update them in subsequent weeks.

All participants were paid $100 for taking part in the study, with a $20 bonus for attaining a forecast accuracy in the top 50% of their group, and an $50 bonus for attaining a forecast accuracy in the top 25%.

Following the study, participants were asked to provide feedback via a distributed survey. The survey was distributed two weeks after the study's completion. Participation was voluntary, and a $15 compensation was provided for completion.

2.3. Procedure

Following recruitment and random assignment to either the treatment or control conditions, participants were then provided with access to the forecasting platform. The stages of the study are described below.

Week 1. In week 1, participants (via email reminder) were directed within the platform to complete the training provided for their condition. Training was presented in a slide-deck format, and participants remained blind to conditions (and therefore differences in training packages).

Training. In both conditions, participants were free to navigate the training slides in a self-paced manner, and could access the content at any point during the remainder of the study. For the control group, participants were provided basic training on the nature of the task (forecasting college football games), probabilistic reasoning, and calibration (see section 2.4 on materials). For the treatment group, training was split into three modules, each building off the last. In the first module, participants were provided with the same background knowledge as the control group. In the second module, participants learnt about the BN tool, with the basics of how to formulate a conditional probability and enter it into their models. For this module, participants were provided with an example model structure to parameterize, such that they could follow along with the training. Finally, in the module three participants were instructed how to add "observations" to their models, so that they could update

[3] Three games were excluded—two from the first week and one from the second week—due to the number of conditional forecasts (i.e., 25 or more) that would have been required to be made by the treatment group.
[4] The link shared in this paper is for the study-specific registration page that participants used. Interested readers may contact the authors for access to a demo version of the BN tool.

forecasts in light of game outcomes. For this, participants were also provided with an example model with which to follow along the training. Along with the training, treatment group participants were also provided with a glossary of terms to refer to at any point during and after the training. Significant effort was made to create a simple training that was accessible for the intended audience of undergraduate university students with an interest in competitive collegiate football.

Forecasts. Following the training, participants were then presented with five questions regarding the outcomes of intraconference games in week 11 of the SEC college football season. Each question was presented on a separate page, and upon entering a forecast, participants could move on to the next page. Those in the treatment group were also provided with a BN model structure showing the target game and earlier games (from weeks 9 and 10—the results of which were yet unknown) which they could parameterize according to their determined conditional probabilities,[5] such that their BN model could generate a forecast for the target game. Participation for the first week of the study had a cutoff time of 11am CT on Saturday, prior to the week 9 SEC games taking place on Saturday and Sunday. This allowed for five full days to complete the training and forecasts.

Weeks 2-5. For each remaining week of the study, participants were sent an email reminder inviting them to participate for the new round of forecasting, with the same cutoff for participation (11am CT on Saturday—prior to SEC game starting times for that week). As outlined in Table I below, for subsequent weeks participants were asked to provide new forecasts at 3 weeks out (i.e., two forecasts in SEC week 12/study week 2, and five forecasts in SEC week 13/study week 3). Further, participants were instructed to update their previous forecasts in light of the past week of game outcomes (i.e., at two weeks out and one week out). For control participants, this meant providing a new estimate for the target game forecasts, whilst for treatment group participants this additionally meant having access to their original parameterized BN model for that target forecast, such that they could update the model with new game observations to inform their updated forecast.

Table I: Target game forecast allocations across weeks of participation.

Study Week	**1**	**2**	**3**	**4**	**5**
SEC Game Week	**9**	**10**	**11**	**12**	**13**
Initial Forecasts (3 weeks out)	5 Games of SEC Week 11	2 Games of SEC Week 12	5 Games of SEC Week 13		
1st Update Forecasts (2 weeks out)		5 Games of SEC Week 11 *(SEC Week 9 game outcomes known*)	2 Games of SEC Week 12 *(SEC Week 10 game outcomes known*)	5 Games of SEC Week 13 *(SEC Week 11 game outcomes known*)	
2nd Update Forecasts (1 week out)			5 Games of SEC Week 11 *(SEC Week 9 & 10 game*	2 Games of SEC Week 12 *(SEC Week 10 & 11 game*	5 Games of SEC Week 13 *(SEC Week 11 & 12 game*

[5] The first week of the study was conducted during week 9 of the SEC football season. Games for weeks 9, 10, and 11 had not occurred yet, so, parameterizing the questions for week 11 required forecasting the necessary games in weeks 9 and 10.

			outcomes known)	*outcomes known*)	*outcomes known*)

Study Completion. Upon completion of study week 5, participants were debriefed, and all participants were paid the base amount for participation ($100) within one week of study completion. Upon resolution of week 13 SEC games, all forecasts (original and updated) were then compared to objective outcomes, and as such overall brier scores for each participant were calculated. From these scores, the top 50% from each group were paid a $20 bonus, and the top 25% were paid a $50 bonus. These bonuses were distributed within two weeks of the end of the study.

Feedback Survey. Two weeks after the end of the study, all participants (control and treatment group) were invited via email to participate in a feedback survey regarding the task, the training they received, and for treatment group participants, perceptions of the BN tool. Participants who accepted the invitation then completed the survey over a one-week period. The survey instrument consisted of a series of likert scale, probability, and open text questions. Specifically, this included nine questions on the training and forecasting task, 14 questions regarding the tool (for the treatment group only). One week after the cutoff date for survey completion, participants were paid $15 in compensation.

2.4. Materials

Materials used in the study can be divided into four components. First, the training materials provided to participants (subsection 2.4.1). Second, the tool used by participants as a platform for forecasting, and within which treatment groups parameterized BN models to augment their forecasts (subsection 2.4.2). Third, the forecasting questions provided to all participants each week of the study (subsection 2.4.3). Finally, the feedback survey questions asked of participants in follow-up to the main study (subsection 2.4.4). In each of these subsections explanations and examples are provided. When practical, full versions of materials are provided.[6] [REPOSITORY LINK]

2.4.1. Training

Both control and treatment group training entailed participants self-guided exploration of a series of slides embedded in the forecasting platform used for facilitating the study. These slides were viewable at all times during the study.

Control Training. The control group received seven slides of training, covering the basics of reasoning under uncertainty, assigning probabilities to uncertain events, placing this within a sports context (game outcomes), formulating a forecast probability, and forecasting calibration over time (see Fig. 1 below).

[6] An online materials and data repository can be found at: <TODO: add link to git repo>

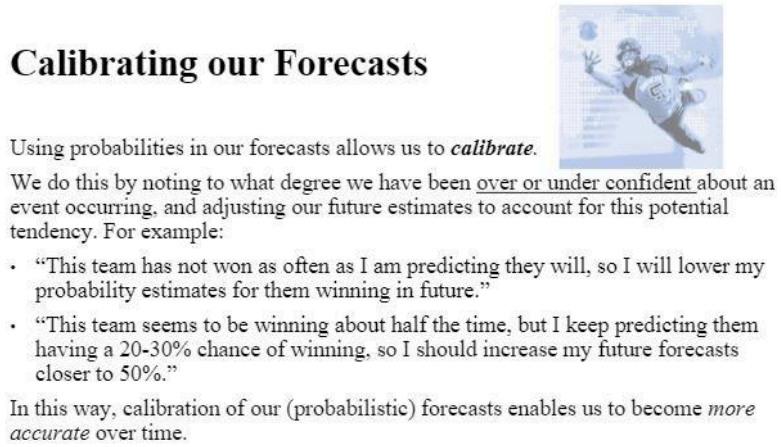

Figure 1: Example slide from control group training outlining calibration, and its value to forecasting.

Treatment Training. Participants assigned to the treatment group were required to complete three separate training modules. The first covered the same task and reasoning basics as did the training that the control group received. The second module—displayed on a separate page covered the basics of formulating prior and conditional probabilities for use within a BN and the specific BN tool used in this study (see Fig. 2).

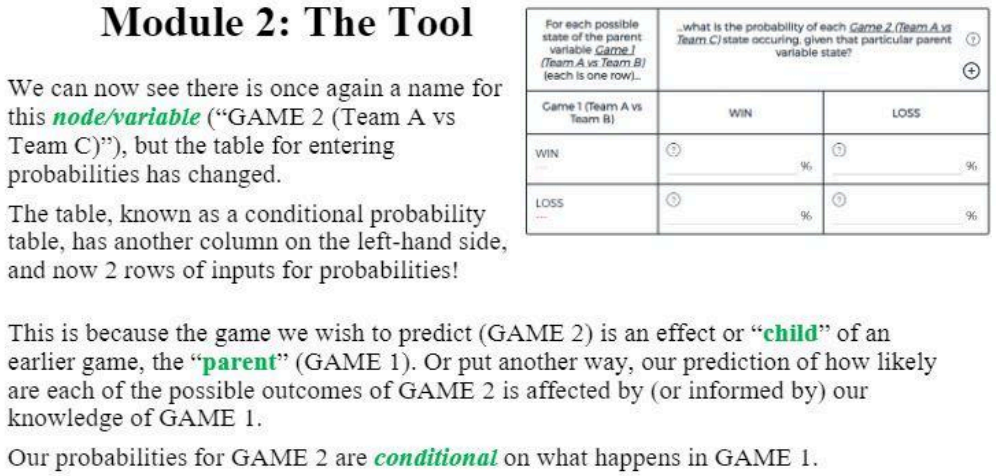

Fig. 2: Treatment group training slide from module 2, covering the basic terminology of BNs, and beginning to instruct on the method of model parameterisation using a conditional probability table.

To help ground this training, participants were also provided with example BNs within the tool (below the slides), with which they could follow along the training (see Fig. 3).

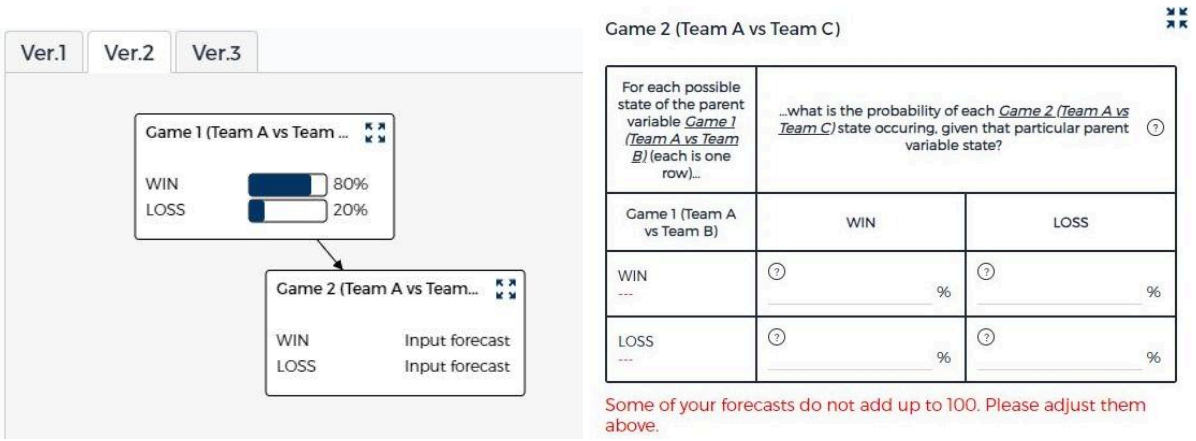

Fig. 3: Treatment group example training model. ***a (left)****. Example model for a BN with two nodes.* ***b (right)****. Having clicked on the bottom node, the conditional probability table for inserting parameters is displayed to participants to follow along the training for parameterising the BN model.*

Finally, the third module covered how to update their models in subsequent weeks, by making "observations" (setting node outcomes) in their previously parameterized networks (see Fig. 4), such that they could "read-off" updated forecasts from their models.

Fig. 4: Treatment group training slide from module 3, covering how to set an observation in the BN tool.

Along with the three pages of training modules (slides + example BN models), treatment group participants were also provided with a glossary of terms (see supplementary materials) to refer to when navigating the training. BN terminology was highlighted to participants in green, indicating descriptions could be found within the glossary.

One additional hour of participant time was allocated for treatment training. Both sets of training slides are available in the online materials repository.

2.4.2. Tool

For those in the treatment group, each forecast question during the tournament was accompanied by a BN model structure which represented the target game, and the games over the two weeks preceding

the target game that involved the target game teams. This model structure did not contain any filled out parameters (i.e., prior and conditional probabilities) for any of the represented games (see Fig. 5), so participants were able to enter their perceived probabilities for each game in accordance with the approach described in training (opening nodes, completing probability tables, closing node).

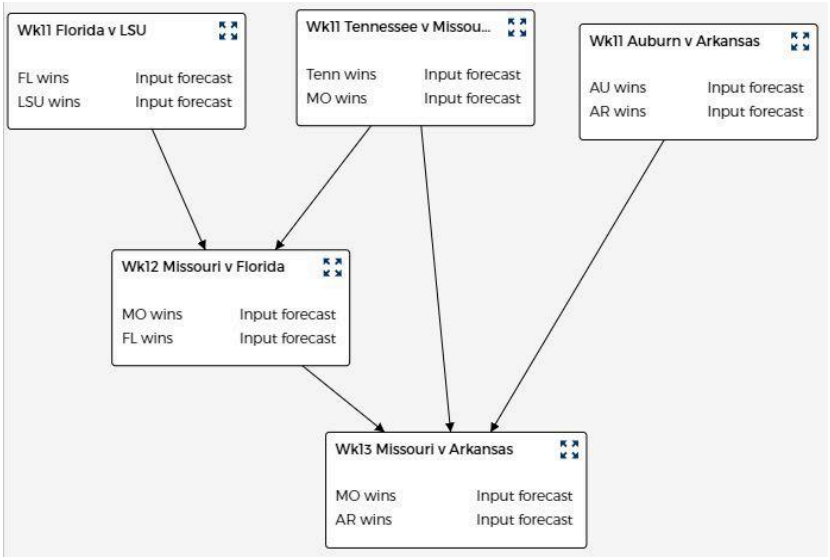

Fig. 5: Example empty BN model structure for (three weeks out) Game Week 13, between teams Missouri and Arkansas, along with preceding games that inform the target forecast from preceding game weeks.

Having completed parameterization, the BN automatically updated the network with the forecast predictions for each game, given the conditional probabilities provided (wedded to the specified model structure). As illustrated in Fig. 6, the target game forecast (bottom node) could then have the probabilities of each team winning "read off" as estimates for the forecast question.

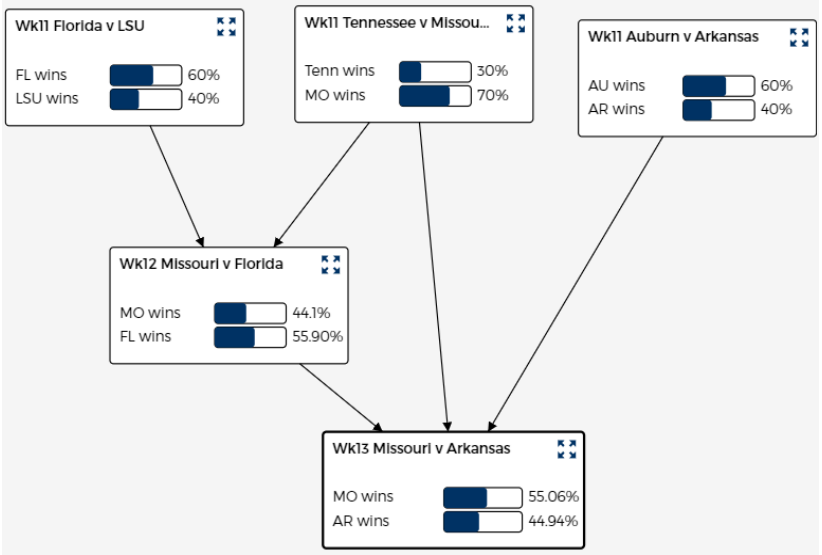

Fig. 6: Example parameterized BN model structure for (three weeks out) Game Week 13, between teams Missouri and Arkansas, along with preceding games that inform the target forecast from preceding game weeks.

In subsequent update weeks, treatment group participants were provided with their completed model. Participants could then update the model as they saw fit, such as by "observing" game outcomes from preceding weeks that had since become known (see 2.4.1 for examples of observation in the updating

section). These updates would also remain for participants when revisiting a forecast on their final week.

Along with access to their training slides, example models, and glossary of terms, the tool incorporated mouseover tooltips for embedded guidance (e.g., in explaining components of the conditional probability table), and warnings for inconsistent/incomplete conditional probability tables.

2.4.3. Forecast Questions

Participants in both groups were presented with forecasting questions in the same format, with reminders regarding the target game in question (which teams are playing, and in which week). The treatment group was additionally provided with their BN for the forecast on the same page (as outlined above), but were free to choose if they wished to use the generated forecast from the model or to provide their own adjusted forecast.

Each of the twelve target games was presented on a separate page, and was only shown to participants for weeks when the specific target game was forecastable. Fig. 7 shows the question format for an initial week (three weeks out) forecast for a target game.

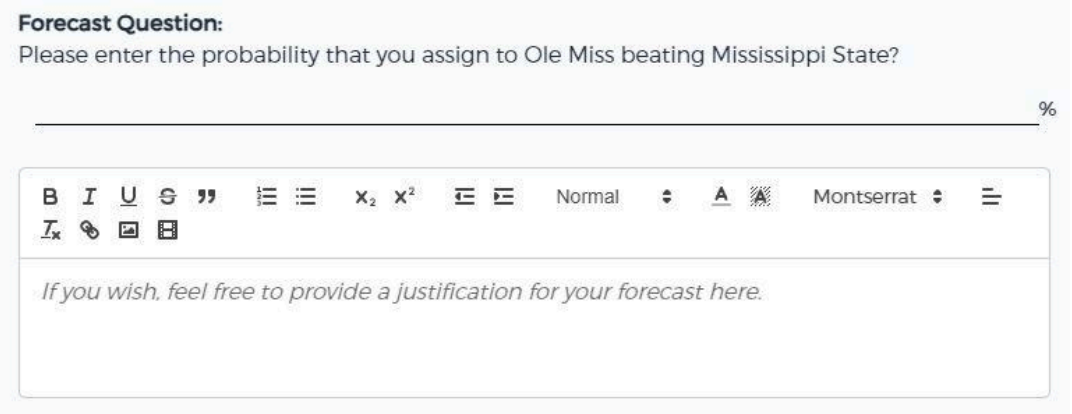

Fig. 7: Example forecast question for an initial forecast (three weeks out) for the SEC week 13 game, between teams Ole Miss (i.e., the University of Mississippi) and Mississippi State (University).

For each forecast question, participants provided their probability of one team beating the other (0-100%), and could provide open text reasoning to justify their forecast if they wished. Participants were free to move between available forecast questions before and after answering, but were instructed to complete all forecasts before submitting their responses for the week. Open text reasoning (wherein participants could provide a justification behind a given forecast) was optional.

As the weeks progressed, participants were instructed to revisit previous target games as intervening game outcomes became known. For these updates, participants revisited the same forecast question page, which included their previous forecast (and if in the treatment group, their BN model), with updated instructions regarding new update questions being provided at the bottom of the page (see Fig. 8), and a provided summary of now-known game outcomes that involve the two teams of the target forecast (in the form of statements regarding which team won and loss, and against whom).

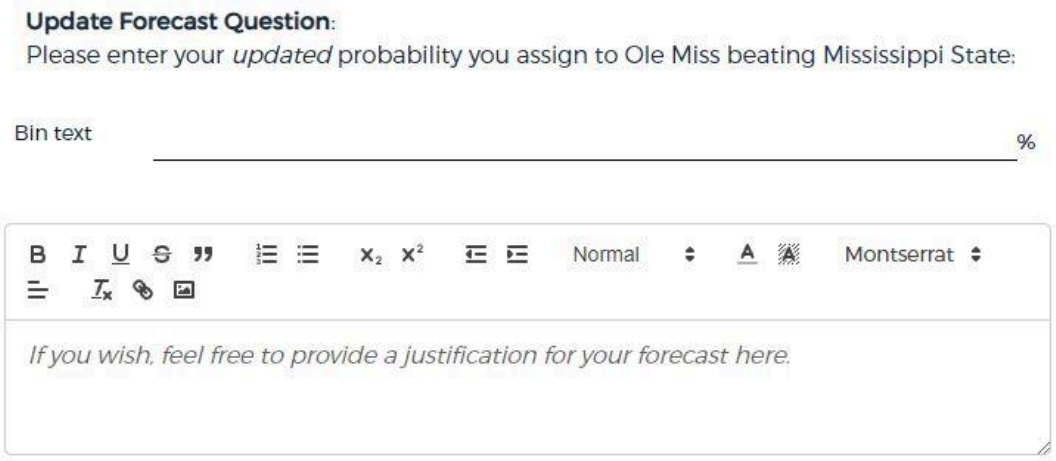

Fig. 8: Example forecast question for the first update forecast (two weeks out) for Game Week 13, between teams Ole Miss (i.e., the University of Mississippi) and Mississippi State (University).

Although participants could see prior forecasts, changing those values made no difference to recorded data for those weeks. Finally, in the last week of possible forecasting (i.e., one week out from the target game), participants were again invited to provide a final update for their forecast of the target game, with instruction indicating the new questions, a reminder of both of their previous forecasts, and summaries of relevant prior games from the intervening week (which teams won or lost, and against whom). Final update question format is shown in Fig. 9.

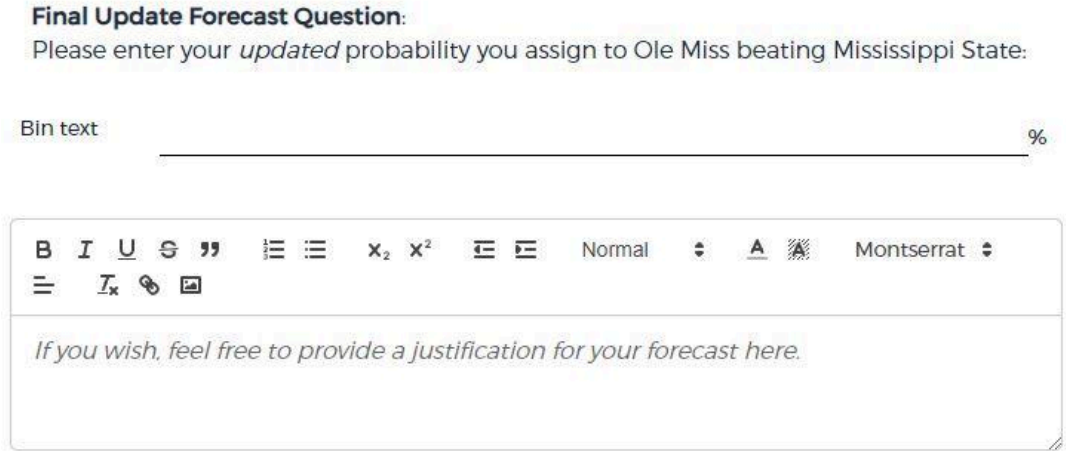

Fig. 9: Example forecast question for the final update forecast (one week out) for Game Week 13, between teams Ole Miss (i.e., the University of Mississippi) and Mississippi State University.

2.4.4. Follow-up Feedback Survey

An email invitation to take part in a feedback survey (see supplementary materials) was circulated to participants two weeks after the conclusion of the study. The survey consisted of nine questions regarding the training and forecasting task, 14 regarding the tool (for the treatment group only). These questions consisted of three formats: likert scale questions, probability scale questions and open text questions. They were as followed:

1. How easy was the content of the training to understand?
   Likert scale: *[Extremely Difficult // Difficult // Fairly Difficult // Fairly Easy // Easy // Extremely Easy]*

2. How much did you apply what you had been shown in your training to the forecasting task?
   Probability scale: *[Enter a percentage: ____ . 0 = "I applied none of it" 100 = "I applied all of it"]*

3. How much did you feel the training assisted you in:
   a. Understanding the structure of the problem/forecast?
   b. Being consistent in your reasoning process
   c. Improving the accuracy of your forecasting

   Likert scale: *[Extremely Unhelpful// Difficult// Fairly Unhelpful// Fairly Helpful// Easy// Extremely Helpful]*

4. If you were to engage in another forecasting tournament/challenge, how likely would you be to use what you have learned during the training?
   Probability scale: *[Enter a number between 0 and 100: ____. 0 = Not at all; 100 = Certainly]*

5. Is there anything you would like or suggest to improve the training?
   Open text: *[_____]*

6. How difficult did you find making the forecasts?
   Likert scale: *[Extremely Difficult // Difficult // Fairly Difficult // Fairly Easy // Easy // Extremely Easy]*

7. How accurate do you believe your forecasts were?
   Probability scale: *[Enter a number between 0 and 100: ____. 0 = Completely Inaccurate; 50 = Just Guessing; 100 = Completely Accurate]*

8. How consistently do you think you applied the same process or approach in your forecasting?
   Probability scale: *[Enter a number between 0 and 100: ____. 0 = completely inconsistent across my forecasts; 50 = consistent across half of my forecasts; 100 = Consistent across all more forecasts]*

9. If you didn't finish this study, can you help us understand why and what we (the research team) might be able to do to improve retention in the future? We do not blame you for not finishing - we just want to improve the way we run these studies. If you did finish, we would also welcome your views on how to improve retention.
   Open text: *[_____]*

The survey respondents who were in the treatment group during the study were asked to answer 14 supplementary questions:

10. Of the 5 original models you were given in the first week of the study, how many did you complete in that first week?
    [Enter a number:_____]

11. Of the 2 new models you were given in the second week of the study, how many did you complete in that second week?
    [Enter a number:_____]

12. Of the 5 new models you were given in the third week of the study, how many did you complete in that third week?
    [Enter a number:_____]

13. If you did not consistently complete the parameters required, what was the reason for this?
    Open text: *[_____]*

14. Of the 5 original models you were given in the first week, how many did you update in the second week to inform your updated forecasts for those questions for that week?
    [Enter a number:_____]

15. Of the 7 models released in the first two weeks, how many did you update in the third week to inform your updated forecasts for those questions for that week?
    [Enter a number:_____]

16. If you did not consistently update your models, what was the reason for this?
    Open text: *[_____]*

17. How consistently did you use the forecast generated by the model as your submitted forecast for that game in the first week?
    [Enter a number between 0 and 100: _____ 0 = Never 100 = Every time]

18. How consistently did you use the forecast generated by the model as your submitted forecast for that game in the second week?
    [Enter a number between 0 and 100: _____ 0 = Never 100 = Every time]

19. How consistently did you use the forecast generated by the model as your submitted forecast for that game in the third week?
    [Enter a number between 0 and 100: _____ 0 = Never 100 = Every time]

20. If you did not consistently use the forecast generated by the model, what was the reason for this?
    Open text: *[_____]*

21. How much did you feel the tool / models assisted you in:
    a. Understanding the structure of the problem/forecast?
    b. Being consistent in your reasoning process?
    c. Improving the accuracy of your forecasting?

    Likert scale: *[Extremely Unhelpful// Difficult// Fairly Unhelpful// Fairly Helpful// Easy// Extremely Helpful]*

22. If you were to engage in another forecasting tournament/challenge, how likely would you be to use a provided model to assist your forecasting?
    [Enter a number between 0 and 100: _____ 0 = Not at all likely; 100 = Certainly]

23. Is there anything you would like or suggest to improve the tool, or the models themselves?
    Open text: *[_____]*

2.5. Hypotheses

Consequently, our research question regarding whether the BN tool resolves previous negative user perceptions, and whether BNs improve forecasting, and can be formulated into the following hypotheses (note: when referring to positive perceptions/ratings below, the implied test is significant deviation from the midpoint of the scale - i.e., neutral - in the positive direction):

Feedback/Training Perceptions Hypotheses

1. Participants from the treatment group perceive the training positively across dimensions of helpfulness, usability, and applicability.
2. Participants from the control group perceive the training positively across dimensions of helpfulness, usability, and applicability.

3. Participants from the treatment group perceive the training as positively as the participants of the control group despite added complexity of the modules 2 and 3.
4. Participants from both treatment and control groups report a high likelihood of using the training in future forecasting scenarios.

Feedback/BN Tool Perceptions Hypotheses

5. The BN tool will be perceived positively across dimensions of helpfulness, usability, and applicability by the treatment group.
6. Participants from the treatment group will report a high likelihood of using the tool in future forecasting scenarios.

Feedback/Task Perceptions Hypotheses

7. Participants from the treatment group will perceive the forecasting task as being easier than participants from the control group.

Forecasting Performance Hypotheses

8. Forecasting Accuracy (as measured by Brier Scores) will be superior (i.e., Brier scores will be lower) in the treatment group than in the control group.

Results

3.1. Data Processing

All 36 forecasts (12 initial, 24 updates) were processed and analyzed, if the participant successfully completed all study weeks (i.e., the 56 who completed the study—24 from the control condition and 32 from the treatment condition). This resulted in a total of 1,152 treatment group forecasts, and 864 control group forecasts.

All 38 responses to the follow-up feedback survey were collected and anonymized. All quantitative data was processed for statistical analysis. For the purposes of this manuscript, text from open text feedback questions have not been analyzed.

3.2. Training, Tool, and Task Perceptions

A total of 38 participants responded to the feedback (25 respondents from the treatment group and 13 from the control group). The survey was distributed two weeks after the study's completion. Participation was voluntary, and additional compensation was provided for completion of the survey (i.e., $15).

We conducted independent samples T-tests (corrected where necessary) for between group comparisons, and replaced one sample T-tests with their Wilcoxon non-parametric alternative for normality-violated single group comparisons. As with Brier score data, attempted transformations (e.g., Log10, ArcSinSqrt, Inverse) were not effective in normalizing the data.

3.2.1. Training Perceptions (Hypotheses 1, 2, 3, 4)

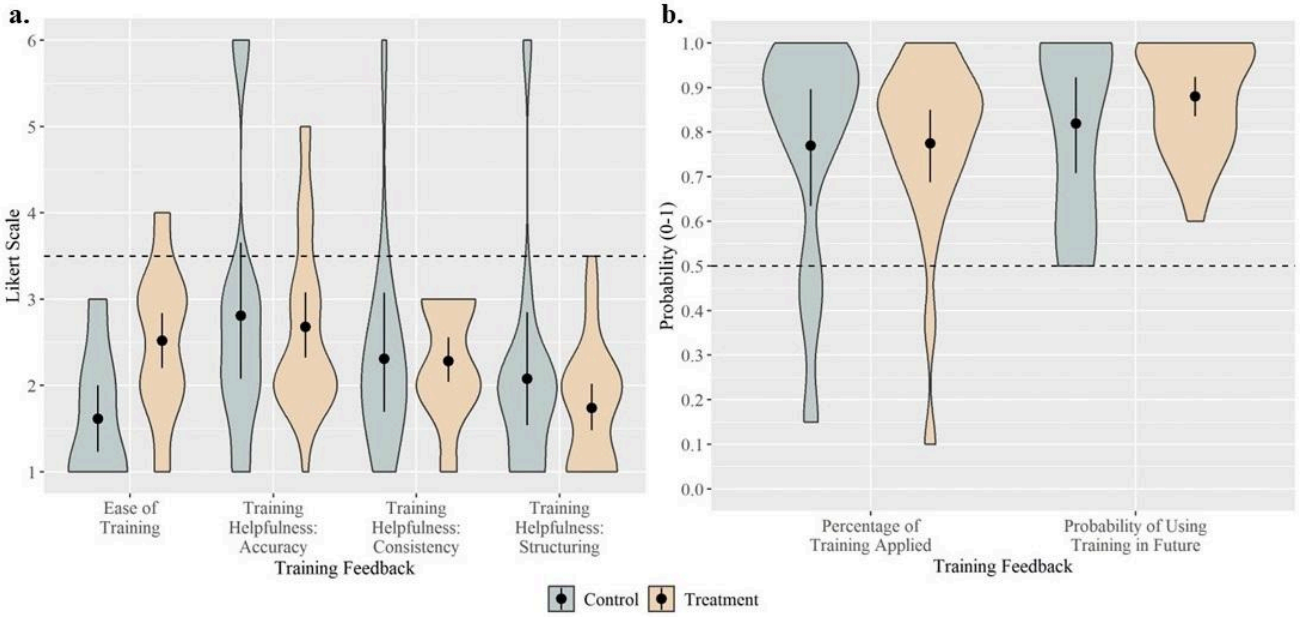

Fig. 13: Participant perceptions of training from feedback survey. Mean points plotted, with crossbars reflecting non-parametrically bootstrapped 95% confidence intervals. Dashed horizontal lines reflect scale mid-points. ***a (left)****. Six point likert scale measures from Extremely [Positive] (Easy, Helpful), 1, to Extremely [Negative] (Difficult, Unhelpful), 6.* ***b (right)****. Probability scale measures from 0-1.*

Across perceptions of training, we found no evidence for differences between control and treatment group training, with the exception of perceived ease of training, for which we found a significant effect for the control group training being perceived as easier than treatment group training, $t(27.33) = -3.286$, $p < .001$, $d = -1.101$.

Focussing on treatment group training perceptions, we found significant positive perceptions across *all dimensions* (based on Wilcoxon tests relative to scale-midpoints test values; $N = 25$): ease of training, $V = 21.00$, $p < .001$, helpfulness of training for forecast structuring, $V = 0.00$, $p < .001$, helpfulness of training for forecast consistency, $V = 0.00$, $p < .001$, helpfulness of training for forecasting accuracy, $V = 49.00$, $p < .001$, percentage of training applied to forecasting, $V = 279.50$, $p < .001$, and the probability of using the training in future, $V = 325.00$, $p < .001$.

Similarly, we find significant positive perceptions of control group training across ease of training ($N = 13$), $V = 0.00$, $p < .001$, helpfulness for forecast structuring, $V = 11.00$, $p = .007$, helpfulness for forecasting consistency, $V = 12.00$, $p = .009$, percentage of training applied to forecasting, $V = 72.00$, $p = .005$, and the probability of using the training in future, $V = 66.00$, $p < .002$. However, we did not find a significant positive perception of control group training for helpfulness for forecasting accuracy ($p = .06$).

3.2.2. Tool Perceptions (Hypotheses 5, 6)

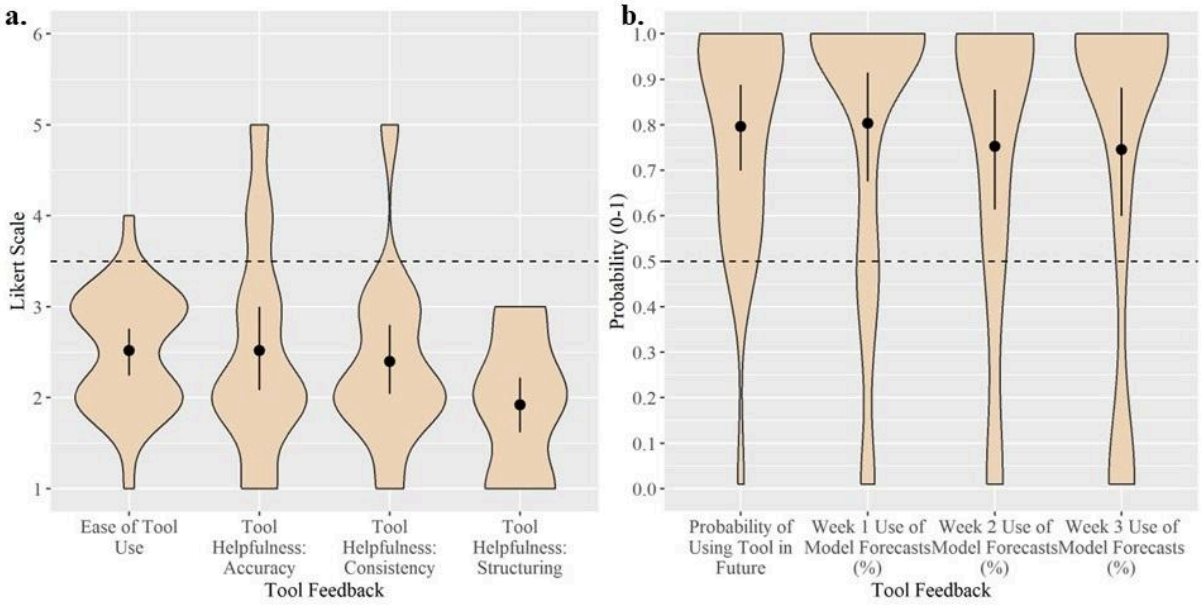

Fig. 14: Participant perceptions of the BN tool from feedback survey. Mean points plotted, with crossbars reflecting non-parametrically bootstrapped 95% confidence intervals. Dashed horizontal lines reflect scale mid-points. ***a (left)****. Six point likert scale measures from Extremely [Positive] (Easy, Helpful), 1, to Extremely [Negative] (Difficult, Unhelpful), 6.* ***b (right)****. Probability scale measures from 0-1.*

As the treatment group alone were asked about their perceptions of the BN tool, here we only compare perceptions within that group to the thresholds of positivity (mid-points in both likert and probability scales). Consequently, we find significant positive perceptions across all tool dimensions: ease of tool use, $V = 7.00$, $p < .001$, tool helpfulness for forecasting accuracy, $V = 43.50$, $p < .001$, tool helpfulness for forecasting consistency, $V = 30.00$, $p < .001$, tool helpfulness for forecasting structuring, $V = 0.00$, $p < .001$, the use of the tool in week 1, $V = 257.00$, $p < .001$, week 2, $V = 259.00$, $p < .001$, and week 3 forecasting, $V = 259.00$, $p < .001$, and the probability of using the tool in future, $V = 240.00$, $p < .001$.

3.2.3. Task Perceptions (Hypothesis 7)

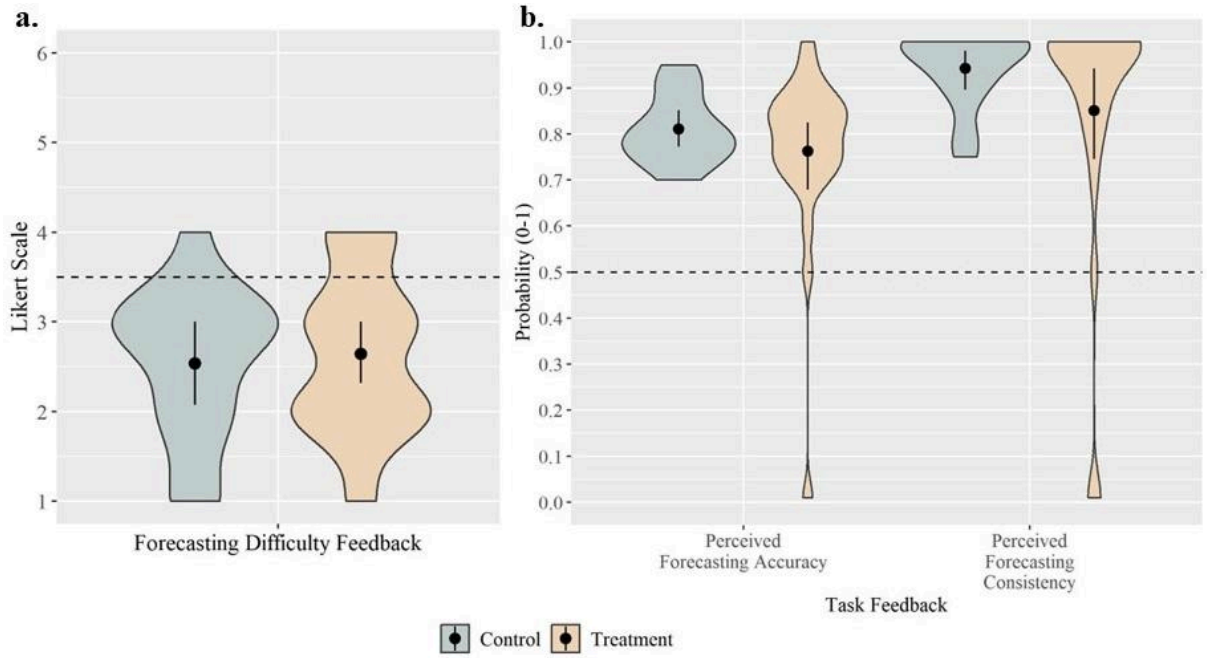

Fig. 15: Participant perceptions of forecasting task from feedback survey. Mean points plotted, with crossbars reflecting non-parametrically bootstrapped 95% CIs. Dashed horizontal lines reflect scale mid-points. ***a (left)****. Six point likert scale measures from Extremely Easy (1), to Extremely Difficult (6).* ***b (right)****. Probability scale measures from 0-1.*

Across perceptions of the forecasting task, we found no significant differences between control and treatment groups.

Among treatment group participants, we found significant positive perceptions of forecasting difficulty (< 3.5), $V = 35.00$, $p < .001$, self-perceived forecasting accuracy (> .5), $V = 277.00$, $p < .001$, and self-perceived forecasting consistency (>.5), $V = 279.00$, $p < .001$. Among control group participants, we also found significant positive perceptions of forecasting difficulty (< 3.5), $V = 4.50$, $p = .002$, self-perceived forecasting accuracy (> .5), $V = 91.00$, $p < .001$, and self-perceived forecasting consistency (>.5), $V = 91.00$, $p < .001$.

The analysis of the training, tool and task perceptions highlights significant positive feedback from both the treatment group and the control group, suggesting effective training delivery across various dimensions including helpfulness, usability and applicability. Notably, while the treatment group found their training to be harder (which was expected considering the more challenging material on BN), they still reported positive perceptions.

Most importantly, the BN tool was overwhelmingly received with positive reviews from the treatment group, demonstrating high ratings in usability, helpfulness and applicability. This aligns well with our hypotheses 5 and 6, affirming the tool's strong usability and potential for future use scenarios.

While, contrary to our hypothesis 7, no significant differences were noted in the perception of task difficulty between groups, the overall high engagement and positive feedback underscore the success of the training and BN tool in enhancing user experience using BN reasoning in forecasting tasks.

3.3. Forecast Accuracy (Hypothesis 8)

Forecast estimates of target game win likelihood (i.e., "Please enter the probability you assign to team X beating team Y") were converted into proportion format (from a 0-100% value to 0-1 value) and compared to actual game outcomes, such that a Brier score of each forecast could be calculated via:

$$Brier = (Forecast - Outcome)^2$$

Where the *Forecast* ranges from 0 (outcome definitely does not happen) to 1 (outcome definitely does happen), and *Outcome* is classified as either 1 (event happened) or 0 (event did not happen). Thus, lower scores reflect a lower squared error. In this way, scores can be averaged to acquire the mean squared error of the forecast or forecaster.

As a result, each participant had a Brier score for each game forecast, across three time-points (three weeks out/week 1; two weeks out/week 2; and one week out/week 3). Across the 12 games being forecast this resulted in 36 Brier scores per participant. The mean Brier score for each participant was used to calculate performance bonuses, split by group. For all forecasting accuracy analyses, aggregations of Brier scores are minimized (unless otherwise specified), such that maximum data points are preserved. Further, Brier scores have homogeneity of variance and non-normality considerations. For variance, all analyses have appropriate corrections applied. For non-normality, transformations have first been attempted (e.g., Log10 and ArcSinSqrt), but non-normality is not resolved. Where relevant, non-parametric analyses are deployed, though we note that between group comparisons using independent samples t-tests are robust to non-normality violations and remain in use (Rasch et al., 2007).

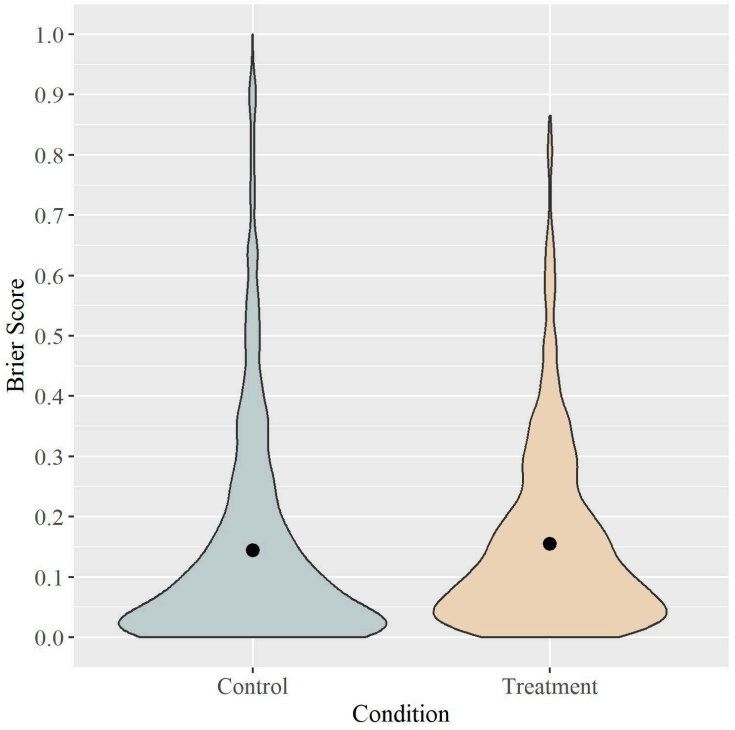

Fig. 10: Violin plots displaying the distribution of Brier scores between the control and treatment groups. Points reflect non-parametrically bootstrapped mean 95% confidence intervals.

Accuracy. Via a one-way (treatment group brier scores < control group brier scores) Welch-corrected independent sample T-test (which accounts for an assumption of different underlying population

variances, compromising some power), we do not find a significant accuray advantage in the treatment group, $t(2099.67) = 1.509$, $p = .934$. We therefore reject Hypothesis 8 (the treatment group forecasts will be more accurate than control group forecasts).

4. Discussion

Our study's findings highlights the successful deployment of a BN training tool that was well-received across various dimensions of user engagement and perceived effectiveness. Our feedback survey data reveals our BN training and tool are considered significantly positive across all dimensions of interest: user friendliness; percentage engagement/use (i.e., actual use of training/tool for forecasting itself); helpfulness for forecasting structuring, accuracy, and consistency; intended future use; and perceptions of forecasting difficulty, accuracy, and consistency. In corroboration of this, we find study retention rates in our treatment group either on par with the control group over the 5 week period (68% retention vs 63% retention, respectively), or more so from study to feedback participation (78% to 54%, respectively).

In addition, BN training is considered equivalently as positive as control group training on forecasting and calibration. These results serve as validation for the careful attention paid throughout training and tool design to overcome previous barriers to positive user engagement.

Despite these positive perceptions, our analysis did not reveal a significant difference in forecasting accuracy between the treatment and control groups as measured by Brier scores. This outcome may be attributed to a potential ceiling effect, where participant forecasting of college football game wins and losses was already near-optimal, given the degree of predictability in those games. Alternatively it remains possible that both the specialized BN training and the basic training provided to the control group already improved forecasting abilities in this domain to a near performance ceiling, leaving little room for observable improvement through additional BN training.

Understanding this effect, if present, is crucial for designing future studies and may point towards the need for a more varied training setting or different forecasting contexts to truly discern the impact of the BN training tool. Further research should explore more variable complex forecasting environments where the capabilities of BN may be more visible. in this area would be beneficial.

Consequently, these initial results provide a solid foundation from which to build up and refine our training and tool to further improve and generalize its forecasting enhancement capabilities.

5. Conclusion

In this manuscript, we described an RCT study that investigated the efficacy of a BN tool designed to enhance forecasting skills amongst university students at a U.S. institution with a strong football culture. A total of 85 participants were divided into two groups, with the control group receiving basic probabilistic reasoning and calibration training, while the treatment group received additional training in using BN reasoning and were provided with a BN tool to use during the forecasting phase. Subsequently, participants engaged in forecasting the outcomes of American football games over a three-weeks period. Although the results did not demonstrate a significant improvement in forecasting

accuracy for the treatment group using the BN training and tool, the feedback was exceptionally positive regarding its ease and lightness of use.

Participants' perception of the BN training and tool – recognizing its value for structuring their forecast and the clarity it brought to understand probabilistic scenarios – suggest that there is a foundation for its utility in more complex forecasting tasks.

Looking ahead, TFI intends to continue developing and refine the BN tool, focusing on more sophisticated features such as enhanced structuring capabilities. We remain committed to maintaining its usability and accessibility for users without extensive statistical backgrounds. By iteratively enhancing and testing these features, we hope to realize the full potential of BNs in improving decision-making across varied and complex forecasting scenarios.

Acknowledgements

Funding: Funding for the study was provided by the Transformative Futures Institute, a nonprofit research organization based in Wichita, Kansas, United States.

Human subjects: This study was approved by Auburn University's Institutional Review Board, Protocol ID: #23-448 EX 2309, Protocol Title: A Tool for Enhancing Reasoning and Improving Forecasting Accuracy. All participants acknowledged an approved Institutional Review Board information letter demonstrating informed consent prior to beginning the survey.

Author contributions: Conceptualization: RG, DP, CM; Methodology: RG, TP, DP, CM, AB; Formal analysis: TP; Data curation: TP, AB, CC; Investigation: TP, CC; Visualization: TP; Project administration: RG, CM; Supervision: RG, TP; Writing—original draft: TP, RG; Writing—review & editing: All authors.

Data and materials availability: Anonymized data from this study can be found at <TODO: insert link>.

References

Abramson, B., Brown, J., Edwards, W., Murphy, A. and Winkler, R.L. (1996). Hailfinder: A Bayesian system for forecasting severe weather. *International Journal of Forecasting*, 12(1), pp.57-71.

Boneh, T., Weymouth, G. T., Newham, P., Potts, R., Bally, J., Nicholson, A. E., & Korb, K. B. (2015). Fog forecasting for Melbourne Airport using a Bayesian decision network. *Weather and Forecasting*, 30(5), 1218-1233.

Cofino, A. S., Cano, R., Sordo, C., & Gutierrez, J. M. (2002). Bayesian networks for probabilistic weather prediction. In *15th European Conference on Artificial Intelligence (ECAI)*.

Constantinou, A. C., Fenton, N., Marsh, W., & Radlinski, L. (2016). From complex questionnaire and interviewing data to intelligent Bayesian network models for medical decision support. *Artificial intelligence in medicine*, 67, 75-93.

Cruz, N., Desai, S. C., Dewitt, S., Hahn, U., Lagnado, D., Liefgreen, A., ...Pilditch, T. D., & Tešić, M. (2020). Widening access to Bayesian problem solving. *Frontiers in psychology*, 11, 521775.

Fenton, N., & Neil, M. (2010). Comparing risks of alternative medical diagnosis under Bayesian arguments. *Journal of Biomedical Informatics*, 43, 485-495.

Fenton, N., & Neil, M. (2018). *Risk assessment and decision analysis with Bayesian networks (2nd Ed.)*. CRC Press.

Gilio, A., & Sanfilippo, G. (2014). Conditional random quantities and compounds of conditionals. *Studia Logica*, 102(4), 709-729.

Gruetzemacher, R., 2022. Bayesian networks vs. conditional trees for creating questions for forecasting tournaments. Bayesian Modeling and Applications Workshop, at the Conference on Uncertainty in Artificial Intelligence.

Hastie, R., & Kameda, T. (2005). The robust beauty of majority rules in group decisions. *Psychological review*, *112*(2), 494.

Juslin, P., Nilsson, H., and Winman, A. (2009). Probability theory, not the very guide of life. *Psychol. Rev.* 116, 856–874.

Korb, K. B., & Nicholson, A. E. (2011). *Bayesian artificial intelligence (2nd. Ed.)*. CRC Press.

Nicholson, A. E., Woodberry, O., Mascaro, S., Korb, K., Moorrees, A., & Lucas, A. (2011, April). ABC-BN: A tool for building, maintaining and using Bayesian networks in an environmental management application. *Proceedings of the 8th Bayesian Modelling Applications Workshop*, 818, 108-116.

Nyberg, E.P., Nicholson, A.E., Korb, K.B., Wybrow, M., Zukerman, I., Mascaro, S., Thakur, S., Oshni Alvandi, A., Riley, J., Pearson, R. and Morris, S. (2022). BARD: A structured technique for group elicitation of bayesian networks to support analytic reasoning. *Risk Analysis*, 42(6), pp.1155-1178.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, US: Morgan-Kaufmann.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. NY, US: Cambridge University Press.

Pettigrew, R. (2016): *Accuracy and the laws of credence*. Oxford, UK: Oxford University Press.

Politzer, G. (2016). Deductive reasoning under uncertainty: A water tank analogy. *Erkenntnis*, 81(3), 479-506.

Rasch, D., Teuscher, F., & Guiard, V. (2007). How robust are tests for two independent samples?. *Journal of statistical planning and inference, 137(8)*, 2706-2720.

Rehder, B., and Waldmann, M. R. (2017). Failures of explaining away and screening off in described versus experienced causal learning scenarios. *Mem. Cogn.* 45, 245–260.

Rottman, B. M., and Hastie, R. (2016). Do people reason rationally about causally related events Markov violations, weak inferences, and failures of explaining away. *Cogn. Psychol.* 87, 88–134

Smit, N. M., Lagnado, D. A., Morgan, R. M., & Fenton, N. E. (2016). Using Bayesian networks to guide the assessment of new evidence in an appeal case. *Crime Science*, 5(9). 10.1186/s40163-016-0057-6

Tetlock, P. E., & Gardner, D. (2016). *Superforecasting: The art and science of prediction*. Random House.

Supplementary Materials

Recruitment Messages

First Solicitation

October 2023

Dear AU student,

I am a professor in the Department of Business Analytics and Information Systems at Auburn. I would like to invite you to participate in my research study on assessing an online tool for improving probabilistic reasoning by forecasting 2023 SEC football games. You may participate if you are a junior or senior-level undergraduate student or a master's level graduate student.

If you choose to participate, you may be asked to complete a training module for the tool ranging from ~30 minutes to ~60 minutes, depending on your pace. You will have until October 28th to complete the training and the first round of forecasts. You will be assigned SEC football games to forecast for the last three weeks of the SEC football season: Week 11, Week 12, and Week 13. This is a total of 15 games, but you will need to begin forecasting these games at least two weeks beforehand.

For participating and completing all assigned forecasts and surveys about your experience you will be compensated with a $100 Visa gift card. If your scores for the tournament are in the top 50% of all participants, you will receive an additional $20. If your scores for the tournament are in the top 25% of all participants, you will receive an additional $50. (Note: the compensation will be distributed based on your performance relative to other participants who received the same type and amount of calibration training as you.)

Your predictions and training results will be kept confidential. Only your email address will be retained in order to contact you to dispense your compensation. Your email address will be deleted after the study. If you would like to know more information about this study, an information letter can be found at this link. If you choose to participate after reading the letter, you can begin the study by clicking here.

If you have any questions, please contact me at dparadice@auburn.edu.

Thank you for your consideration,

David Paradice
Business Analytics and Information Systems
Harbert College of Business
Auburn University

Updated Solicitation

Dear AU student,

I am a professor in the Department of Business Analytics and Information Systems at Auburn. I would like to invite you to participate in my research study on assessing an online tool for improving probabilistic reasoning by forecasting 2023 SEC football games. You may participate if you are a junior or senior-level undergraduate student or a master's level graduate student.

If you choose to participate, you may be asked to complete a training module for the tool ranging from ~30 minutes to ~60 minutes, depending on your pace. You will have until October 28th to complete the training and the first round of forecasts. You will be assigned SEC football games to forecast for the last three weeks of the SEC football season: Week 11, Week 12, and Week 13. This is a total of 15 games, but you will need to begin forecasting these games at least two weeks beforehand.

For participating and completing all assigned forecasts and surveys about your experience you will be compensated with a $100 Visa gift card. If your scores for the tournament are in the top 50% of all participants, you will receive an additional $20. If your scores for the tournament are in the top 25% of all participants, you will receive an additional $50. (Note: the compensation will be distributed based on your performance relative to other participants who received the same type and amount of calibration training as you.)

Your predictions and training results will be kept confidential. Only your email address will be retained in order to contact you to dispense your compensation. Your email address will be deleted after the study. If you would like to know more information about this study, an information letter can be found at this link. If you choose to participate after reading the letter, you can begin the study by clicking here.

If you have any questions, please contact me at dparadice@auburn.edu.

Thank you for your consideration,

David Paradice
Business Analytics and Information Systems
Harbert College of Business
Auburn University

Mid-study participation email reminder

Hello!

You're receiving this email because you registered for the SEC Football Forecasting Challenge on quorumapp.com and your participation last week qualifies you to continue forecasting this week. First and most importantly: thank you so much for your thoughtful engagement with this challenge so far. We hope you've found it interesting.

When you return to the study, you'll find three changes:

- First, the five week 11 forecasts now have a final section at the bottom for your final forecast update, in light of the week 10 game results.
- Second, the two week 12 forecasts now have a forecast update at the bottom in light of the week 10 game results
- Third, there is a new set of 5 forecast questions for the week 13 games. These will be your final forecasts

To qualify for next week, you must **complete ALL the new update questions (as specified above), and the 5 new forecast questions for week 13**. You must do this no later than 11am CT on Saturday, November 11th.

Note that you are permitted to revisit and revise your forecasts throughout the week to account for new information. The last forecast you submit before Saturday November 11th at 11am CT will be the one we evaluate for accuracy and bonus prizes.

Finally, keep in mind that those who complete all requirements **for all four weeks** are entitled to $100 in base compensation. The top 50% most accurate forecasters will receive an additional $20. And the top 25% most accurate forecasters will receive an additional $50.

To get back to forecasting, go to quoruampp.com.

Thank you so much, once again, for your time and effort. Good luck – and, as always, War Eagle!!

Amory Bennett (on behalf of the Transformative Futures Institute)
Technical Assistance
quorumapp.com

Feedback Survey Recruitment email

Hello!

You're receiving this email because you registered for the SEC Football Forecasting Challenge on Quorum, a study that was run by the Transformative Futures Institute (TFI) in October and November of this year.

We're reaching out because the study has concluded and we have one more 30min (max!) survey about your experience during the challenge. **Those who complete this survey before December 20th will get $15.** The person who submits the most useful feedback -- as judged by the TFI research team -- will earn an extra $50.

You do not need to have completed the study to qualify for payment for this final survey. You just need to have (at minimum) participated during the first week (when we provided training). We want to learn, among other things, how to improve retention in future studies like this, so we want to hear from you even if you dropped out!

Here is the survey. We'll pay base comp ($15) by December 27 (within a week of the due date). We may take more time to decide who gets the $50 for the most useful feedback - but we'll do our best to be quick!

Huge thanks, once again, for participating, to whatever extent you participated -- and thanks in advance to those of you who take the time to respond to this survey!

Amory Bennett (on behalf of the Transformative Futures Institute)
Technical Assistance
quorumapp.com

Methods

Table: Methods.1. Glossary of terms for treatment group training.

Prediction	Reasoning "forwards" from likely causes to probable effects.
Diagnosis	Reasoning "backwards" from observed effects to probable cause.
Claim/hypothesis	A belief/topic under investigation. Often, we are determining whether this is true or false. Here these are football game outcomes.
Evidence	Relevant pieces of information (e.g., prior games) that can inform a claim/hypothesis.
Variable	A hypothesis/cause/claim or item of evidence (think of this as the name/identifier representation). Here these are the names of games.
State	The possible values (or outcomes) a variable can occupy: E.g., True or False; Present or Absent. Here these are the outcomes of games (Win; Loss/Draw).
Node	The representation of a variable and its states in the Bayesian Network formalism.
Relation	The link from one variable to another. It represents the possible relationship between the two variables. Relations are directional, and flow from cause to effect (e.g., Claim Variable -> Evidence Variable).
Parent/Child Variable	A way of describing two variables in relation to one another. The relation flows from a "parent" variable, to a "child" variable. The parent occupies the role of cause, and the child occupies the role of effect.
Prior Probability	The initial/background probability of a variable state, given that variable has no parents (i.e., is a parentless claim/cause).
Conditional Probability	The probability of a state occurring in a child variable, given the (assumed) state of a parent variable.
Conditional Probability Table (CPT)	The table within a child node wherein the probability of each child variable state is specified for each combination of possible parent variable states.
True Positive Rate	The probability of observing a child variable (e.g., item of evidence) state (e.g., present, positive, true), given the target parent variable state is true (e.g., positive test, given disease present).
False Positive Rate	The probability of observing a child variable (e.g., item of evidence) state (e.g., present, positive, true), given the target parent variable state is not true (e.g., positive test, given disease absent).
Likelihood Ratio	The diagnostic value of a child variable (e.g., item of evidence) towards its parent variable(s). Calculated via (True Positive / False Positive). Indicates evidence strength.
Observation	The "setting" or observing of a variable as known to be in a particular state.
Posterior Probability	The updated probability of a variable state, given the observation of a related variable state.

Table: Methods.2. Survey Questions for both Groups

Variable	Corresponding Survey Question	Unit
EaseOfTraining	"How easy was the content of the training to understand?"	[Extremely Difficult // Difficult // Fairly Difficult // Fairly Easy // Easy // Extremely Easy]
PrcntTrainingApplied	"How much did you apply what you had been shown in your training to the forecasting task?"	[0-100% slider; 1% increments || Labels: 0% = I applied none of it; to 100% = I applied all of it]
TrainingHelpStructure	"How much did you feel the training assisted you in: [Understanding the structure of the problem/forecast?]"	[Extremely unhelpful // unhelpful // fairly unhelpful // fairly helpful // helpful // Extremely helpful]
TrainingHelpConsistency	"How much did you feel the training assisted you in: [Being consistent in your reasoning process]"	
TrainingHelpAccuracy	"How much did you feel the training assisted you in: [Improving the accuracy of your forecasting]"	
ProbUseTrainingInFuture	"If you were to engage in another forecasting tournament/challenge, how likely would you be to use what you have learnt during the training?"	[0-100% slider; 1% increments || Labels: 0% = Not at all, to 100% = Certainly]
ForecastDifficulty	"How difficult did you find making forecasts?"	[Extremely Difficult // Difficult // Fairly Difficult // Fairly Easy // Easy // Extremely Easy]
SelfForecastAccuracy	"How accurate do you believe your forecasts were?"	[0-100% slider; 1% increments || Labels: 0% = complete inaccurate; 50% = just guessing; 100% = completely accurate]
SelfForecastConsistency	"How consistently do you think you applied the same process or approach in your forecasting?"	[0-100% slider; 1% increments || Labels: 0% = completely inconsistent across my forecasts; 50% = consistent across half of my forecasts; 100% = Consistently across all my forecasts]

Table: Methods.3. Survey Questions for Treatment Group Only

Variable	Corresponding Survey Question	Unit
EaseOfTool	"How easy was the tool (and models) to understand?"	[Extremely Difficult // Difficult // Fairly Difficult // Fairly Easy // Easy // Extremely Easy]
Week1UseOfModelForecasts	"Of the 5 original models you were given in the first week of the study, how many did you complete in that first week?"	Enter Number (% score calculated out of 5 possible models)
Week2UseOfModelForecasts	"Of the 2 original models you were given in the second week of the study, how many did you complete in that second week?"	Enter Number (% score calculated out of 2 possible models)
Week3UseOfModelForecasts	"Of the 5 original models you were given in the third week of the study, how many did you complete in that third week?"	Enter Number (% score calculated out of 5 possible models)
ToolHelpStructure	"How much did you feel the tool / models assisted you in: [Understanding the structure of the problem/forecast?]"	
ToolHelpConsistency	"How much did you feel the tool / models assisted you in: [Being consistent in your reasoning process]"	[Extremely unhelpful // unhelpful // fairly unhelpful // fairly helpful // helpful // Extremely helpful]
ToolHelpAccuracy	"How much did you feel the tool / models assisted you in: [Improving the accuracy of your forecasting]"	
ProbUseToolInFuture	"If you were to engage in another forecasting tournament/challenge, how likely would you be to use a provided model to assist your forecasting?"	[0-100% slider; 1% increments || Labels: 0% = Not at all, to 100% = Certainly]