A neglected dimension of good forecasting judgment: The questions we choose also matter

Edgar C. Merkle
University of Missouri

Mark Steyvers
University of California, Irvine

Barbara Mellers and Philip E. Tetlock
University of Pennsylvania

Abstract

Forecasters are typically evaluated via proper scoring rules such as the Brier score. These scoring rules use only the reported forecasts for assessment, neglecting related variables such as the specific questions that a person chose to forecast. In this paper, we study whether information related to question selection influences our estimates of forecaster ability. In other words, do good forecasters tend to select questions in a different way from bad forecasters? If so, can we capitalize on these selections in estimating forecaster ability? To address these questions, we extend a recently-developed psychometric model of forecasts to include question selection data. We compare the extended psychometric model to a simpler model, studying its unidimensionality assumption and highlighting the unique information it can provide. We find that the model can make use of the fact that good forecasters select more questions than bad forecasters, and we conclude that question selection data can be beneficial above and beyond reported forecasts. As a side benefit, the resulting model can potentially provide unique incentives for forecaster participation.

In many areas of forecasting, question selection is an issue of considerable importance. Does a forecaster look good because he/she chose to forecast only easy questions? Should

Correspondence to Edgar C. Merkle, Department of Psychological Sciences, University of Missouri, Columbia, MO 65211. Email: merklee@missouri.edu. For data access, please contact Phillip Rescober at rescober@goodjudgment.com. This research was supported by the Intelligence Advanced Research Projects Activity (IARPA) via the Department of Interior National Business Center (DoI/NBC) Contract No. D11PC20061. The U.S. government is authorized to reproduce and distribute reprints for government purposes notwithstanding any copyright annotation thereon. The views and conclusions expressed herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. government.

we reward a forecaster for attempting difficult questions, even if her forecasts on those questions are poor? Is forecasting ability related to question choice; that is, are good forecasters better able to select questions on which they will excel? These questions cannot be answered via classical metrics such as proper scoring rules (e.g., Gneiting & Raftery, 2007), which generally assume that all forecasters have reported on all questions.

Instead of proper scoring rules, model-based approaches to forecast evaluation make it feasible to study issues related to question selection. A prime candidate is a recently-proposed psychometric model of probabilistic forecasts (Merkle, Steyvers, Mellers, & Tetlock, 2016), which is related to previously-proposed item response models for doubly-bounded variables (Bejar, 1977; Ferrando, 2001; Müller, 1987; B. Muthén, 1989; Noel & Dauvier, 2007; Samejima, 1973). This model simultaneously provides estimates of forecaster ability and of question difficulty and discrimination. For example, if a particular question has ambiguous wording, then good forecasters' judgments may be indiscernable from bad forecasters' judgments. The model can recognize this, discounting forecasters' judgments on ambiguous questions as we estimate the forecasters' general abilities across questions. Conversely, certain questions may be particularly suitable for discriminating between forecasters of different abilities. Forecaster judgments on these good questions would then be weighted more heavily, as compared to judgments on other questions.

In addition to addressing novel substantive issues, model-based forecaster assessment allows us to make explicit our assumptions related to missing forecasts. That is, by excluding (or including) question selection data from a model, we implicitly make assumptions about why forecasters do not respond to some questions. For example, the *missing completely at random* (MCAR; e.g., Little & Rubin, 2002) assumption says that missingness is independent of the data (both observed and unobserved). This assumption, which is unlikely to be fulfilled in practice, generally implies that we can ignore missing data.

The Merkle et al. (2016) models instead employed the *missing at random* assumption, whereby all observed forecasts (even those from forecasters with incomplete data) are used for model estimation. The MAR assumption states that the probability of missingness can be predicted exclusively from the observed data; if we could observe the missing data, our predictions would not improve. This assumption excludes the possibility that forecasters of greater/lesser ability differ in frequency of responding or in the types of questions that they choose. When forecaster ability is related to question selection, then models employing the MAR assumption may lead to suboptimal substantive conclusions regarding forecaster ability or question attributes.

To study question selection issues in this paper, we develop a psychometric model of forecasts that jointly accommodates question selections and reported forecasts. This model draws from the psychometric literature on explicitly modeling (as opposed to ignoring) missing data (e.g., Chang, Tsai, & Hsu, 2014; Holman & Glas, 2005; O'Muircheartaigh & Moustaki, 1999; Rose, von Davier, & Xu, 2010; Wang, Jin, Qiu, & Wang, 2012), which explores the idea that information can be gained from missing data in standardized testing contexts. Following model development, we apply the model to data from a recent forecasting tournament. This allows us to study a major model assumption related to unidimensionality of forecasting ability, and it also allows us to compare the proposed model to a previous model that employs the MAR assumption. We additionally compare the model-based estimates to other forecaster ability estimates that are based on the Brier score and illustrate the general use of question selection data in forecaster assessment.

In the pages below, we first provide technical detail on the models, starting with previous developments and continuing with novel developments. Next, we apply the model to data from a recent forecasting tournament. The application includes an examination of model assumptions, a small example that provides readers with an intuition of the model's estimates, and a larger example involving the full data. Finally, we report on a simulation that further illustrates the benefits of modeling question selection data.

Models

Assume that I forecasters each respond to some subset of J questions, with each forecaster's subset possibly being unique. Let y^*_{ij} be forecaster i's probit-transformed forecast for the realized outcome of question j (with the possibility that it is missing), and let d_{ij} be a 0/1 variable indicating whether or not y^*_{ij} is missing (0 for missing, 1 otherwise). We first briefly review the MAR model proposed by Merkle et al. (2016), and we then introduce a new model that handles the d_{ij} in addition to the y^*_{ij}.

MAR Model

The models described by Merkle et al. (2016) focused on the observed y^*_{ij}, providing estimates of forecasters' abilities and questions' difficulties and discriminations. Because the model focuses exclusively on the observed y^*_{ij}, it employs the MAR assumption.

Most of the concepts underlying the Merkle et al. (2016) model derive from the classical item response literature (e.g., Embretson & Reise, 2000; Lord & Novick, 1968; McDonald, 1999), with the application to probabilistic forecasts being relatively novel. That is, instead of being applied to binary data reflecting whether or not a student correctly answers a test question (say), the models are applied to probability judgments. The model can be written as

$$y^*_{ij}|t_{ij}, \theta_{a,i}, d_{ij} = 1 \sim N(\mu_{ij}, \sigma^2_j) \tag{1}$$

$$\mu_{ij} = \beta_{0j} + (\beta_{1j} - \beta_{0j}) \exp(-\beta_2 t_{ij}) + \lambda_j \theta_{a,i} \tag{2}$$

$$\theta_{a,i} \sim N(0, 1), \tag{3}$$

where t_{ij} is the time at which person i forecasted question j (measured as days until the question expires), $\theta_{a,i}$ is person i's forecasting ability (the a subscript stands for "ability"), and the β_j and λ_j parameters are related to item j's difficulty and discrimination, respectively.

The above model is related to a factor analysis model, with extra parameters (the βs) that allow question difficulty to change over time. This is necessary because forecasters often report on a question at different points in time, and information relevant to the question changes over time. For example, imagine two forecasters predicting the chance of rain for February 1. A forecaster responding on January 31 will have a natural advantage over a forecaster responding on January 28, because the question is easier on January 31. The model can account for this issue by allowing difficulty to change over time, based on the way that the full group of forecasters is responding over time.

Merkle et al. (2016) used Bayesian methods to fit the above model to data from a forecasting tournament (data from the same source used in this paper, further described below) and found that (i) the model could successfully predict out-of-sample forecasts; (ii) the forecaster ability estimates were more highly related to a forecaster's future ability, as compared to the Brier score, and (iii) the item parameter estimates were related to external covariates in a way that was theoretically expected. In the following section, we extend this model to handle question selection data, resulting in a model that allows for *missing not at random* (MNAR) data.

MNAR Model

The MNAR model allows for the possibility that missing data provide information about forecaster ability (and about item attributes), over and above the observed data. It is a generalization of the above model that simultaneously accounts for the missingness indicators d_{ij} and the reported forecasts y^*_{ij}. In developing the model, we adopted an approach similar to that of O'Muircheartaigh and Moustaki (1999) and Holman and Glas (2005), both of whom studied methods for handling missing data in traditional item response contexts. For each person i, we simultaneously model $2 \times J$ variables: the probit-transformed forecasts for the J items (y^*_{ij}, some of which are missing), along with the missingness indicators for the J items (d_{ij}).

The J forecast variables are all modeled in a manner similar to Merkle et al. (2016):

$$y^*_{ij}|t_{ij}, \theta_{a,i}, d_{ij} = 1 \sim N(\mu_{ij}, \sigma^2_j) \tag{4}$$

$$\mu_{ij} = \beta_{0j} + \beta_{1j} t_{ij} + \lambda_{j,1}\theta_{a,i}. \tag{5}$$

This is a simplification of the Merkle et al. (2016) model, where the "time" covariate has a linear influence on μ_{ij} instead of an exponential curve. This function is simpler than the exponential function while still allowing for curvilinear influences of time on reported forecasts (because we are modeling the probit-transformed forecasts, as opposed to the original forecasts). To identify this part of the model, we fix a single $\lambda_{j,1}$ parameter (in $j = 1, \ldots, J$) to 1.

In addition to the forecast variables, the J missingness indicators are handled via a two-factor model

$$d_{ij}|\theta_{a,i}, \theta_{r,i} \sim \text{Bernoulli}(p_{ij}) \tag{6}$$

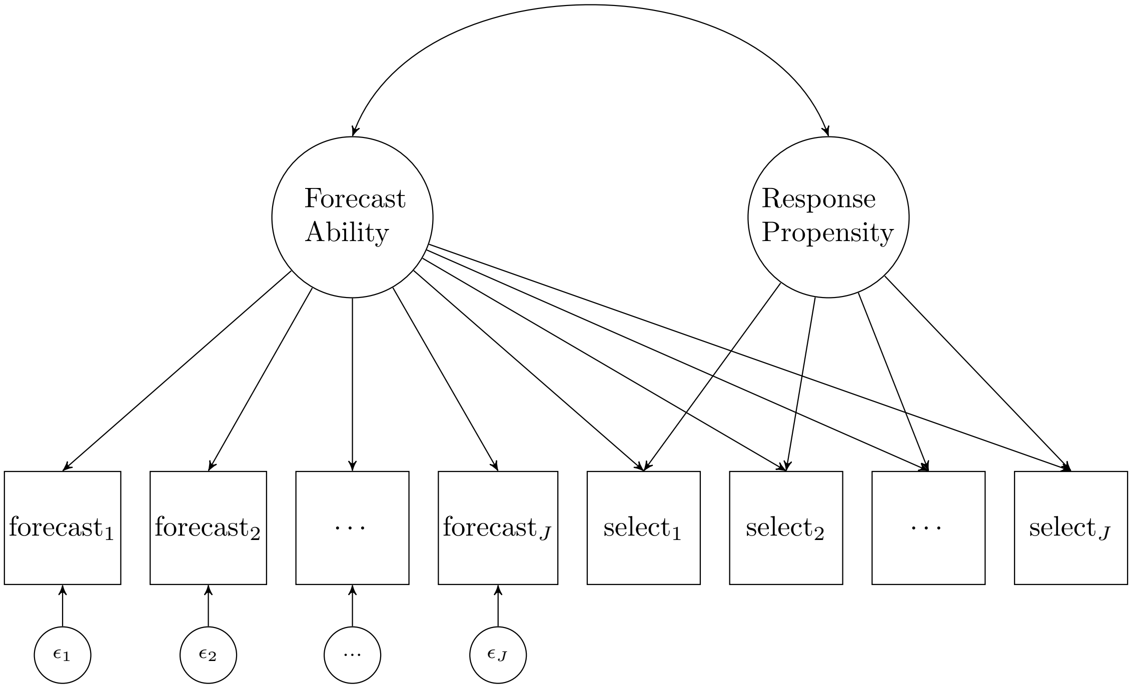
$$\text{probit}(p_{ij}) = \beta_{0,(J+j)} + \lambda_{(J+j),1}\theta_{a,i} + \lambda_{(J+j),2}\theta_{r,i}, \tag{7}$$

where $\theta_{r,i}$ is person i's response propensity. This equation implies that a person's forecasting ability can play a role in both the questions that he/she selects and the forecasts that he/she reports. There is additionally a response propensity factor that accounts for a person's general level of activity in making forecasts. The subscripts above are based on fact that the missingness variables can be treated as new questions within the model. That is, for person i, questions 1 to J include the reported forecasts, while questions $(J+1)$ to $2J$ include the binary missingness indicators (for completeness, we define the parameters $\lambda_{1,2}$ to $\lambda_{J,2}$ to all equal zero). To identify this part of the model, we fix a single $\lambda_{(J+j),2}$ parameter (where j is in $1, \ldots, J$) to 1.

Along with the above constraints, parameter identification is completed by assuming that

$$\boldsymbol{\theta}_i = (\theta_{a,i}\ \theta_{r,i})' \sim N(\mathbf{0}, \boldsymbol{D}_\psi), \tag{8}$$

Figure 1. Path diagram of the proposed model. For simplicity, the time covariate t_{ij} is excluded.



where $\boldsymbol{D}_\psi$ is a diagonal covariance matrix with unique entries ψ_1 and ψ_2. The assumption of diagonality here can potentially be relaxed, though we would require parameter constraints elsewhere in the model to trade off with this relaxation. In preliminary testing, we found that the model without diagonality was slow to converge, so we did not pursue it further in this paper. Holman and Glas (2005) show that parameter estimates under the above constraints can be linearly transformed to parameter estimates under the alternative constraints (with the diagonality assumption relaxed), implying that the parameter estimates under the two approaches are related to one another.

Estimation

The model can be represented as a path diagram, illustrated in Figure 1. Each forecaster potentially contributes $2J$ observed variables: forecasts for the J questions, along with selection indicators for the J questions. These observed variables are shown in the boxes labeled forecast$_1$ to forecast$_J$ and select$_1$ to select$_J$. The former consist of probit-transformed forecasts, with each forecast variable being observed only if the corresponding select variable equals 1. For example, a forecaster only supplies forecast$_1$ if select$_1$ equals 1.

The path diagram further shows the two latent variables labeled "forecast ability" and "response propensity," with "forecast ability" influencing both the reported forecasts and question selections. In terms of notation, the λ parameters represent the paths from the latent variables to the observed variables, the θ parameters represent the latent variables, and the β parameters (corresponding to the time covariate) are excluded for simplicity (we would have unique β parameters for each observed variable, cluttering the diagram).

To incorporate the time covariate in the model and to easily obtain θ estimates, we rely on Bayesian methods of model estimation. We specifically employ Markov chain Monte Carlo (MCMC) methods, adopting an approach that is similar to existing MCMC methods for estimating psychometric models (e.g., Ghosh & Dunson, 2009). We used the following prior distributions on classes of model parameters (subscripts are absent because the same prior was used on each free parameter):

$$\beta_0 \sim \mathrm{N}(0, 2) \tag{9}$$
$$\beta_1 \sim \mathrm{N}(0, 2) \tag{10}$$
$$\lambda \sim \mathrm{N}(0, 1) \tag{11}$$
$$\psi \sim \mathrm{Gamma}^{-1}(.01, .01) \tag{12}$$
$$\sigma^2 \sim \mathrm{Gamma}^{-1}(.01, .01), \tag{13}$$

where the second parameter of each normal distribution is a variance, as opposed to a precision.

These priors were intended to place high density in sensible parameter ranges, which can improve model convergence and sampling efficiency. The parameter ranges are sensible because the model parameters are generally used to make predictions on the probit scale, meaning that the predictions are akin to z-scores. Thus, we would be surprised to observe values of β_0 or β_1 drastically outside of $(-2, 2)$ because these values would correspond to extreme probabilities near .025 and .975, respectively. We would also be surprised to observe values of λ much larger than 1, given the diverse questions in our dataset (further discussion below). Finally, the priors on ψ and σ^2 are traditional, noninformative priors on variance parameters.

Unless otherwise mentioned, we burned in the models for three chains of 2,000 iterations each, then sampled parameters for an additional 2,000 iterations each. Chain convergence was fast and was monitored using time series plots and the Gelman-Rubin potential scale reduction statistic (Gelman & Rubin, 1992).

Parameter Interpretation

The model parameters supply many pieces of information about relationships between forecasting problems, forecaster abilities, and forecaster selection. In particular, the model allows us to address the following questions (relevant parameters in parentheses):

- Which questions are more popular than others? ($\beta_{0,J+1}$ to $\beta_{0,2J}$)
- Who are the frequent/infrequent responders? ($\theta_{r,i}$)
- Do good forecasters tend to select/avoid certain questions? ($\lambda_{J+1,1}$ to $\lambda_{2J,1}$)
- Do frequent forecasters tend to select/avoid certain questions? ($\lambda_{J+1,2}$ to $\lambda_{2J,2}$)

The first two issues are easily addressed by examining the raw data (i.e., response proportions), but the last two issues are more difficult to address via simple, data-based metrics. This is an advantage of the model-based approach described here.

Along with the above topics, the new model can address the same issues that the Merkle et al. (2016) model addressed. These include:

- Which questions are easier/harder than others? ($\beta_{0,1}$ to $\beta_{0,J}$)
- Who are the better/worse forecasters? ($\theta_{a,i}$)
- Are some questions better than others for discriminating between forecasters of varying abilities? ($\lambda_{1,1}$ to $\lambda_{J,1}$)

In the applications below, we will focus on the $\theta_{a,i}$ parameters, examining how estimated forecaster abilities change from the MAR model to the MNAR model. While we eventually fit the model to a large dataset, we initially fit the model to data from only 4 questions because it is easier to illustrate the model's behavior. First, however, we describe the data source and study the extent to which model assumptions are fulfilled.

Application: Geopolitical Forecasting

The forecasts used in this paper arise from a four-year geopolitical forecasting tournament sponsored by IARPA. The tournament involved five research teams, each of which was required to forecast hundreds of diverse questions related to world events. Example questions include

- Will Australia formally transfer uranium to India by 1 June 2012?
- Will Mario Monti resign, lose re-election/confidence vote, or vacate the office of Prime Minister of Italy before 1 January 2013?
- Will there be a significant outbreak of H5N1 in China in 2012?
- Will the Yuan to Dollar exchange rate on 31 December 2012 be more than 5% different than the 31 August 2012 exchange rate?

For each question, the research teams elicited forecasts from large groups of individuals. The teams then aggregated the forecasts via statistical methods and reported them to the funder on a daily basis.

We focus here on assessing individual forecasters who were part of the winning team in the tournament (the Good Judgment Project). This team collected forecasts from thousands of individuals, each of whom was active for at least one of the four tournament years. We first provide some background detail on the dataset (also see Mellers et al., 2014; Mellers, Stone, Atanasov, et al., 2015; Mellers, Stone, Murray, et al., 2015), and we then discuss issues of dimensionality related to the dataset. The dimensionality issues are important because the proposed model makes specific assumptions here.

Data

Adult forecasters of all ages were recruited from across the United States via email lists, professional societies, university organizations, and social media. The forecasters voluntarily logged on to a website and selected questions that they wished to forecast.

Forecasters were motivated to participate in various manners, including monetary payment for participation and leaderboards of the best forecasters.

For each question, forecasters read the question details and reported a probability of event occurrence (from 0 to 1 inclusive; forecasts of exactly 0 and 1 were transformed to .001 and .999, respectively, for modeling). Forecasters were randomly assigned to experimental conditions that differed in whether, e.g., the forecasters worked individually (vs on teams) and the types of training the forecasters received. For the purposes of this paper, we ignore experimental conditions and model only individuals' reported forecasts and question selections. This is facilitated by the fact that even forecasters who worked on teams reported their own individual forecasts.

Below, we use a data set containing 775 forecasters who each report on a subset of 157 binary (event occurs/does not occur) questions. To speed model estimation, forecasters were initially included if they responded to 70 or more questions; we later apply the model to forecasters with sparser data. While the forecasters were free to respond to the same question multiple times (i.e., to update their forecasts), we maintained only the first forecast supplied on a given question for simplicity.

Unidimensionality

The model studied in this paper assumes a single "forecaster ability" dimension and a single "response propensity" dimension, with the reported forecasts being influenced only by the "ability" dimension and the question selections being influenced by both dimensions. The assumption of a single "forecaster ability" dimension is almost certainly violated for the application considered here, which involves forecasts of diverse world events. For example, we could imagine a forecaster having expertise on a specific topic like European politics, so that his/her forecasts are better on questions related to that topic than on questions unrelated to that topic. The proposed MNAR model would assign this forecaster a single ability estimate, representing some combination of his/her ability at forecasting European politics and his/her ability at forecasting other questions. This single estimate will not be an optimal assessment of the forecaster's true ability, which requires two dimensions to fully describe (one for European politics and one for other questions). Likewise, if a forecaster's ability improves over time, his/her single model estimate will not reflect this. However, use of a single model estimate does mimic applied forecaster assessments where the Brier score is indiscriminately averaged across all available questions (e.g., Carvalho, in press).

Beyond mimicking practical assessments, we can draw on the psychometric literature to explicitly assess dimensionality. Researchers here have pointed out that, even in the case of standardized educational tests, strict unidimensionality will not hold in practice (e.g., Reise, Scheines, Widaman, & Haviland, 2013; Thissen, 2016; Zhang, 2007). Thus, considerable effort has been devoted to assessing the magnitude of unidimensionality violation, as opposed to assessing whether or not unidimensionality is violated (e.g., Bonifay, Reise, Scheines, & Meijer, 2015; Stout et al., 1996; van Abswoude, van der Ark, & Sijtsma, 2004; Zhang, 2007). This effort provides metrics that can tell us whether or not a set of questions is "unidimensional enough" to be useful. The metrics are nonparametric in nature, because model-based assessments tend to be overly sensitive to minor violations of unidimensionality.

One of the most popular metrics resulting from this literature, which we adapt to forecasting data here, is called DETECT (Zhang & Stout, 1999; Zhang, 2007). This metric makes use of the fact that, for unidimensional tests, nonzero covariances/correlations between questions should all be due to the single, underlying ability dimension. Thus, partial covariances/correlations between questions (conditioning on the single underlying dimension) should all equal zero. While this is the idea underlying DETECT, the specific algorithm is more complex than simple covariance calculation. Further computational details are provided in Appendix A.

The DETECT index is useful for our purposes because previous researchers have provided rules of thumb for its interpretation. Roussos and Ozbeck (2006) state that values below 0.2 are often taken to represent approximate unidimensionality, whereas values greater than 1.0 are taken to represent strong multidimensionality. As we move from 0.2 to 1.0, multidimensionality increases in strength. Thus, for the unidimensionality assumption to be useful, we should look for D values below 1.0, with values closer to 0 being better.

We computed this statistic separately for the reported forecasts y and for the question selections d. For the question selections, the DETECT index indicated strong multidimensionality, achieving a maximum value of 2.3 at 2 clusters (subgroups) of questions. The subgroups detected here had a strong temporal component: when we re-computed the index using only data from a single year, the maximum DETECT value was 0.58 (indicating moderate multidimensionality). For the reported forecasts, we obtained a maximum DETECT statistic of 0.72 at 3 subgroups.

These results provide some evidence that, for this particular dataset, multidimensionality is moderate and results from changes in the forecasters over time, as opposed to forecasters having specific expertise or interest in particular question topics. To address these findings, we later fit the model to subsets of the data arising from only a single year of the tournament and compare it to a model fitted to the full dataset.

Simple Example

For an initial example of the proposed model's behavior, we use data from only four questions. The four questions used here (with identification numbers in parentheses) were all open during 2012–2013; they are:

- Will Traian Basescu resign, lose referendum vote, or vacate the office of President of Romania before 1 April 2012? (1067)
- Will Kim Jong-un resign or otherwise vacate the office of Supreme Leader of North Korea before 1 April 2013? (1106)
- Before 1 April 2013, will the Egyptian government officially announce it has started construction of a nuclear power plant at Dabaa? (1147)
- Will Mohammed Morsi cease to be President of Egypt before 1 April 2013? (1177)

In the tournament, 771 of the 775 forecasters in our dataset responded to at least one of the four questions. We use all data supplied from these 771 forecasters, including missing observations. Below, we further describe the questions and the model before examining the results.

Table 1

Brier scores and response rates of four questions.

Question	Mean Brier	Response Rate
1067	.08	.87
1106	.01	.77
1147	.09	.59
1177	.07	.59

Table 2

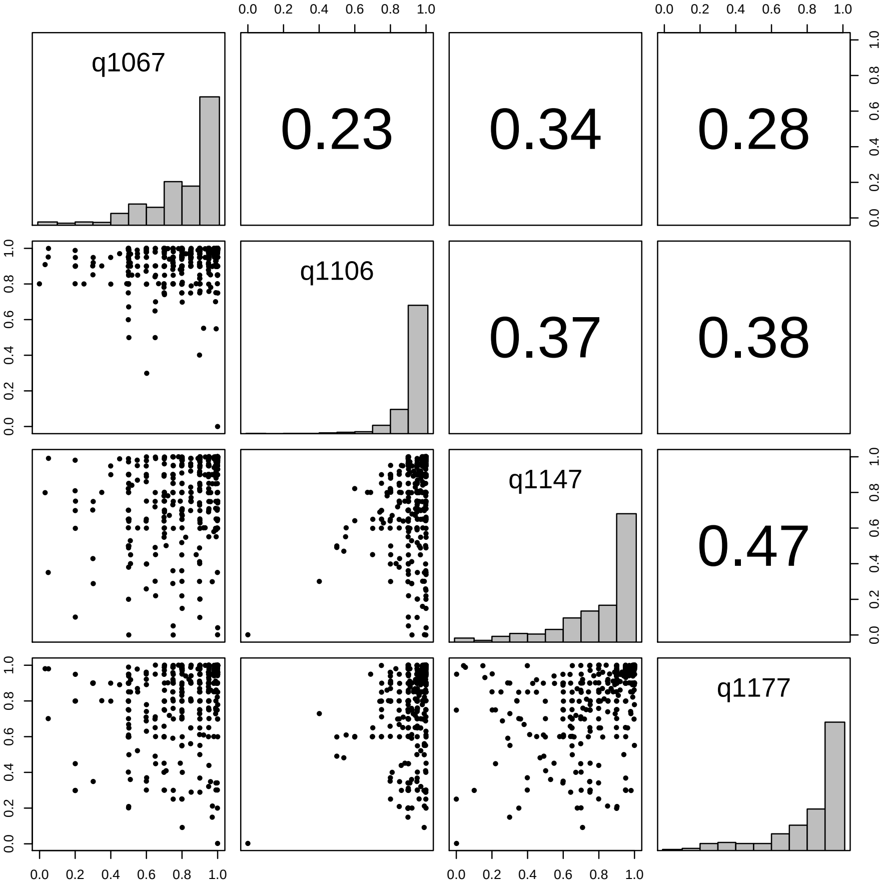
Simple example, response pattern frequencies and mean Brier scores. The four numbers in the "Response pattern" column correspond to questions 1067, 1106, 1147, and 1177, respectively, equaling 0 for question nonresponse and 1 otherwise.

Response pattern	Frequency	Mean Brier
0001	1	0.022
0010	1	0.000
0011	1	0.006
0100	15	0.053
0101	11	0.063
0110	8	0.027
0111	64	0.056
1000	173	0.102
1010	1	0.061
1100	87	0.043
1101	31	0.052
1110	34	0.066
1111	345	0.060

Data Summary. The questions' mean Brier scores and response rates (out of the number of people who responded to any of the four questions) are displayed in Table 1; scatterplots and distributions of reported forecasts are displayed in Figure 2; and response pattern frequencies and mean Brier scores are displayed in Table 2. Questions 1067 and 1147 had the worst Brier scores, and questions 1147 and 1177 were less popular than the other two. Figure 2 (most notably, the panels for question 1067) also shows that there is some overuse of "nice" numbers like .5, which indicates that, e.g., some participants might be reporting .5 to reflect complete uncertainty, as opposed to a probability of event occurrence. Our model does not account for this phenomenon, and it is unclear whether accounting for it is worth the additional model complexity that would be required.

Finally, Table 2 shows that 345 forecasters responded to all four questions, with 173 forecasters responding only to the first question (1067). The people responding only to question 1067 appear to be worse than other forecasters in terms of the Brier score, though this result is clouded by differences in question difficulty and in response pattern frequencies. The estimated model, described below, can help to provide a clearer assessment of these issues.

Figure 2. Simple example, visual summaries of forecasts for each question's realized outcome. The upper triangle displays Pearson correlations associated with the scatterplots in the lower triangle.

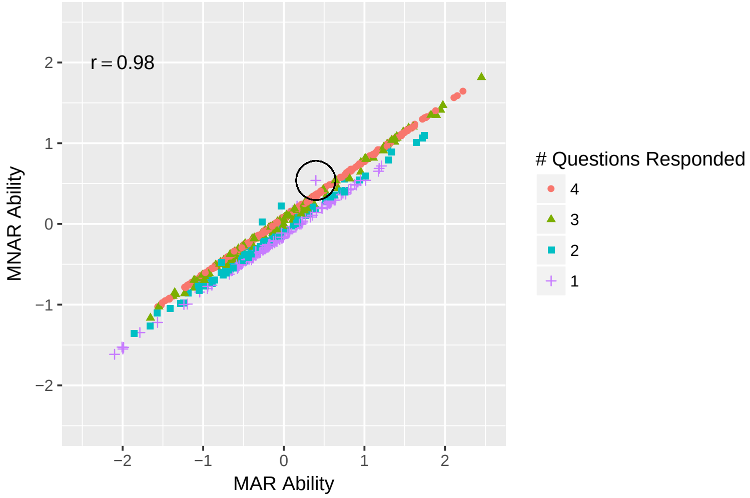


Results. In examining the estimated IRT model of forecasts and question selections, several results are notable. We start with the λ parameters that describe the influence of forecaster ability on reported forecasts and on question selection. We then move to the forecaster ability estimates.

The λ parameters that related to question discrimination ($\lambda_{1,1}$ to $\lambda_{4,1}$) were all close to 1, which (unsurprisingly) means that better forecasters tended to do better on all four questions. Perhaps more surprisingly, better forecasters were more likely to select certain questions (as judged by $\lambda_{5,1}$ to $\lambda_{8,1}$). This was particularly the case for the two questions with lower response rates and worse Brier scores, 1147 and 1177. Question 1106 showed a smaller influence of forecaster ability on question selection, while question 1067 showed virtually no influence.

Figure 3 compares the ability estimates from the Merkle et al. (2016) MAR model (x-axis; note that this model included a linear effect of time similar to Equation (5)) to the new model of forecasts and question selection (y-axis). Each point represents a single

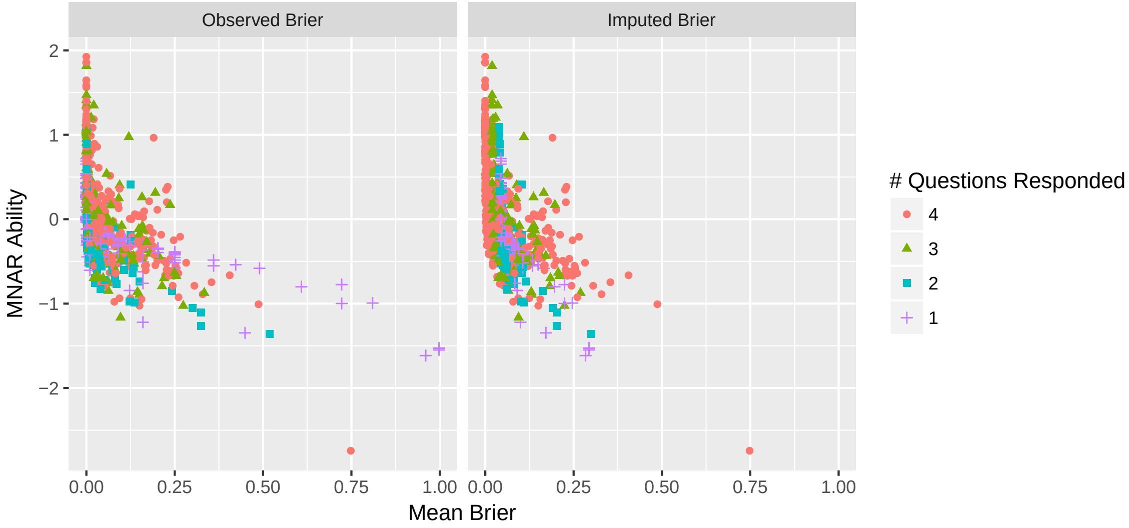
Figure 3. Simple example, comparison of MAR ability estimates to MNAR ability estimates that incorporate question selection. The Spearman correlation appears in the upper left.



forecaster, with the point's color and shape representing the total number of questions answered (out of 4 possible). We can roughly see three diagonal lines going from bottom left to top right: one line of red circles and green triangles, one line of blue squares, and one line of purple plusses. The red circles and green triangles tend to be closest to the top, which means that the forecasters who answered 3 or 4 questions generally received the highest ability estimates under the MNAR model, followed by the forecasters who answered 2 questions, followed by forecasters who only answered 1 question. The MNAR model has automatically penalized non-responders, because the non-responders tended to supply worse forecasts than the frequent responders.

The figure also includes a small number of forecasters who stand out; one such forecaster is circled in the middle of the plot. The circled forecaster answered only one question (a purple plus) yet, under the new model, obtained a higher ability estimate than similar people who responded to all four questions. This is a person who responded only to the question that was most highly associated with forecaster ability (question 1147). This is also a very uncommon response pattern: this is the only person who responded to question 1147 and no others. The person additionally made a near-perfect forecast of .99 in favor of the realized outcome on that question. Thus, the model has rewarded the person for making a good forecast on the question that was most associated with good forecasting. This reward is relative to the person's MAR ability estimate; that is, the person's new ability estimate is still in the middle of the pack, as compared to the full set of forecasters. In order to obtain the highest ability estimate, a forecaster must report exceptional forecasts on most or all of the questions. This is because the shrinkage of each forecaster's ability estimate is related to the amount of data available on a forecaster: as a forecaster responds to more questions, his/her ability estimate can become more extreme.

Figure 4 further compares the new model's ability estimates to two types of Brier scores: a mean observed Brier score, and a mean imputed Brier score. These reflect heuristic methods for handling missing data while still using a proper scoring rule. For each forecaster, *Figure 4*. Mean observed Brier score (x-axis, left panel) and mean imputed Brier score (x-axis, right panel) versus MNAR ability estimates.



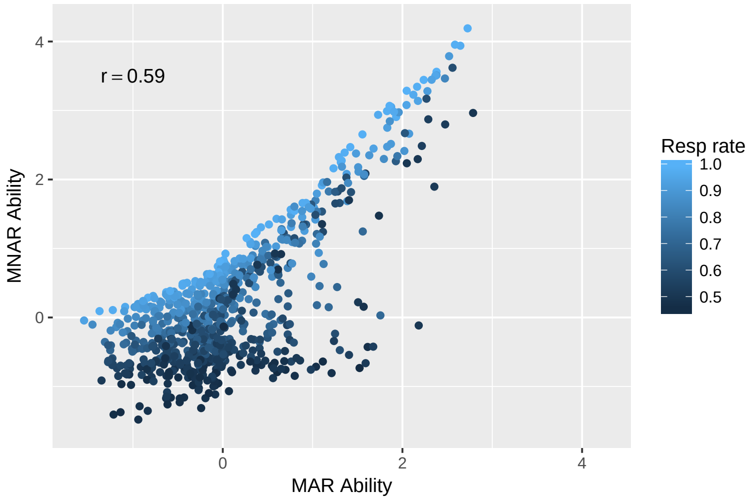
the mean observed Brier score averages Brier scores only across the questions to which the forecaster responded (similar to, e.g., treating missing forecasts as "not reached"). The mean imputed Brier score, on the other hand, fills in the missing observations; these missing observations receive the corresponding question's mean Brier score based on the observed forecasts for that question (similar to, e.g., treating missing forecasts as "incorrect").

In Figure 4, the x-axis reflects the Brier scores and the y-axis reflects the model estimates. For reference, the red circles in both figures are located in exactly the same places; Brier score imputing has no impact on people who forecasted all four questions. The figure shows that the ability estimates from the model are generally related to the Brier scores, with correlations in the $-.6$ to $-.7$ range. Comparing the two panels with one another, we see that the Brier score imputing helped many people with bad Brier scores. In the left panel, these people are generally closer to the right side of the x-axis with points that are triangles, squares, or plusses. In the right panel, these people have all moved further left on the x-axis (improved) while the people who responded to all four questions stayed in the same location. Perhaps the most striking result of this figure involves the fact that we observe multiple vertical "lines" of points. This shows that the model assigns different ability estimates to people who receive nearly the same Brier scores. The specific questions that were selected, along with time that the forecasts were reported, are responsible for these differences.

Full Data

Now that we have illustrated the model's application to a small number of questions, we fit the model to the larger data set of 775 forecasters responding to 157 questions (again maintaining only the first forecast reported by each person on each question). We focus on comparing the MNAR model to the Merkle et al. (2016) model that does not handle question selection. This comparison provides information about the impact of the "missing

Figure 5. Comparison of MAR ability estimates versus MNAR ability estimates obtained from data across all four years of the tournament. The Spearman correlation appears in the top left.



at random" assumption on model estimates.

A comparison of the Merkle et al. (2016) ability estimates (missing at random) and the new ability estimates (missing not at random) is displayed in Figure 5. Points are now displayed in various shades of blue based on response rate; blue points represent forecasters who responded to nearly all the questions, while black points represent forecasters who responded to fewer questions. The figure clearly shows that response rate influences ability estimates in the new model: forecasters who received similar ability estimates under the old model can now receive very different estimates from the new model. The extent to which the new ability estimates change is dependent on response rate: light blue points are always closest to the top of the graph, and darker points are further below. Just like the simple example, the extent to which the darker points are penalized is dependent on the specific questions to which forecasters responded: if a "low response rate" forecaster responded to many questions that good forecasters selected, then that forecaster is not penalized as much. If the "low response rate" forecaster responded in other ways, then his/her penalty is larger.

Figure 6 displays a histogram of λ estimates corresponding to paths from "Forecaster ability" to the question selection variables (see Figure 1). These estimates provide information about whether good forecasters are more/less likely to select certain questions. The histogram indicates that the chance of responding to each question increases with forecaster ability, regardless of that question's difficulty. This result has at least two further implications. First, there is a deviation from the MAR assumption, because the MAR model is obtained when all these λ parameters equal zero. Second, a forecaster can improve his/her ability in two ways: by reporting good forecasts and by responding to many questions. This provides an incentive that is especially useful to forecast consumers: the model developed here can incentivize forecasters to increase response rates. We return to this issue in the General Discussion.

Figure 6. Histogram of λ estimates corresponding to paths from the "forecaster ability" latent variable to "question selection" variables.

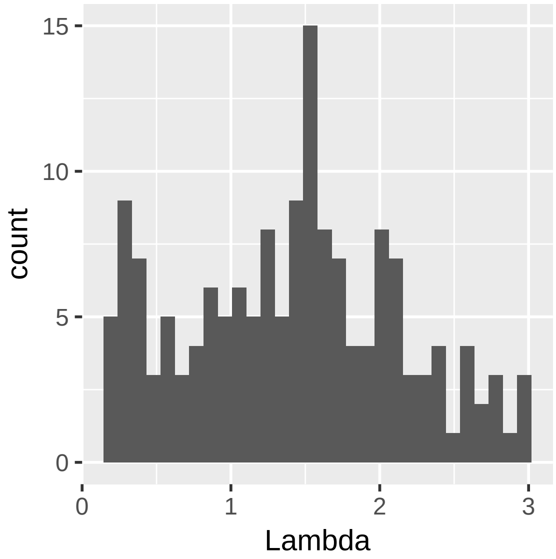


Table 3

Notable questions illuminated by the model estimates.

Questions related to high ability	Question text
1174	Will the Turkish government release imprisoned Kurdish rebel leader Abdullah Ocalan before 1 April 2013?
1177	Will Mohammed Morsi cease to be President of Egypt before 1 April 2013?
1183	Will the United Nations Security Council pass a new resolution directly concerning Iran between 17 December 2012 and 31 March 2013?

Questions unrelated to ability	Question text
1004	Will the United Nations General Assembly recognize a Palestinian state by 30 September 2011?
1010	Will the 30 Sept 2011 "last" PPB for Nov 2011 Brent Crude oil futures* exceed $115?
1022	Will the South African government grant the Dalai Lama a visa before 7 October 2011?

Finally, the histogram in Figure 6 indicates question variability: some estimates are close to zero, indicating that forecaster ability is nearly unrelated to the selection of certain questions, whereas other estimates are far from zero. Table 3 shows some specific questions that fell at each extreme. The bottom section contains questions whose λ estimates were near zero, indicating that their selection was "unrelated to ability." These questions were all open near the start of the tournament, when people were first getting accustomed to forecasting. Some of these people became good forecasters and some dropped out, likely explaining the model results. Conversely, the top section contains questions whose selections were "related to high ability." These questions were open later in the tournament, and they comprise less-popular topics that beginning forecasters may have avoided.

Impact of Dropouts

As shown in an earlier section, the multidimensionality in forecasts and question selections is partially related to the fact that the forecasting tournament was divided into four separate years. At the end of each year, many existing forecasters dropped out and many new forecasters entered for the subsequent year. Thus, the results in the previous section (Figure 5) were influenced by two types of missing data: dropouts who only reported forecasts during a subset of the tournament, and selective responders who forecasted a subset of questions across the entire tournament.

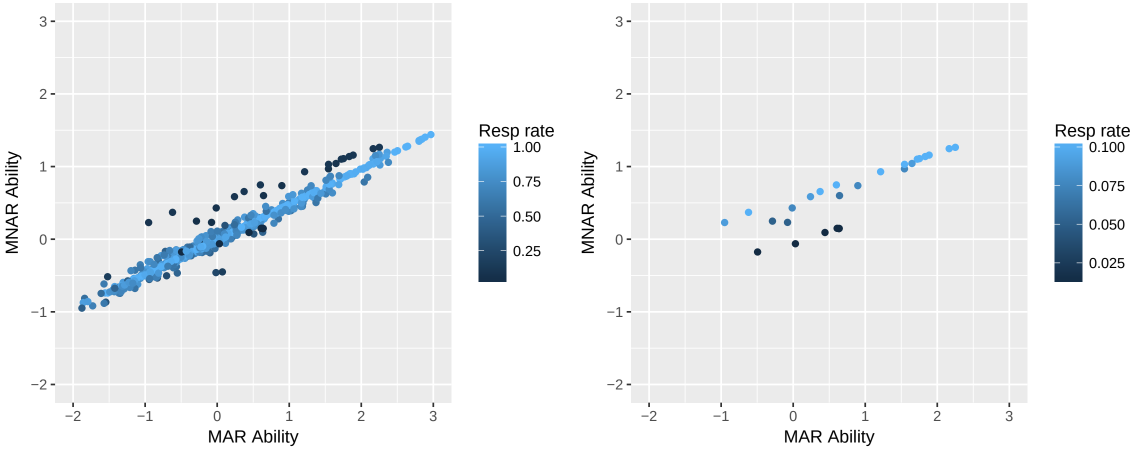
The dropouts may influence the model differently from the selective responders. This is because the best forecasters (the "superforecasters;" see Mellers, Stone, Murray, et al., 2015) tended to continue reporting forecasts during the entire tournament, and worse forecasters were more likely to drop out. The fact that bad forecasters dropped out more often implies that bad forecasters had more missing data, so that the model learned to penalize forecasters with low response rates. If we can avoid the bad forecasters who dropped out after each year, however, then the model may penalize/reward forecasters differently. Thus, in this section, we fit the model to only Year 1 forecasts, which eliminates year-to-year dropout effects in our analysis.

Method. We fit the model to 771 forecasters who made at least one forecast during Year 1. This is a subset of the original data and includes some forecasters with very sparse data (who reported infrequently during Year 1 and more frequently during subsequent years). We restricted ourselves to 78 questions that both opened and closed during Year 1.

Results. The left panel of Figure 7 contains the main results, with the MAR ability estimates on the x-axis and the MNAR ability estimates on the y-axis. The light blue points form a diagonal line, showing that people who responded to nearly all questions receive similar ability estimates across models (except for some rescaling). A small number of darker points cluster around the main diagonal line, showing that some people who responded to fewer questions received small rewards or penalties depending on their response patterns. Aside from this, we see a small number of dark points that are farther from the diagonal line, with many of these points receiving higher ability estimates under the MNAR model.

The dark points above the line represent people who responded to a small number of questions that tended to be selected by good forecasters. The right panel of Figure 7 contains a closer look at the dark points from the left panel. The right panel contains forecasters who responded to 10% (seven) of the questions or fewer, so that the shading reflects response rates that go from 0 to .1 instead of from 0 to 1. It is seen that the

Figure 7. Comparison of MAR ability estimates to MNAR ability estimates during Year 1 (left panel). The right panel contains a subset of points on the left panel from infrequent responders.



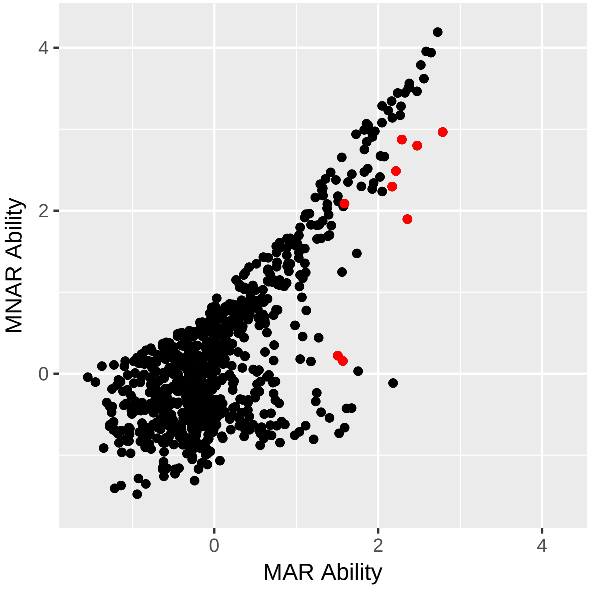
nonresponders who received the largest boost generally answered six or seven questions (close to 10% of the questions). These questions were ones that good forecasters tended to answer, and the nonresponders generally reported good forecasts on these questions. The model deemed this sufficient evidence to give the forecasters a boost.

Do these forecasters deserve the boost? To answer this question, we looked at how the forecasters performed in the full dataset (including data from other years). We focused on the nine nonresponders in the right panel of Figure 7 whose ability estimates from the new model were greater than .8. We then re-created Figure 5 in Figure 8, except that the 9 nonresponders are now highlighted in red. It is seen that, when we compare forecasters on ability across years, the people who originally received a boost now receive a penalty. This is likely because the nonresponders had larger amounts of missing data across years. Despite this finding, the nonresponders who originally received a boost during Year 1 all remain in the top half of forecasters, with seven of nine being above the 90th percentile on ability. This suggests that the new model can help us identify good forecasters who have only responded to a small number of questions. We further explore this suggestion in the next section.

Improvements in Ability Estimates

While the previous sections have illustrated that the new IRT model rewards/penalizes (non)response in an intuitive fashion, we ultimately wish to know whether the resulting ability estimates are better than those of the model that employs the missing at random assumption. This issue is more complex than it initially appears because it requires us to explicitly define what we mean by "ability." For example, imagine that we estimate forecaster ability via three metrics: the mean Brier score, the MAR model, and the MNAR model. It is likely that, if we compute each of these metrics in a training sample, they will be most highly correlated with the analogous metric in a test sample: the training Brier score will be most correlated with the test Brier score, the training MAR estimates

Figure 8. Display of the Year 1 infrequent responders in the full dataset (red points), as compared to other forecasters.

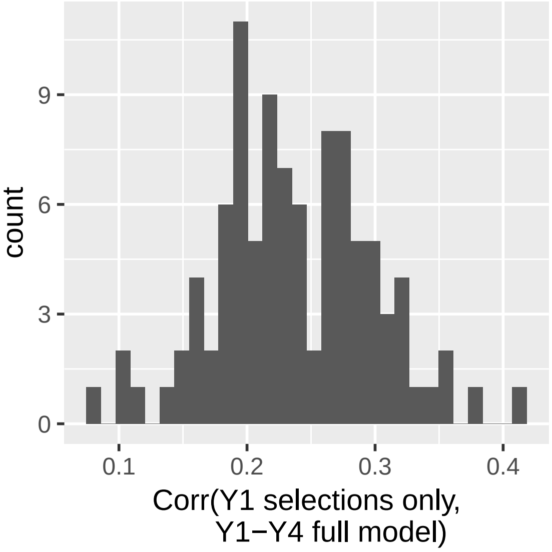


will be most correlated with the test MAR estimates, and the training MNAR estimates will be most correlated with the test MNAR estimates. In order to say which model is best, we need to explicitly decide which metric counts as the "official" measure of ability. This amounts to dealing with the ability metrics' validities (e.g., Borsboom, Mellenbergh, & van Heerden, 2004), which is a difficult topic to address in the current context.

We sidestep validity issues here, showing that, if we provide the MNAR model only with the questions that some forecasters selected (and not with their actual forecasts), then those forecasters' ability estimates are related to those that would be obtained if we used the full data. This implies that there is information to be leveraged from the item selections, separately from the reported forecasts. This, in turn, illustrates the utility of the proposed model in practice.

Method. We conducted a simulation study of the MNAR model, using only data from Year 1 of the forecasting tournament. Similar to the previous section, this was done so that the model could not capitalize on year-to-year dropout effects. For each of 100 replications, we randomly selected 25% of the 775 forecasters in the data and deleted all their forecasts. We maintained the questions that these forecasters selected (i.e., the d_{ij}), however, fitting the model to these selections along with the full data provided by the remaining 75% of forecasters. Following model estimation (2,000 burn-in samples followed by 2,000 posterior draws), we computed the posterior mean ability estimates of the forecasters whose forecasts were deleted. Finally, we examined relationships between these ability estimates and the ability estimates associated with the forecasters' full data from Years 1 to 4. We included the data from Years 2 to 4 in our comparison because it served as a more stringent generalizability measure. That is, because data from Years 2–4 were completely held out of the initial model estimation, it is more impressive if the resulting ability estimates are correlated with the estimates that include data from Years 2–4.

Figure 9. Simulated correlations between ability estimates under two models: a model that only uses question selections from Year 1, and a model that uses both reported forecasts and question selections from Years 1–4.



Results and Discussion. Figure 9 contains a histogram of correlations between (i) the ability estimates resulting from the Year 1 deleted dataset (where 25% of forecasters had only question selection data) and (ii) the ability estimates resulting from the model developed in this paper (utilizing reported forecasts and question selections from all four years). There are 98 correlations depicted in the histogram, as the model failed to converge for two of the one-hundred simulation replications. This is likely due to bad, randomly-generated initial values in these replications.

The histogram shows that the ability estimates from the two models are positively correlated across all replications, with a mean correlation of 0.24 and an interquartile range of (0.2,0.28). This result provides evidence that the question selections contain information that is related to the full ability estimates (that would be obtained if we included reported forecasts in the model).

The result is weakened by the fact that the question selection data were included in both models; we might expect a positive correlation between the models' estimates because they were partially based on the same data. To explore this criticism, we also examined the relationship between the "Year 1, question selection" ability estimates and the MAR ability estimates (based on the model from Merkle et al. (2016)). The latter model utilizes only the reported forecasts from Years 1–4 (without question selection data), so that the forecasters with deleted data contribute unique data points to each model. These correlations are nearly always positive (in 97 of 98 replications), with a mean of 0.12 and an interquartile range of (0.08,0.16). This mean (and range) is lower than that of the correlations from Figure 9, potentially illustrating the impact of repeating the data across models. However, based on the fact that correlations remain positive, we conclude that there exists useful information in the question selection data. This information may not always lead to major, practical improvements in ability estimates, but it is worthwhile to consider in scenarios where forecasters are free to select their own questions.

Discussion

In this paper, we first developed a psychometric model that allows us to assess forecasters' abilities while simultaneously handling data on question selection. This is potentially useful in situations where forecasters are free to select the questions that they wish to forecast, so that the selected questions provide information about forecasting ability above and beyond the forecasts reported on the questions. After model development and assumption checking, we illustrated the extent to which the proposed model differed from a previous model that did not account for question selection. Results from the new model implied that good forecasters tended to select more questions, regardless of question difficulty, and that specific question selections had an influence on forecaster ability estimates. We also studied the extent to which we can estimate forecaster ability based on question selections alone (and not forecasts), finding that these ability estimates exhibited correlations of .24 (on average) with the full data ability estimates. This implies that there is information in the question selections that can be capitalized upon, a result that has also been studied in other contexts (e.g., Rubin & Steyvers, 2009). In the Discussion, we provide further ideas on missingness mechanisms, relationships to traditional scoring rules, model assumptions, and methods of model estimation.

Missingness Mechanisms

One appeal of the proposed MNAR model involves the fact that it handles missingness in a manner that agrees with intuition: good forecasters select questions differently from bad forecasters (in the specific context of the current data, good forecasters selected more questions than bad forecasters), and we should be able to leverage these differences in forecaster assessment. On the other hand, the statistical literature on missing data (e.g., Little & Rubin, 2002) clearly states that (i) there are an infinite number of missingness mechanisms that qualify as "missing not at random," and (ii) if the mechanism in the model does not match the truth, then parameter estimates may exhibit more bias than the corresponding "missing at random" estimates. The implication is that the extra complexity of the proposed model may hurt us.

At least for the model proposed in this paper, there appears to be little danger in employing the MNAR model instead of the MAR model. This is because the MAR model is a special case of the MNAR model, being obtained by fixing a subset of the λ parameters to zero. Thus, if the MAR assumption is approximately fulfilled, the model should automatically account for this during estimation.

Relationship to Scoring Rules

The model described here may also be used to develop new types of model-based scoring rules (see Budescu & Bo, 2015, for related ideas). Existing scoring rules (such as the Brier score or logarithmic score; see, e.g., Gneiting & Raftery, 2007) work only on the forecasts themselves, requiring that every forecaster responds to exactly the same questions. This seldom holds true in practice, and it is awkward to tailor these scoring rules to missing data. For example, for each question that a forecaster fails to answer, we might substitute the mean observed Brier score on the corresponding question. This substitution is related to IRT procedures that code unanswered questions as incorrect (though, in a forecasting context, the notion of "incorrect" is unclear).

Beyond substitution of missing observations, we can consider new scoring rules in which the question selections and missing data play a role. A rough definition, corresponding to the models estimated in this paper, is as follows. A forecaster receives the highest expected score if:

- He/she consistently makes better forecasts than the crowd, and
- He/she responds to more questions.

This definition requires the best forecasters to be the best on both question selection and forecast reporting. Further, middling forecasters could receive the same abilities through different routes. For example, say that Forecaster A and Forecaster B receive the same ability estimate from the model. Forecaster A may obtain this estimate through selecting many questions but providing relatively-bad forecasts on those questions, while Forecaster B may obtain this estimate through selecting few questions but providing relatively-good forecasts on those questions. Further work could examine the extent to which these two criteria simultaneously incentivize honest forecasting and frequent responding. A game-theoretic framework similar to that of Prelec (2004) may be useful here, because we can depict each forecaster as striving to do the minimal amount of forecasting required to be the best. Under these conditions, forecasters might be motivated to respond to all questions when they do not know other forecasters' response patterns.

Model Assumptions

As mentioned throughout, the model proposed here assumes a single dimension of forecaster ability; that each forecaster's ability can be summarized via a single number. While the analyses in this paper suggest that this assumption is not grossly violated in our dataset, there remains the possibility that it is grossly violated in other datasets. At an extreme, we could imagine a forecaster who only follows local occurrences and knows nothing about broader world events. If this forecaster only responds to questions related to her locale, then she may receive a good ability estimate despite the fact that her forecasts would be awful on other, unanswered questions.

Despite this violation, the model's handling of this extreme forecaster could still be reasonable. First, if other forecasters of high ability tend to respond to questions that do not involve this particular locale, then the model will temper the extreme forecaster's ability estimate so that it is not as high as others. Second, if the extreme forecaster does not respond to many questions (i.e., there are few questions about the forecaster's locale), then the model will again temper her ability estimate: the model requires large amounts of data from the forecaster before it is "willing" to assign an extremely-good ability estimate. While these results do not guarantee that the model is robust to all dimensionality violations, they seem applicable to many situations where evaluators wish to rank order forecasters across all questions.

Related to the dimensionality issue, the model also assumes that each forecaster has a static level of forecasting ability and response propensity. In contrast, forecasters tend to change over time, gaining (losing) interest in the forecasting tournament and reporting improved (diminished) judgments. As proposed here, the model cannot accommodate forecaster attributes that change over time, though it may be possible to directly model changes in forecaster ability over time via new parameters and/or increased dimensions of forecaster ability. It would also be of interest to relax distributional assumptions, employing, say, t distributions instead of normal distributions or mixture models that accommodate subclasses of homogeneous forecasters. As further described in the next section, traditional psychometric modeling frameworks can be helpful for including these model extensions.

Model Estimation

The estimation of traditional item response models with multiple ability dimensions is generally difficult (e.g., Cai, 2010), and this result holds true for the two-dimensional model proposed here. The Bayesian approach that we adopted introduces an additional complication in that we must employ Markov chain Monte Carlo, sampling the forecaster ability parameters instead of integrating them out (e.g., Lee, 2007). This means that we must be careful to ensure that the model parameters are identified and that the model converges (e.g., Ghosh & Dunson, 2009; Merkle & Wang, in press; Peeters, 2012), which may introduce an undesirable practical complication.

Depending on the data, simplifications are available. In particular, we adopted the Bayesian approach in this paper so that we could easily include the "time of reported forecast" covariate in the model. This covariate is not necessary, however, when all forecasters report their judgments at approximately the same time. If this covariate is not necessary, then the model proposed here can often be estimated via Maximum Likelihood, using popular SEM software such as Mplus (L. K. Muthén & Muthén, 1998–2012) or lavaan (Rosseel, 2012). These approaches would make use of ideas related to the path diagrams from Figure 1. When the data are very sparse (i.e., each forecaster reports on a small proportion of questions), however, these programs may fail in situations where the Bayesian approach can succeed. This failure is again related to the fact that ML estimation methods integrate the forecaster latent variables out of the likelihood, whereas Bayesian estimation methods directly sample the forecaster latent variables (and are "smoothed" by the prior distributions). Integration of the latent variables requires us to work with the covariance matrix of a multivariate normal likelihood, which can often become non-positive definite during model estimation (resulting in failed estimation).

Sample size is an additional consideration for all the models discussed here. Because the proposed model is related to traditional psychometric models (including factor analysis and item response models), we can draw on the psychometric literature for sample size recommendations. In that literature, it is customary to observe hundreds or thousands of participants reporting on a small number of items. Researchers proposing models similar to ours have followed this trend: Holman and Glas (2005) applied their model to 171 participants responding to 32 items, whereas O'Muircheartaigh and Moustaki (1999) applied their model to two datasets, one of which had 2,691 participants responding to five items and one of which 1,270 participants responding to four items. While our application had many more items than the others, we generally recommend large numbers of participants and suggest artificial data simulation as a way to determine whether one's particular sample size is appropriate for estimating parameters of interest. The Bayesian approach of directly sampling forecaster latent variables can again be helpful here, allowing us to bypass non-positive definite covariance matrices.

Summary

In situations where respondents are free to select their own questions or stimuli, the specific selections can provide valuable information about the latent respondent attributes that we wish to measure. While these selections are often viewed as nuisance characteristics of the data that cause difficulties for analysis, we have illustrated here a model-based approach to capture the information inherent in the selections. The ability to incorporate multiple types of variables (forecasts, question selections) in forecaster assessment is a major advantage of model-based approaches over data-based metrics (i.e., scoring rules), which rely exclusively on the reported forecasts. In forecasting scenarios and beyond, detailed consideration of selection/missingness mechanisms can lead to improved estimation of latent traits of interest.

Computational Details

All results were obtained using the R system for statistical computing (R Core Team, 2016) version 3.3.2 and JAGS software for Bayesian computation (Plummer, 2003) version 4.2.0, employing the add-on package runjags 2.0.4-2 (Denwood, in press). R and the package runjags are freely available under the General Public License 2 from the Comprehensive R Archive Network at `http://CRAN.R-project.org/`. JAGS is freely available under the General Public License 2 from Sourceforge at `http://mcmc-jags.sourceforge.net/`.

References

Bejar, I. (1977). An application of the continuous response level model to personality measurement. *Applied Psychological Measurement*, *1*, 509–521.

Bonifay, W. E., Reise, S. P., Scheines, R., & Meijer, R. R. (2015). When are multidimensional data unidimensional enough for structural equation modeling? An evaluation of the DETECT multidimensionality index. *Structural Equation Modeling*, *22*, 504–516.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*, 1061–1071.

Budescu, D. V., & Bo, Y. (2015). Analyzing test-taking behavior: Decision theory meets psychometric theory. *Psychometrika*, *80*, 1105–1122.

Cai, L. (2010). High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm. *Psychometrika*, *75*, 33–57.

Carvalho, A. (in press). An overview of applications of proper scoring rules. *Decision Analysis.*

Chang, Y.-W., Tsai, R.-C., & Hsu, N.-J. (2014). A speeded item response model: Leave the harder till later. *Psychometrika*, *79*, 255–274.

Denwood, M. J. (in press). runjags: An R package providing interface utilities, model templates, parallel computing methods and additional distributions for MCMC models in JAGS. *Journal of Statistical Software*. Retrieved from `http://runjags.sourceforge.net`

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum Associates.

Ferrando, P. J. (2001). A nonlinear congeneric model for continuous item responses. *British Journal of Mathematical and Statistical Psychology*, *54*, 293–313.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, *7*, 457–511.

Ghosh, J., & Dunson, D. B. (2009). Default prior distributions and efficient posterior computation in Bayesian factor analysis. *Journal of Computational and Graphical Statistics*, *18*, 306–320.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*, 359–378.

Holman, R., & Glas, C. A. W. (2005). Modelling non-ignorable missing-data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Psychology*, *58*, 1–17.

Lee, S.-Y. (2007). *Structural equation modeling: A Bayesian approach.* Chichester: Wiley.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: Wiley.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

McDonald, R. P. (1999). *Test theory: A unified treatment.* Mahwah, NJ: Erlbaum.

Mellers, B., Stone, E., Atanasov, P., Rohrbaugh, N., Metz, S. E., Ungar, L., … Tetlock, P. (2015). The psychology of intelligence analysis: Drivers of prediction accuracy in world politics. *Journal of Experimental Psychology: Applied*, *21*, 1–14.

Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., … Tetlock, P. (2015). Identifying and cultivating superforecasters as a method of improving probabilistic predictions. *Perspectives on Psychological Science*, *10*, 267–281.

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., … Tetlock, P. E. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*.

Merkle, E. C., Steyvers, M., Mellers, B., & Tetlock, P. E. (2016). Item response models of probability judgments: Application to a geopolitical forecasting tournament. *Decision*, *3*, 1–19.

Merkle, E. C., & Wang, T. (in press). Bayesian latent variable models for the analysis of experimental psychology data. *Psychonomic Bulletin and Review*.

Müller, H. (1987). A Rasch model for continuous ratings. *Psychometrika*, *52*, 165–181.

Muthén, B. (1989). Tobit factor analysis. *British Journal of Mathematical and Statistical Psychology*, *42*, 241–250.

Muthén, L. K., & Muthén, B. (1998–2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Noel, Y., & Dauvier, B. (2007). A beta item response model for continuous bounded responses. *Applied Psychological Measurement*, *31*, 47–73.

O'Muircheartaigh, C., & Moustaki, I. (1999). Symmetric pattern models: A latent variable approach to item non-response in attitude scales. *Journal of the Royal Statistical Society A*, *162*, 177–194.

Peeters, C. F. W. (2012). Rotational uniqueness conditions under oblique factor correlation metric. *Psychometrika*, *77*, 288–292.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, & A. Zeileis (Eds.), *Proceedings of the 3rd international workshop on distributed statistical computing.*

Prelec, D. (2004). A Bayesian truth serum for subjective data. *Science*, *306*, 462–466.

R Core Team. (2016). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2013). Multidimensionality and structural coefficient bias in structural equation modeling: A bifactor perspective. *Educational and Psychological Measurement*, *73*, 5–26.

Robitzsch, A. (2016). sirt: Supplementary item response theory models [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=sirt` (R package version 1.12-2)

Rose, N., von Davier, M., & Xu, X. (2010). *Modeling nonignorable missing data with item response theory (IRT)* (Tech. Rep.). ETS Research Report.

Rosseel, Y. (2012). *lavaan*: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. Retrieved from `http://www.jstatsoft.org/v48/i02/`

Roussos, L. A., & Ozbeck, O. Y. (2006). Formulation of the DETECT population parameter and evaluation of DETECT estimator bias. *Journal of Educational Measurement*, *43*, 215–243.

Rubin, T. N., & Steyvers, M. (2009). A topic model for movie choices and ratings. In *Proceedings of the 9th International Conference on Cognitive Modeling – ICCM2009.* Manchester, UK.

Samejima, F. (1973). Homogeneous case of the continuous response model. *Psychometrika*, *38*, 203–219.

Stout, W., Habing, B., Douglas, J., Kim, H. R., Roussos, L. A., & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement*, *20*, 331–354.

Thissen, D. (2016). Bad questions: An essay involving item response theory. *Journal of Educational and Behavioral Statistics*, *41*, 81–89.

van Abswoude, A. A. H., van der Ark, L. A., & Sijtsma, K. (2004). A comparative study of test dimensionality assessment procedures under nonparametric IRT models. *Applied Psychological Measurement*, *28*, 3–24.

Wang, W.-C., Jin, K.-Y., Qiu, X.-L., & Wang, L. (2012). Item response models for examinee-selected items. *Journal of Educational Measurement*, *49*, 419–445.

Zhang, J. (2007). Conditional covariance theory and DETECT for polytomous items. *Psychometrika*, *72*, 69–91.

Zhang, J., & Stout, W. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika*, *64*, 213–249.

Appendix A
DETECT technical details

We computed the DETECT statistic separately for the reported forecasts y and for the question selections d. The statistics were calculated via the `expl.detect()` function in R package *sirt* (Robitzsch, 2016).

Calculation of the DETECT statistic for question selections was straightforward. This is because all forecasters had complete data corresponding to standard item response data. That is, each forecaster's data were composed of a series of 0s and 1s, with 0 indicating that he/she did not respond to a particular question and 1 indicating the opposite. In addition to the observed data, the DETECT statistic also requires unidimensional estimates of person ability. For this, we used the weighted likelihood estimates arising from a Rasch model.

To compute a DETECT statistic for the reported forecasts, we first restricted ourselves to a subset of 241 forecasters who responded to at least 136 of 176 questions (with most forecasters responding to at least 160 of the questions). We did this so that we could ignore missing data mechanisms while examining forecast dimensionality. Next, in an attempt to maximize the DETECT statistic, we transformed the data to account for the fact that forecasts were reported at different points in time. In particular, for each question j, we regressed the y^*s associated with question j on the time at which the forecast was reported (i.e., the t_{ij}). We then used the fitted model to push each person's reported forecast to the the question's "halfway" point (i.e., the time where the question is halfway between introduction and resolution). Finally, to compute the statistic, we created binary variables from the aligned forecasts (equal to 0 if the forecast was less than .5, 1 otherwise). For estimates of person ability, we used the average forecast reported for each question's realized outcome.

Appendix B
JAGS model estimation

JAGS code to estimate the model is displayed below. The probit-transformed forecasts ystar are in long format, while the missingness indicators d are in a data matrix where rows are forecasters and columns are questions. Following the JAGS code, we provide R code to illustrate usage.

```
model{
  for (i in 1:nr){  ## Rows of forecast data
    ystar[i] ~ dnorm(mu[i], invsig2[qidx[i]])

    mu[i] <- b0[qidx[i]] + b1[qidx[i]]*nd[i] + lambda[qidx[i], 1] * theta[pidx[i], 1]
  }

  for (i in 1:n){ ## Forecasters
    for (j in 1:J){ ## Questions
      d[i, j] ~ dbern(pd[i, j])

      probit(pd[i, j]) <- b0[(J + j)] + lambda[(J + j), 1] * theta[i, 1] +
                          lambda[(J + j), 2] * theta[i, 2]
    }

    ## Person parameters
    theta[i, 1] ~ dnorm(0, invpsi[1])
    theta[i, 2] ~ dnorm(0, invpsi[2])
  }
  invpsi[1] ~ dgamma(.01, .01)
  invpsi[2] ~ dgamma(.01, .01)

  ## Equality constraints + priors for question parameters
  lambda[1,1] <- 1
  lambda[1,2] <- 0
  lambda[(J+1), 1] ~ dnorm(0, 1)
  lambda[(J+1), 2] <- 1

  b0[1] ~ dnorm(0, .5)
  b0[(J + 1)] ~ dnorm(0, .5)
  b1[1] ~ dnorm(0, .5)
  invsig2[1] ~ dgamma(.01, .01)

  for (j in 2:J){
    ## loadings for forecasts
    lambda[j, 1] ~ dnorm(0, 1)
    lambda[j, 2] <- 0

    ## loadings for d parameters
    lambda[(J + j), 1] ~ dnorm(0, 1)
    lambda[(J + j), 2] ~ dnorm(0, 1)

    ## Intercept priors
    b0[j] ~ dnorm(0, .5)
    b0[(J + j)] ~ dnorm(0, .5)
    b1[j] ~ dnorm(0, .5)

    ## Error precision prior
    invsig2[j] ~ dgamma(.01, .01)
  }
}
```

The R code below gives an example with artificial data, showing how the JAGS code can be run from within R using the runjags package.

```
library("runjags")
set.seed(1080)

## Generate data
n <- 500
K <- 100

## Probability judgments
b0 <- runif(K, -1, 2)
b1 <- runif(K, 0, 3)

lambda <- runif(K, -.5, 3.5)
theta1 <- rnorm(n, 0, 1)
nd <- runif(n*K, -.5, 0)

dat <- expand.grid(uidx=1:n, ifpidx=1:K)
dat$nd <- nd
mny <- b0[dat$ifpidx] + b1[dat$ifpidx]*nd + lambda[dat$ifpidx]*theta1[dat$uidx]
dat$ystar <- rnorm(n*K, mny, .4)
dat$ystar[dat$ystar < -3.5] <- -3.5
dat$ystar[dat$ystar > 3.5] <- 3.5
dat$ystar[dat$ystar > -.1 & dat$ystar < .1] <- 0

dat$fcast1 <- pnorm(dat$ystar)

## Missingness indicators
b0 <- runif(K, 0.5, 2)
lambda <- matrix(runif(K*2, -.5, 2.5), K, 2)
theta2 <- rnorm(n, 0, 1)

ppd <- lambda[,1] %*% matrix(theta1, 1, n) + lambda[,2] %*% matrix(theta2, 1, n)
ppd <- apply(ppd, 2, function(x) x + b0)
d <- apply(ppd, 2, function(x) rbinom(length(x), 1, pnorm(x)))
for(i in 1:K){
  subs <- which(d[i,] == 0)
  dat$ystar[dat$ifpidx == i & dat$uidx %in% subs] <- NA
}

rmrows <- which(is.na(dat$ystar))
dat <- dat[-rmrows,]

## Data formatted for JAGS
data <- list(nr = nrow(dat), n = length(unique(dat$uidx)),
             J = length(unique(dat$ifpidx)), ystar = dat$ystar,
             qidx = dat$ifpidx, pidx = dat$uidx, nd = dat$nd,
             d = t(d))

## Starting values
inits <- list(b0 = rep(0, 2*data$J), theta = matrix(0, data$n, 2),
              b1 = rep(.1, data$J), invsig2 = rep(1, data$J),
              invpsi = rep(1, 2))

## MCMC run, will take some time
runjags.options(force.summary = TRUE)
mdraws <- run.jags("paper_model.jag", data=data, inits=inits, monitor=c("theta","b0","b1","lambda"),
                   n.chains=3, burnin=5000, sample=1000)

## Parameter summaries, posterior means
mdraws$summaries
mdraws$summaries[,"Mean"]
```