Item Response Models of Probability Judgments: Application to a Geopolitical Forecasting Tournament

Edgar C. Merkle

University of Missouri

Mark Steyvers

University of California, Irvine

Barbara Mellers and Philip E. Tetlock

University of Pennsylvania

Author Note

Correspondence to Edgar C. Merkle, Department of Psychological Sciences, University of Missouri, Columbia, MO 65211. Email: merklee@missouri.edu. 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.

Abstract

In this paper, we develop and study methods for evaluating forecasters and forecasting questions in dynamic environments. These methods, based on item response models, are useful in situations where items vary in difficulty, and we wish to evaluate forecasters based on the difficulty of the items that they forecasted correctly. Additionally, the methods are useful in situations where we need to compare forecasters who make predictions at different points in time or for different items. We first extend traditional models to handle subjective probabilities, and we then apply a specific model to geo-political forecasts. We evaluate the model's ability to accommodate the data, compare the model's estimates of forecaster ability to estimates of forecaster ability based on scoring rules, and externally validate the model's item estimates. We also highlight some shortcomings of the traditional models and discuss some further extensions. The analyses illustrate the models' potential for widespread use in forecasting and subjective probability evaluation.

Item Response Models of Probability Judgments: Application to a Geopolitical Forecasting Tournament

The assessment of forecaster ability is often of interest in applied contexts, especially when there exist conflicting opinions and we wish to know who to believe. This assessment is relevant to in applications such as weather forecasting, sports betting, and political predictions, where the forecasts and the forecasters often exhibit large variability.

In the situation where all forecasters predict the same events at the same point in time, comparisons are relatively straightforward: one waits for the events to resolve themselves, then computes an accuracy metric (often a proper scoring rule) for each forecaster. Forecasters can then be ranked according to their average metric across all problems, providing information about the forecasters' relative ability. These metrics have been traditionally applied to situations where (i) the items are equally "good" at measuring forecasters' abilities, (ii) the items are equally difficult for all forecasters, and (iii) forecasters have responded to the same set of items. Condition (i) is likely violated in domains involving geopolitical events, where novel forecasting questions must be manually developed and old questions cannot be re-used. There is no guarantee that the newly-developed questions are as good (or as bad) as the previously-developed questions. Condition (ii) is violated when forecasters make judgments at different points in time: political forecasters may provide predictions of international events at different times, sports pundits may predict this season's basketball champion at different points in time, and so on. When this happens, forecaster ability may be artificially inflated by the easiness of the forecasted events. For example, a weather forecaster predicting rain for next week will naturally look worse than a weather forecaster predicting rain for tomorrow. Condition (iii) is violated in most real applications, where busy forecasters fail to respond to all inquiries. It is difficult to compare two weather forecasters, one of whom forecasted rain in Seattle and the other of whom forecasted rain in Phoenix.

When the above three conditions are violated, scoring rules and other traditional

forecaster assessments fail because they cannot handle items that differ from forecaster to forecaster. Differing items occur when forecasters have predicted few common items, when forecasts are reported at different points in time, or when some items are better than others (for assessing forecaster ability). To address these issues, we adopt a model-based approach using ideas from item response theory (IRT; Lord & Novick, 1968; Embretson & Reise, 2000; de Ayala, 2009). IRT methods were originally designed to compare school students, as opposed to forecasters, on intellectual ability. In this context, overlapping sets of items are administered to a large number of students. On each item, each student's response is coded as either 0 or 1, where 0 means "incorrect" and 1 means "correct" (though there exist extensions to other types of items). Models arising from IRT then generally decompose the probability that individual i is correct on item j, p_{ij}, into effects of item difficulty and effects of individual ability. Students' ability estimates can be influenced by the specific items that were correctly answered; correct responses can be weighted by the extent to which the correctly-answered item is diagnostic of ability. Further, IRT is developed to the point that, even if some individuals answer no common items, it is still possible to scale them on ability (e.g., Fischer, 1981). This latter feature is especially applicable to forecasting scenarios, where the forecasted events often vary widely.

While IRT models appear useful for comparing forecasters on ability, the traditional models are not immediately applicable to forecasting data. This is because traditional IRT models are fit to binary or ordinal data that reflect whether or not a particular individual correctly answered a particular item. This differs from forecasting contexts, where forecasters assign a probability in $[0, 1]$ reflecting their certainty that a particular event occurs (assuming only two possible outcomes, event occurrence and event non-occurrence). Once the outcome is known, forecaster accuracy is usually represented by a continuous number such as the Brier or logarithmic score (e.g., Merkle & Steyvers, 2013; O'Hagan et al., 2006). Alternatively, components of the Brier score involving calibration or discrimination are sometimes used (Lichtenstein, Fischhoff, & Phillips, 1982; Yates, 1982).

Budescu and Johnson (2011) present a model for assessing calibration that is similar in spirit to the models considered here.

In the pages below, we first formally review some standard IRT models and describe extensions to probability judgments. Next, we apply these models to geo-political forecasts, highlighting differences between IRT ability estimates and simple estimates based on average Brier scores. We also examine out-of-sample model predictions and the model's ability to accommodate major data patterns. Finally, we discuss further model issues and extensions.

Traditional IRT Models and Extensions

As mentioned above, traditional IRT models predict the probability that individual i is correct on item j, denoted here as p_{ij}. One of the simplest models, the Rasch model, makes this prediction via the equation

$$\text{logit}(p_{ij}) = \theta_i - \delta_j, \tag{1}$$

where θ_i is the ability of person i and δ_j is the difficulty of item j. The logit operator, $\text{logit}(p_{ij}) = \log(p_{ij}/(1-p_{ij}))$, ensures that the model's accuracy predictions (i.e., predictions of the p_{ij}) are bounded between 0 and 1. According to the model, the chance that a person is correct depends on her ability relative to item difficulty. If a person's ability equals item difficulty, her probability of being correct is .5. Conversely, as a person's ability increases (assuming fixed item difficulty), her probability of being correct goes to 1.

Many Rasch model extensions have been proposed (e.g., De Boeck et al., 2011), and some of the extensions are useful for modeling forecasts. When forecasters make predictions at different points in time, we must account for the facts that item difficulty changes over the life of an item and that forecasters may update their forecasts multiple times for a single item. To handle the "changing item difficulty" feature of the data, a *dynamic* Rasch model has been developed (Verhelst & Glas, 1993). This model was

originally designed to account for feedback effects on multiple-choice tests, whereby accuracy feedback on previous items impacts one's accuracy on the current item. In this context, the Rasch model was extended to include a "number of previous items correct" covariate. The previous model becomes

$$\text{logit}(p_{ij}) = \theta_i - \delta_j + \beta t_{ij}, \tag{2}$$

where t_{ij} is, say, the number of items that subject i has correctly answered when she reaches item j. In forecasting contexts, t_{ij} could instead represent the time at which a forecast is reported, relative to the time that the item expires. The β parameter describes the extent to which the subject's within-test learning impacts item difficulty. This change is nonlinear on the probability scale, because β is modeled on the logit scale. The β parameter could possibly be given an i subscript to reflect individual differences in feedback effects

Second, to handle the "forecast update" feature of the data, we need to make some assumptions about how updated forecasts are influenced by previous forecasts. In the simplest case, we can assume that item difficulty changes over time and that this fully accounts for the forecaster's updated response. In this paper, we generally employ this simplifying assumption. In a more complex case, we must jointly model the number of updates and the forecast reported at each update. Some issues underlying the estimation of these models were considered by Böckenholt (1993). In particular, we may assume that the frequency of forecast updates is impacted by the forecaster and/or by the item, and that a forecaster's updated forecast is dependent on her previous forecast. The models that we describe below could potentially be extended in these manners.

Now that we have discussed some traditional IRT models and extensions, we discuss novel model extensions to handle probability judgments.

IRT Models of Probability Judgments

A key feature of probability judgments involves the fact that they are bounded from below by 0 and from above by 1. There exists some previous work on IRT models for these types of doubly bounded data, initiated by Samejima (1973) and applied/refined by a small number of others (Bejar, 1977; Ferrando, 2001; Müller, 1987; Muthén, 1989; Noel & Dauvier, 2007). The recent work of Noel and Dauvier (2007) is notable because it employs a beta-distributed likelihood, whereas the others employ logit- or probit-normal likelihoods. We take an approach that has similarities to Samejima, in that we first transform the probability judgments to be unbounded, then fit a model with normally-distributed error (i.e., a one-factor model) to the transformed data. Once this model has been estimated, the parameter estimates can be transformed to IRT parameter estimates using theory that connects factor analysis models to IRT models (Takane & de Leeuw, 1987). This allows us to extract parameter estimates that have simple interpretations.

Model Conceptualization

We generally fit models to respondent i's probit-transformed forecast for item j, denoted Y^*_{ij}. This is essentially equivalent to the logit transformation described above (e.g., McDonald, 1999). We assume there are exactly two possible outcomes for each item, and that Y^*_{ij} always holds respondent i's forecast associated with the realized outcome (so that larger forecasts are always better).

The models that we estimate are variants of a traditional, one-factor model. As further described later, these models are highly related to the IRT models described in the previous section. Assuming J items and I forecasts, a preliminary model may be written as

$$y^*_{ij} = \mu_j + \lambda_j \theta_i + e_{ij} \quad i = 1, \ldots, I; \; j = 1, \ldots, J, \tag{3}$$

where y^*_{ij} is individual i's observed, probit-transformed forecast on item j, μ_j is the mean

forecast for item j, θ_i is respondent i's ability, λ_j reflects the extent to which ability is useful on item j, and e_{ij} is the residual associated with individual i's forecast on item j. Because larger forecasts are always better, the μ parameters are related to item easiness: larger values of μ reflect easier items, and smaller values of μ reflect harder items. As is standard for factor models, we further assume that

$$\theta_i \sim N(0, 1) \tag{4}$$

$$e_{ij} \sim N(0, \psi_j^{-1}), \tag{5}$$

where the θ_i and e_{ij} are assumed independent. The fixed mean and variance on the hyperdistribution of the θ_i act as identification constraints; these parameters are undefined because the θ_i parameters are unobserved, so that the fixed mean and variance "set the scale" for the θ_i. As further described in the "Model Estimation and Use" section, the variance of 1 is enforced on each iteration of the MCMC algorithm through a parameter expansion technique.

A variant of the above model allows time to have an impact on difficulty:

$$y_{ij}^* = b_{0j} + (b_{1j} - b_{0j}) \exp(-b_2 t_{ij}) + \lambda_j \theta_i + e_{ij}, \tag{6}$$

where t_{ij} is the time at which individual i forecasted item j (measured as days until the item expires), b_{0j} reflects item j's easiness as days to item expiration goes to infinity, b_{1j} reflects item j's easiness at the time the item resolves (i.e., the item's "irreducible uncertainty"), and b_2 describes the change in item easiness over time. This exponential function is advantageous in that its parameter interpretations are meaningful to practitioners, but many others could be conceptualized. A linear function of time is simpler and still allows for nonlinear change in probabilistic forecasts over time (because the linear equation is on the probit scale). Alternatively, autoregressive/moving average functions may help account for autocorrelation in forecasts over time.

For stability in model estimation, the t_{ij} from Equation (6) are scaled so that the values stay relatively close to zero. While this scaling can ease model estimation, it can also make it more difficult to accurately estimate the b_0 parameters (because the scaling leads to few data points with large, negative values of t_{ij}). We have still found the scaling to be useful, however, because fast model convergence has been more important to us than the accuracy of b_0 estimates.

The above model assigns four unique parameters to each item: two item easiness parameters b_0 and b_1, a loading λ, and an error variance ψ. We could also allow b_2 to be unique to each item, though experience has indicated that this is not necessarily helpful (further discussion on this point later). Regardless of this issue, the model parameters can be transformed to IRT parameters using theoretical results on the equivalence between factor models and item response models (e.g., Takane & de Leeuw, 1987). In effect, the probit forecasts y^*_{ij} are treated as latent variables giving rise to the binary outcomes associated with item resolution. The parameter transformations are useful for interpretation purposes, whereby one wishes to describe items via their difficulty and discrimination rather than the four parameters from the factor model.

To convert the factor model parameter estimates to item response estimates, we take (Takane & de Leeuw, 1987)

$$\gamma_{0j} = b_{0j}\psi_j^{-1/2} \tag{7}$$

$$\gamma_{1j} = b_{1j}\psi_j^{-1/2} \tag{8}$$

$$\gamma_j = b_{0j} + (b_{1j} - b_{0j})\exp(-b_2 t_{ij}) \tag{9}$$

$$\alpha_j = \lambda_j\psi_j^{-1/2}, \tag{10}$$

with the ability estimates θ_i remaining unchanged. Equation (9) assumes a fixed timepoint t_{ij}, so that the equation is related to item j's difficulty at the time individual i provided a forecast. These transformed parameters then correspond to a two-parameter IRT model

(i.e., two unique parameters per item) for binary outcomes y_{ij}:

$$y_{ij} \sim \text{Bernoulli}(p_{ij}) \tag{11}$$

$$\text{probit}(p_{ij}) = \gamma_j + \alpha_j \theta_i, \tag{12}$$

where, as just mentioned, γ_j is related to item j's easiness at time t_{ij} and α_j is the extent to which item j discriminates between forecasters of different ability. These transformed parameters provide no more information than the original, factor analysis parameters. However, as mentioned above, the transformed parameters' interpretations are advantageous for forecasting situations. Additionally, the parameter transformations illustrate that previous IRT model extensions are relevant to the modeling of forecasting data.

We now mention some additional issues before moving to model estimation. First, as shown in Equation (4), the person ability parameter θ_i is typically treated as a random effect. Therefore, these parameters are not typically estimated using ML methods but can be obtained via, e.g., regression-based methods or Bayesian methods (we use the latter below). Second, while the models outlined above are related to the traditional, 2-parameter IRT model, we can also obtain a Rasch-like model as a special case (using a probit, as opposed to a logit, link function). This is generally accomplished by restricting the λ_j to all be equal and the ψ_j to all be equal.

Finally, we can define some simpler models that are useful for model comparison. These models include only item effects or only forecaster effects:

$$y^*_{ij} = b_j + e_{ij} \tag{13}$$

$$y^*_{ij} = \theta_i + e_{ij}, \tag{14}$$

where the top equation includes fixed item effects and the bottom equation includes fixed forecaster effects (and the residual in both equations follows a $N(0, \sigma^2_e)$ distribution).

These models can be easily estimated via standard regression or ANOVA methods.

Model Estimation and Use

Estimation of the above models is handled via Bayesian methods. In the analyses below, we generally fit Bayesian versions of the model from (6) via MCMC, transforming the resulting parameters to IRT parameters via Equations (7) to (10). JAGS code (Plummer, 2003) for model estimation is provided as supplemental material to this article.

To improve chain mixing and convergence, we use the parameter expansion technique described by Ghosh and Dunson (2009). In short, we sample from a model with unidentified λ and θ parameters. At each sampling iteration, we transform these unidentified parameters to a unique solution. The unidentified parameters, which we denote λ^* and θ^*, have prior distributions of

$$\lambda_j^* \sim N(0, 1) \ \ \forall \ j \tag{15}$$

$$\theta_i^* \sim N(0, \phi^{-1}) \ \ \forall \ i \tag{16}$$

$$\phi^{-1} \sim \text{Gamma}(.01, .01). \tag{17}$$

These sampled parameters are then transformed to the desired model parameters via

$$\lambda_j = \text{sign}(\lambda_1^*)\lambda_j^*\phi^{-1/2} \tag{18}$$

$$\theta_i = \text{sign}(\lambda_1^*)\phi^{1/2}\theta_i^*, \tag{19}$$

where $\text{sign}(\lambda_1^*)$ equals either 1 or -1, depending on whether λ_1^* is positive or negative.

Other model parameters are not involved in the parameter expansion approach. They

receive priors of

$$b_{0j} \sim N(0, .5) \;\; \forall \; j \tag{20}$$

$$b_{1j} \sim N(0, .5) \;\; \forall \; j \tag{21}$$

$$b_2 \sim N(0, .5) \tag{22}$$

$$\psi_j \sim \text{Inv-Gamma}(.01, .01) \;\; \forall \; j, \tag{23}$$

where the normal distributions are parameterized with precisions, as opposed to variances. In the analyses below, the priors on the bs appear to have little impact on the outcomes unless the precision terms are very small. In that case, parameter convergence sometimes fails when forecasts associated with an item become extreme (close to 0 or 1). This is because the model parameters are operating on the probit scale, where they can stray towards $-\infty$ or $+\infty$.

To monitor convergence, we use time series plots and the Gelman-Rubin statistic (Gelman & Rubin, 1992). Throughout the analyses below, these statistics always indicate that the model parameters converge to the posterior distribution. Missing data are assumed to be missing at random (e.g., Little & Rubin, 2002), so that the missingness mechanism can be ignored. This assumption is generally valid here (described further below) but is not likely to be valid in all applications. We expand on this issue in the General Discussion.

Application: Geo-political Forecasts

To study the model's application to real data, we used geopolitical forecasts collected from September, 2011 to April, 2013 as part of a forecasting tournament. The tournament included five university-based teams and was sponsored by the Intelligence Advanced Research Projects Activity. The forecasts used here arise from the team that won the tournament.

Methods

1,593 national and international participants submitted probability judgments associated with 199 questions on the `goodjudgmentproject.com` web site. Participants were randomly assigned to different conditions (involving whether or not they received probability training and whether or not they worked in groups), and they were encouraged to forecast as many questions as possible over time. Further details on the data collection methods can be found in Mellers et al. (2014).

In the current paper, we focus on a subset of the full dataset: 241 participants who responded to nearly all of 133 two-alternative forecasting questions. We chose this subset because we expected it to yield the most reliable results. In particular, because these subjects forecasted nearly all of the questions, they have the most experience and should therefore have the most stable ability levels. Additionally, we focused on two-alternative questions for ease in modeling: while questions with more than two alternatives could be incorporated into the model, these questions may differ from the others in difficulty or in response strategies. We wished to avoid this heterogeneity. Finally, the focus on participants who responded to nearly all the questions helped us to avoid complicated missing data issues. In particular, it is unlikely that the full dataset fulfills the missing at random assumption (e.g., Little & Rubin, 2002) that is commonly employed. While reliance on frequent forecasters solves the missing data issues, it also creates a selection effect: the frequent forecasters undoubtedly differ from infrequent forecasters, so that parameter estimates are likely to change if we included infrequent forecasters.

In the following sections, we first examine the model's predictive ability. In traditional IRT contexts, such an examination is complicated by the facts that (i) the observed data consist of only 0s and 1s, and (ii) traditional estimation methods (e.g., marginal ML) do not directly estimate the θ parameters. However, study of out-of-sample predictions is straightforward in the current, Bayesian context.

Following the examination of predictions, we compare the models' estimates of

forecaster ability to the average Brier score, which is often used as a metric of forecaster ability. After making a general comparison, we further examine the precision of ability estimates when a small number of items have been forecasted. Finally, we use the IRT models to study the impact of time and of item covariates on item difficulty.

Out-of-Sample Predictions

There are many ways to study the model's ability to accommodate the observed data, including posterior predictive checks, residual analyses, and out-of-sample predictions. We focus on the latter option here, randomly deleting 30% of the forecasts from the original dataset. Following this deletion, there still exists some data from each item and each judge. These data are used to fit the model from Equation (6), obtaining estimates of γ and α for each item and of θ for each judge (see Equation (12)). The estimates are then used to predict the deleted forecasts, allowing us to compare the predictions to the observed, held-out data.

Results. Figure 1 displays observed forecasts (x-axis) vs out-of-sample model predictions (y-axis), both of which are probit transformed. The right panel displays predictions for the IRT model described previously, while the other panels display predictions for the "fixed item effects only" model (Equation (13)) and the "fixed user effects only" model (Equation (14)). This figure displays both some successes and shortcomings of the IRT model. Focusing on successes, we see a general positive trend in the right panel, with a correlation of 0.66 between observed and predicted forecasts. This implies an R^2 of 0.43. In contrast, the model that only includes effects of items (and not of forecasters or time) has an R^2 around 0.15, and the model that only includes effects of forecasters (and not of items or time) has an R^2 of about 0.05. Thus, we can conclude that the dynamic IRT model is making better predictions than simpler counterparts.

Focusing on shortcomings, the model predictions appear to be worst for probit forecasts around -3 or 3, which correspond to probabilistic forecasts of 0 or 1. This

partially reflects the fact that forecasters over-use forecasts of 0 or 1 because these forecasts have clear verbal interpretations (see, e.g., Fischhoff & Bruine de Bruin, 1999). Relatedly, the model predictions have a lower bound around -2, which corresponds to a probability of about .12. This means that forecasters are never predicted to provide incorrect, extreme forecasts. Later in the paper, we consider some model extensions to address these issues.

To further examine the model's fit, Figure 2 displays residual plots of model predictions (out of sample) by item and by forecaster, respectively. The x-axes contain item and forecaster IDs, which are ordered by average Brier score (the user/item with best Brier score has ID 1, and higher IDs reflect users/items with worse Brier scores). The forecaster plot indicates some heteroscedasticity, with points on the right side having greater variability than points on the left side. Additionally, some outlying negative residuals appear in both plots. These come from situations where the forecaster assigned probabilities near zero to the realized outcome; these issues are further incantations of the previously-discussed issue associated with model predictions.

Ability Estimates

The previous section examined out-of-sample model predictions using a partially-deleted dataset. In this section, we fit the same model to the full dataset and to small numbers of items. We generally compare model-based ability estimates to Brier scores, as the Brier score is often the default forecaster evaluation metric in practice. The Brier score weights each item equally, however, which may be suboptimal in forecasting environments. The item response model, on the other hand, differentially weights items depending on the extent to which they discriminate forecasters of different abilities. The prior distributions used to estimate the models were the same as the prior distributions used in the previous section.

Full Data Estimates. Figure 3 compares model-based ability estimates to mean Brier scores for the 241 forecasters in the data. There is a clear relationship between the

two metrics, with a Spearman correlation of -0.43 (the correlation is negative because larger Brier scores are bad, while larger ability estimates are good). The model-based ability estimates are more closely related to each forecaster's median Brier score (not shown), with a Spearman correlation of -0.84. Use of the median Brier score diminishes the impact of bad forecasts, so this strong correlation implies that the model also exhibits a diminished impact of bad forecasts.

In addition to the above, we examined the relationship between the model-based ability estimates and mean standardized Brier scores (i.e., Brier scores converted to z-scores on an item-by-item basis). The standardized Brier score is a heuristic method for adjusting scores based on item difficulty; a forecaster with a bad Brier score can still get a good standardized score if her prediction was better than the crowd. We expected that the standardized Brier scores be more similar to the IRT estimates because they are both relative measures. Nonetheless, the scatter plot (not pictured) looks similar to Figure 3, with the Spearman correlation between standardized Brier scores and model-based estimates being -0.59.

To further study relationships between the ability metrics, we identified four individuals in Figure 3 for closer comparison. The first two individuals are the person with the best Brier score (red plus) and the person with the best IRT-based ability estimate (red X). We also selected two individuals who are related to these first two: the person who is most similar to the best Brier score on ability while being most *dissimilar* on Brier score (black plus), and the person who is most similar to the best ability estimate on Brier score while being most dissimilar on ability (black X). We plotted these individuals' forecasts to examine their response styles and the items for which they provided good/bad forecasts. These plots are displayed in Figure 4; the x-axis displays item IDs (ordered by mean Brier score), while the y-axis displays the Brier score that each person obtained for each item. The top row displays plots for the "plusses" (similar ability estimates, dissimilar Brier scores), while the bottom row displays plots for the "Xs" (similar Brier scores, dissimilar

ability estimates). Points are shaded to represent the time at which they were made; lighter points represent forecasts that were made closer to event resolution.

Comparing the two panels in the top row, we see that the individuals arrived at similar IRT ability estimates in different manners. The person in the left panel generally made good forecasts up to item 100 (approximately), then made bad forecasts. In contrast, the person in the right panel generally made some bad forecasts across the full range of items, but the bad forecasts were often made far from event resolution (i.e., many points are dark). The IRT model takes into account the time at which each forecast is made, which resulted in the similar ability estimates assigned to these two individuals. This reflects a major advantage of the model-based estimates: they allow us to account for time of reported forecast.

Comparing the two panels in the bottom row, we see that the IRT model rewarded the individual on the left who generally made extreme forecasts. The person on the right, on the other hand, generally reported forecasts closer to .5, which resulted in a lower ability estimate. Comparison of the two panels in the left column reinforces these ideas: we see that the person with the best Brier score (top left panel) avoided extreme errors, whereby one assigns a probability of 1 to the incorrect outcome. However, the person also made less-extreme judgments when she was uncertain, and she had particular trouble with the most difficult items (as measured by the Brier score; those on the far right side of the graph). In contrast, the person with the best ability estimate (bottom left panel) used extreme judgments and often made a perfect forecast on the most difficult items. The model therefore rewarded this person for usually making perfect forecasts, even for many difficult items. As compared to the Brier score, we see that the model rewards correct, extreme judgments and lightly penalizes incorrect, extreme judgments. In short, the model rewards forecasters who are willing to take risks, while the Brier score rewards forecasters who are more conservative. While different decision makers will value different types of forecasters, we found the model's handling of extreme forecasts to be appealing. In the

General Discussion, we provide further thoughts about model-based ability estimates and proper scoring rules.

Estimates from Few Items. Are the IRT ability estimates related to future performance on unseen items? To answer this question, we created a scenario where we observe a small number of forecasts from a small group of individuals (and many forecasts from a large group of individuals). We estimate the ability of the small group via IRT models and Brier scores. We then examine whether the small group's ability estimates are related to their ability estimates from a separate set of test forecasts.

In this analysis, we partitioned the original data to create twenty sets of training and test data. Within each set, the training data were comprised of all forecasts from 90% of forecasters and ten randomly-chosen forecasts from the remaining 10% of forecasters (the "test" forecasters). Using the training data, we measured the test forecasters' abilities via three methods: mean Brier score, mean standardized Brier score (described in the previous section), and the IRT model (Equation (6)).

After obtaining ability estimates with the training data, we used the test data to re-estimate test forecasters' abilities. The test data included 66 new forecasts from the test forecasters and all forecasts from the other forecasters. This analysis was set up to mimic applied situations where we have large amounts of data on some forecasters, and we wish to accurately measure new forecasters' abilities using small amounts of data (here, ten forecasts).

The results are most efficiently described via Spearman correlations between each training measure and its test measure. These results are displayed in Table 1, containing the mean correlation (with interquartile ranges) across the twenty replications. We generally observe greater consistency in the IRT ability measures from the training to the test set, as compared to the Brier scores. In particular, the mean correlation between training and test Brier scores is .48; between training and test standardized Brier scores is .59; and between training and test IRT estimates is .73. While the intervals associated with

each measure overlap, the ordering is consistent across all twenty replications: the IRT estimates always exhibit the largest correlation, followed by the standardized Brier scores, followed by the regular Brier scores.

These results suggest that, when a forecaster provides only a small number of forecasts, the IRT model provides an ability estimate that better generalizes to future IRT ability estimates. This is because the IRT model has two advantages over the Brier score: it can account for the time at which a forecast was made, and it uses information about all forecasters in assigning an ability estimate to any single forecaster. The former advantage can help the model handle incorrect, long-term forecasts; a forecaster's ability estimate is not penalized as much if he/she makes a bad forecast at a time when the item is difficult. The latter advantage can help the model account for item difficulty; it lightly penalizes bad forecasts on items that are difficult for everyone, and it heavily penalizes bad forecasts on items that are easy. While standardized Brier scores address the latter advantage, they do not address the former advantage.

Relation of Model Parameters to Covariates

Finally, we examine covariates both within and outside the model. We first examine the estimated impact of time on item difficulty. We then examine the extent to which the estimated item parameters are related to external ratings of item "surprisingness."

Effects of Time. In comparing the b_0 and b_1 estimates associated with each item, we found that $b_1 > b_0$ for all but one item. Because b_1 represents easiness at item resolution and b_0 represents easiness as days to resolution go to ∞, this implies that items usually become easier over time. In Figure 5, we graph the estimated effect of time on item easiness for three randomly-selected items. As we move from left to right in the figure, we get closer to event resolution.

It is seen that, at 80 days before item resolution, the item represented by the solid (dashed) line is easiest (hardest). These difficulties change differentially over time, so that,

around 30 days before item resolution, the dashed line is no longer the most difficult. Finally, near 0 days before resolution, the solid and dashed lines converge. The items that varied greatly in easiness around 80 days to resolution have now become equally easy.

These results show that the model estimates support our intuition of items becoming easier over time. Additionally, they illustrate why we employ a b_2 parameter that is fixed across items: the b_0 and b_1 parameters specify the magnitude of change in item difficulty over time, which in turn influences the speed at which items change. We do not need an additional, item-specific b_2 parameter to obtain different rates of change across items.

Relation to External Covariates. We also compared the model's estimated parameters to an external variable: the extent to which the outcome of each item was surprising. After each item resolved (i.e., ex post), surprisingness was rated by two subject matter experts (who did not create the items) on a scale from 0–10, where 10 means "most surprising" and 0 means "least surprising." An item was given a high rating if its outcome would have been surprising to an informed observer at the time the item was intially created (when the outcome was unknown). Ratings were averaged across the two raters for the analyses here.

Model parameters for comparison included each item's easiness (γ_1), discrimination (α), and residual variability (ψ). Surprising items should be related to these parameters: items whose outcomes are surprising may be more difficult, they may fail to discriminate between forecasters of differing abilities, and they may induce greater residual variability. Therefore, if the model's item estimates match reality, they should be related to the external ratings.

Figure 6 displays scatterplots between the model's item estimates and the external ratings (below the main diagonal), with correlations between each pair of variables appearing above the main diagonal. Focusing on the signs of model parameters' correlations with surprise ratings, we see that less-surprising items are easier, discriminate better between forecasters of different ability, and have lower residual variability. The

magnitudes of these correlations are relatively small, being −.09 (easiness), −.26 (discrimination), and .14 (residual variability). The previously-discussed model shortcomings likely dampen the correlations, as well as the residual variability in the ratings of surprisingness. Overall, however, the correlations between item parameters and the external variable are in the expected directions and provide further evidence that the model is capturing the data in a meaningful way.

Model Evaluation Summary

In summary, the model predictions are reasonably accurate, and the model parameter estimates correspond to our expectations about their behavior (as illustrated through the relationship with external variables). The main weakness of the model lies in its inability to accommodate incorrect, extreme forecasts. The model is also unable to account for the over-use of forecasts of .5. These issues are potentially addressed via an item response model of ordinal judgments, which is studied in the next section.

Ordinal Model

As developed thus far, the IRT model does a poor job of handling the extreme forecasts of 0 and 1. Furthermore, the model cannot handle the overuse of .5 that is commonly observed in many datasets (Bruine de Bruin, Fischhoff, Millstein, & Halpern-Felsher, 2000). To handle these issues, multiple model alterations are possible. The alteration that we consider in detail here (with others being described in the General Discussion) involves treating the reported forecasts as ordered categories instead of as continuous judgments. This mimics the manner by which confidence judgments are typically collected and modeled in cognitive psychology tasks (Pleskac & Busemeyer, 2010; Ratcliff & Starns, 2009; Van Zandt, 2000). For example, in recognition memory experiments, subjects must report whether or not a focal word appeared on a previous list of words. Their confidence is not typically reported on a continuous scale; instead it may be reported on a 6-category scale ranging from “certain the word previously appeared” to

"certain the word did not appear." The rationale for this scale is that subjects can only discriminate between a small number of confidence levels (say, uncertain, somewhat certain, and certain). This implies that geopolitical forecasts of, e.g., .2 and .4 might be treated equivalently because forecasters have no cognitive basis for discriminating between such judgments.

Model Detail and Method

The model described here is an ordinal version of the continuous model that we described previously. We assume a continuous, latent variable that drives the judgments Y^*. This variable is similar to the probit-transformed forecasts used previously, except that the variable is now unobserved. Multiple threshold parameters determine the reported, ordinal category. The model employs ideas related to the graded response model that is popular in psychometrics (Samejima, 1969, 1997), the signal detection models that are popular in cognitive science (DeCarlo, Kim, & Johnson, 2011; Wickens, 2002), and the proportional odds model that is popular in statistics (Fullerton, 2009; McCullagh, 1980). The thresholds change over time in the current application, reflecting the fact that problem difficulty tends to change over time.

Formally, we define the latent variable as

$$y^*_{ij} = \lambda_j \theta_i + e_{ij}, \tag{24}$$

where the e_{ij} and the θ_i arise from separate standard normal distributions. Assuming m ordered categories, we then compute $(M-1)$ threshold parameters via the equation

$$q_{jm} = b_{0jm} + (b_{1j} - b_{0jm}) \exp(-b_2 t_{ij}) \quad m = 1, \ldots, (M-1), \tag{25}$$

where the bs are free parameters, t_{ij} was defined previously (as time until the problem

resolves), and $b_{0j1} < b_{0j2} < \ldots < b_{0j(M-1)}$. The observed response y_{ij} is then determined via

$$y_{ij} = \begin{cases} 1 & \text{if } y^*_{ij} < q_{j1} \\ 2 & \text{if } y^*_{ij} > q_{j1} \text{ and } y^*_{ij} < q_{j2} \\ \vdots & \\ M & \text{if } y^*_{ij} > q_{j(M-1)}. \end{cases} \tag{26}$$

This model can also be written via probit-transformed, cumulative probabilities of responding to each category or below; further detail on the multiple ways of writing these models is provided by Smithson and Merkle (2013).

We fit the model to the full dataset used earlier in the paper, with $M = 5$: probability judgments of 0 were coded as category 1, probability judgments in $[.01, .49]$ were category 2, .5 was category 3, $[.51, .99]$ was category 4, and 1 was category 5. This coding collapses away much information, and codings for larger values of M could be conceptualized. We were interested in the extent to which forecaster ability estimates differed when treating the forecasts as ordered categories, as compared to continuous judgments. This comparison provides information about the extent to which the original model's (mis)handling of extreme forecasts impacts ability estimates, as well as the extent to which continuous forecasts are more useful than categorical forecasts. Prior distributions on model parameters were the same as those used for the original model.

Results

A scatter plot of forecaster ability under the continuous model (x-axis) and under the ordinal model (y-axis) is displayed in Figure 7. This illustrates a surprisingly-strong correlation of 0.87 between the models' ability estimates, with the best forecasters being especially similar across the models. The low-ability forecasters exhibit greater variability between models, and this is likely related to the original model's problems with incorrect, extreme judgments: bad forecasters provide more of these judgments, so that the resulting

ability estimates diverge.

The similarity between the two models' ability estimates provides interesting results about the original model and about the forecasters. Focusing on the former, the results imply that the original model's shortcomings involving extreme judgments do not have a large impact on the resulting ability estimates. Focusing on forecasters, it was surprising that results remained similar when we discard a large amount of data; when all continuous judgments from .01 to .49 (and from .51 to .99) are collapsed into a single category. While this implies that it may be sufficient to elicit forecasts on an ordinal, as opposed to continuous scale, ordinal scales may induce other problems. First, the optimal number of ordered categories (M) to use in any particular application is unclear, and fixing M to a small value may be too restrictive in some applications. A forecaster may report a "certain" judgment on one item, then encounter a new item for which she is more certain than she was on the original item. Second, the use of ordered categories may disallow base rate information that is important for some geopolitical items. For example, items involving elections may benefit from polling data, results of similar elections, and so on. This information may yield a quantitative estimate that is more precise than an ordered category, and the forecaster may use this estimate as a prior probability for judging the focal election. Forcing the forecaster to use an ordered category may not allow the forecaster to communicate all of her knowledge.

General Discussion

In this paper, we first tailored item response models to probability judgments. These models can handle many issues that traditional forecast evaluation methods cannot, including the facts that (i) forecasters forecast the same item at different points in time, (ii) forecasters forecast different subsets of items, and (iii) items differ in the extent to which they measure ability. Therefore, the models offer novel methods for evaluating forecasters and questions in realistic situations. In fitting the models to data from a

geo-political forecasting tournament, we empirically observed both strengths and weaknesses of the models. Strengths included the facts that the model's out-of-sample predictions were reasonable and that the model's item estimates were related to external measures of item closeness and surprisingness. Weaknesses were related to extreme, incorrect forecasts: the model could not accommodate these forecasts, and it may not adequately penalize forecasters who were frequently extreme and incorrect. In the sections below, we describe extensions to handle these and other issues. These extensions are related to extreme forecasts, "proper" ability estimates, item evaluation, missing data, and dimensionality of forecaster ability. We also consider some methodological issues involved in the evaluation of probability judgments.

Extreme forecasts

As described previously, we implemented an ordinal model to address the original model's inability to accommodate extreme forecasts (and overuse of .5). While the ordinal model handled extreme forecasts, it was unsatisfactory from the standpoint that data were being discarded (i.e., many distinct forecasts were collapsed into a single category). It may instead be worthwhile to consider alternative models that treat the forecasts as continuous while simultaneously accounting for overuse of forecasts in $\{0, .5, 1\}$. One alternative involves development of a two-component mixture model that can handle overuse of particular forecasts, which is related to the traditional three-parameter item response models (Birnbaum, 1968). The traditional model assumes that individuals sometimes respond correctly by guessing, regardless of their ability. In a forecasting mixture model, we might instead assume that forecasters sometimes provide judgments of 0, .5, or 1 that are unrelated to the specific question's difficulty or to the specific forecaster's ability. There are likely to be estimation difficulties associated with this model due to the parameter identification issues inherent in mixture models (see Smithson, Merkle, & Verkuilen, 2011, for a general discussion of mixture models for probability judgments).

Other researchers have addressed overuse of forecasts through rounding or through modified link functions. For example, Kleinjans and Van Soest (2014) develop a model whereby subjects arrive at probability judgments through various types of rounding. A judgment of, say, .07 implies that the subject's true belief lies somewhere between .065 and .075; her judgment is rounded to the nearest multiple of .01. An extreme judgment of 1, however, may be obtained through other types of rounding: the subject may be rounding to the nearest multiple of .1, so that the judgment of 1 merely implies that her true belief is greater than .95. The overuse of $\{0, .5, 1\}$ arises because subjects can arrive at these judgments through multiple types of rounding (to the nearest multiple of .01, .05, .1, .25, or .5). These multiple types of rounding lead to a five-component mixture model, which makes it challenging to implement rounding within the model proposed here.

Related to rounding, others have employed truncated or censored distributions to account for the extreme judgments of 0 and 1. Muthén (1989) uses a censored model to handle extreme judgments, whereby all judgments that are more extreme than the censoring point simply assume an extreme value of 0 or 1. Ferrando (2001) relatedly describes use of a truncated normal distribution to model bounded responses (such as probability judgments), using an identity link function instead of the logit or probit. Neither of these models immediately handle the abundance of .5 judgments, though additional censoring points may be employed to handle these judgments.

Proper ability estimates

Scoring metrics such as the Brier score or logarithmic score are popular in forecasting contexts because they are *proper*: forecasters can expect to receive the best score when their forecasts match the true probability of event occurrence. This, in turn, is thought to motivate forecasters to be honest; to report their honest belief about the probability of event occurrence.

The model-based ability metric proposed in this paper (from the IRT model of

continuous judgments) is not proper. For a traditional, two-parameter IRT model applied to binary data, the sufficient statistic for ability (θ_i) is $\sum_j \alpha_j y_{ij}$, where α_j is item j's discrimination and y_{ij} indicates whether or not subject i correctly answered item j (e.g., Baker & Kim, 2004). Translating these results to the current model (in Equation (6)), we can show that the sufficient statistic for θ_i involves a weighted sum of the y^*_{ij}. This amounts to an "absolute error" metric, which encourages reporting of a median judgment instead of a mean judgment (Gneiting, 2011). That is, assuming one's uncertainty about the forecast is represented by a probability distribution, then the proposed model encourages reporting of that distribution's median instead of the mean. Models whose ability estimates are related to the Brier score or to other proper scoring rules require further study, but the inclusion of specific scoring rules into traditional IRT models has recently been considered by Bo, Lewis, and Budescu (2015).

Item attributes

In traditional IRT applications, models provide estimates of item attributes (typically difficulty and discrimination) in addition to estimates of respondents' ability. The item attributes are especially useful because the items can be re-used: we can administer test items to an initial group of students, then select the best items for wide-scale administration. This is different from many forecasting situations, where, once we know an item's outcome, the item is expired and cannot be reused. As a result, we cannot immediately use IRT models to pre-determine which items are of suitable difficulty. Instead, we could extend the models to include additional item covariates that are related to difficulty (e.g., De Boeck & Wilson, 2004). This would allow us to estimate the impact of covariates on item difficulty, providing information about the relative difficulty of different item types.

Missing data

Many forecasting scenarios lead to large amounts of missing data: forecasters may respond to only a small proportion of items, new forecasters may enter the panel, and old forecasters may leave the panel. In the current paper, we bypassed these issues by examining a subset of frequent forecasters. Here, we consider some model extensions to handle the missing data issues.

Focusing on forecasters who respond to a small proportion of items, it is easiest to ignore the missing data points and model only the observed data. This amounts to assuming that the data are missing at random (e.g., Little & Rubin, 2002), an assumption that can be violated if, e.g., the good forecasters are able to select easy questions. When this assumption is violated (so that the data are missing not at random), then we must simultaneously model the reported forecasts and a binary variable denoting whether or not each forecaster responded to each item. The approach of O'Muircheartaigh and Moustaki (1999) appears promising for accomplishing this. Their approach involves a two-dimensional ability model, where the first dimension is forecaster ability and enters the model in the same way as the model used in the current paper. The second dimension is then "response propensity," and it is used to model the binary response data (indicating whether or not a forecaster responded to an item). Importantly, this second dimension also influences the reported forecasts, which allows for situations where the reported forecasts are related to item selection. This approach is an extension of the model reported in this paper and can potentially be estimated via Bayesian methods.

The situation where new forecasters enter and old forecasters leave can be addressed via *linking* and *equating* methods that have been thoroughly studied in traditional IRT contexts (e.g., von Davier, 2013). In short, these methods allow us to place forecasters' abilities on a common scale even when different forecasters respond to different items. To do so, there is typically a requirement that either (i) everyone has responded to at least one common item, or (ii) some people have responded to all items. These requirements may be

partially relaxed, however, for situations where each person potentially responds to a unique subset of items. Fischer (1981) describes necessary and sufficient conditions under which the Rasch model can be fit to these types of incomplete data. Importantly, responses must be "connected," in the (rough) sense that each person's chosen items must overlap with other people's chosen items. For forecasting contexts, it is perhaps safest to present incoming forecasters with some practice items to which everyone has responded. Alternatively, if there are a subset of committed forecasters who respond to all items, then this should be sufficient to employ the methods described here.

Dimensionality of ability

The IRT models described in this paper assume that ability is unidimensional; that each forecaster's ability can be described by a single number. In the previous section, we discussed adding a dimension to accommodate missing data issues. We could more generally assume multiple dimensions of forecasting ability, where different dimensions reflect different types of ability. For example, one dimension might reflect ability to make long-term estimates, a second dimension might reflect subject-matter knowledge, and so on. These models are increasingly difficult to fit as the number of dimensions increase. Additionally, the dimensions may be modeled as either *compensatory* or *non-compensatory*. In the former case, low ability on one dimension can be offset by high ability on another dimension. In the latter case, one must have high ability on all relevant dimensions in order to exhibit good performance. These extensions, along with the previous extensions to address model shortcomings, may lead to important advances in forecast assessment. Further work is needed to ensure that these complex models can be reliably estimated in general forecasting situations.

Methodological issues

Finally, we remark on some issues surrounding the statistical analysis of probability judgments. Previous researchers (Erev, Wallsten, & Budescu, 1994; Wallsten, 1996) have

noted that different probability judgment analysis methods can lead to conflicting conclusions about the judges/forecasters. These researchers have focused on calibration, which is the correspondence between the probability judgments and the outcomes. For example, if a well-calibrated forecaster repeatedly reports forecasts of 60%, then 60% of those events should occur. Erev et al. (1994) showed that one's conclusions about calibration can be reversed, depending on whether probability judgments are treated as response variables or as predictor variables in the analysis. The modeling of probability judgments conditioned on objective probability of event occurrence is associated with underconfidence, and the modeling of event outcomes conditioned on probability judgments is associated with overconfidence. Wallsten (1996) further summarizes the issues and notes that "The most useful analyses generally will be those that rely on the response distributions conditional on true or false statements or on objective probabilities" (pp. 225–226).

The models proposed here implicitly condition on event outcomes (which Wallsten calls "true or false statements"), because our Y^* is defined as the probability judgment for the outcome that occurred. However, the impact of this conditioning on our conclusions is unclear because we are not characterizing judges' calibration in an absolute sense (i.e., we are not characterizing judges as underconfident, overconfident, or well-calibrated). Instead, we are characterizing judges' relative forecasting ability, which is partially based on calibration but also includes aspects of discrimination, noise, and so on. Further, we draw no conclusions about judges' absolute ability: the model can tell us, e.g., which judge is better than all the others, but it does not tell us whether that judge's forecasts are highly predictive of the event outcomes. We conjecture that the relative nature of the proposed model resolves some of the above-noted statistical issues.

Conclusion

In summary, IRT models of forecasts afford a useful framework for evaluating forecasters and items in realistic environments. They formalize reseachers' intuitions about question difficulty dynamically changing over time and about questions' discriminations between forecasters of varying abilities. Further, many model extensions are available that allow for data analyses that were impossible with other traditional analyses (say, scoring rules or traditional regression/ANOVA models). These extensions provide avenues for future research, resulting in a family of models that can potentially handle many dynamic forecasting situations.

References

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). Boca Raton, FL: CRC Press.

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

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–472). Reading, MA: Addison-Wesley.

Bo, Y., Lewis, C., & Budescu, D. V. (2015). An option-based partial credit item response model. In R. E. Millsap, D. M. Bolt, L. A. van der Ark, & W.-C. Wang (Eds.), *Quantitative psychology research* (Vol. 89, pp. 45–72). Springer.

Böckenholt, U. (1993). Estimating latent distributions in recurrent choice data. *Psychometrika*, *58*, 489–509.

Bruine de Bruin, W., Fischhoff, B., Millstein, S. G., & Halpern-Felsher, B. L. (2000). Verbal and numerical expressions of probability: "It's a fifty-fifty chance". *Organizational Behavior and Human Decision Processes*, *81*, 115–131.

Budescu, D. V., & Johnson, T. R. (2011). A model-based approach for the analysis of the calibration of probability judgments. *Judgment and Decision Making*, *6*, 857–869.

de Ayala, R. J. (2009). *The theory and practice of item response theory.* New York, NY: Guilford.

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software*, *39*, 1–28.

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach.* New York: Springer-Verlag.

DeCarlo, L. T., Kim, Y., & Johnson, M. S. (2011). A hierarchical rater model for constructed responses, with a signal detection rater model. *Journal of Educational*

Measurement, *48*, 333–356.

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

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review*, *101*, 519–527.

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

Fischer, G. H. (1981). On the existence and uniqueness of maximum-likelihood estimates in the Rasch model. *Psychometrika*, *1*, 59–77.

Fischhoff, B., & Bruine de Bruin, W. (1999). Fifty-fifty=50%? *Journal of Behavioral Decision Making*, *12*, 149–163.

Fullerton, A. S. (2009). A conceptual framework for ordered logistic regression models. *Sociological Methods & Research*, *38*, 306–347.

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. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, *106*, 746–762.

Kleinjans, K. J., & Van Soest, A. (2014). Rounding, focal point answer and nonresponse to subjective probability questions. *Journal of Applied Econometrics*, *29*, 567–585.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (p. 306-334). Cambridge, England: Cambridge University Press.

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.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society B*, *42*, 109–142.

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

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. (2013). Choosing a strictly proper scoring rule. *Decision Analysis*, *10*, 292–304.

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.

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

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., ... Rakow, T. (2006). *Uncertain judgements: Eliciting experts' probabilities*. Hoboken: Wiley.

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.

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection theory: A theory of choice, decision time, and confidence. *Psychological Review*, *117*, 864–901.

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.

Ratcliff, R., & Starns, J. (2009). Modeling confidence and response time in recognition memory. , *116*, 59–83.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Tech. Rep. No. 17). Psychometrika Monograph Supplement.

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

Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). New York, NY: Springer.

Smithson, M., & Merkle, E. C. (2013). *Generalized linear models for categorical and continuous limited dependent variables*. Boca Raton, FL: Chapman & Hall/CRC.

Smithson, M., Merkle, E. C., & Verkuilen, J. (2011). Beta regression finite mixture models of polarization and priming. *Journal of Educational and Behavioral Statistics*, *36*, 804–831.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*, 393–408.

Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 582–600.

Verhelst, N. D., & Glas, C. A. W. (1993). A dynamic generalization of the Rasch model. *Psychometrika*, *58*, 395–415.

von Davier, A. A. (2013). Observed-score equating: An overview. *Psychometrika*, *78*, 605–623.

Wallsten, T. S. (1996). An analysis of judgment research analyses. *Organizational Behavior and Human Decision Processes*, *65*, 220–226.

Wickens, T. D. (2002). *Elementary signal detection theory*. New York, NY: Oxford

University Press.

Yates, J. F. (1982). External correspondence: Decompositions of the mean probability score. *Organizational Behavior and Human Performance*, *30*, 132-156.

Table 1
Matrix reflecting mean Spearman correlations (over twenty replications) between training and test ability estimates (Brier scores, standardized Brier scores, and IRT estimates). Interquartile ranges are displayed in parentheses.

	Train Brier	Train Std Brier	Train IRT
Test Brier	0.48 *(0.36–0.56)*		
Test Std Brier		0.59 *(0.47–0.7)*	
Test IRT			0.73 *(0.67–0.79)*

Figure 1. Observed forecasts (probit transformed) vs out-of-sample model predictions. Predictions are based on item effects only (left panel), forecaster effects only (middle panel), and the dynamic IRT model (right panel).

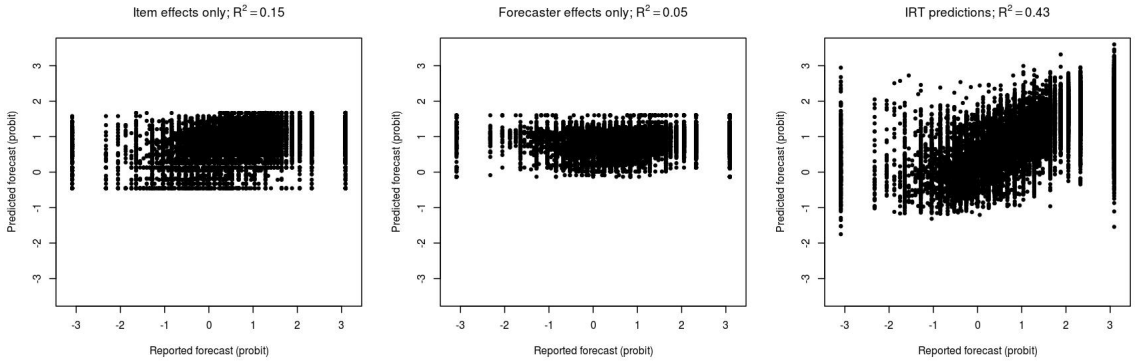

Figure 2. Out-of-sample model residuals by item (left panel) and forecaster (right panel).

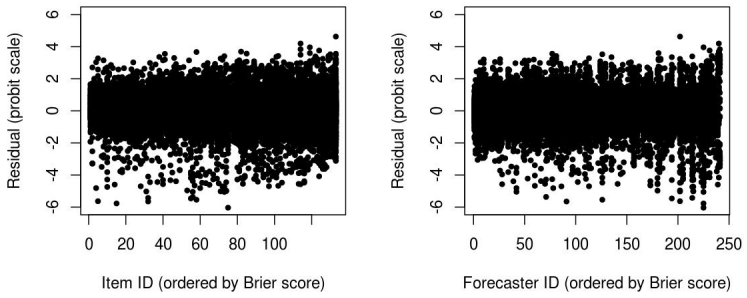

Figure 3. Model-based ability estimates versus mean Brier scores.

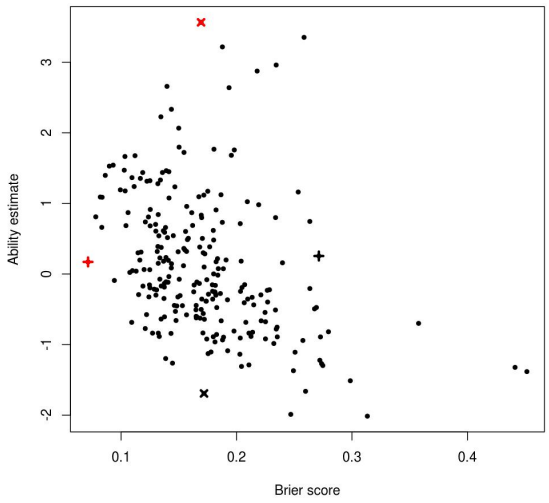

Figure 4. Comparison of best-ranking forecasters, as determined by average Brier score and by the model. Each panel corresponds to one of the points in Figure 3 (see panel labels). Light-colored points represent forecasts that were made closer to event resolution.

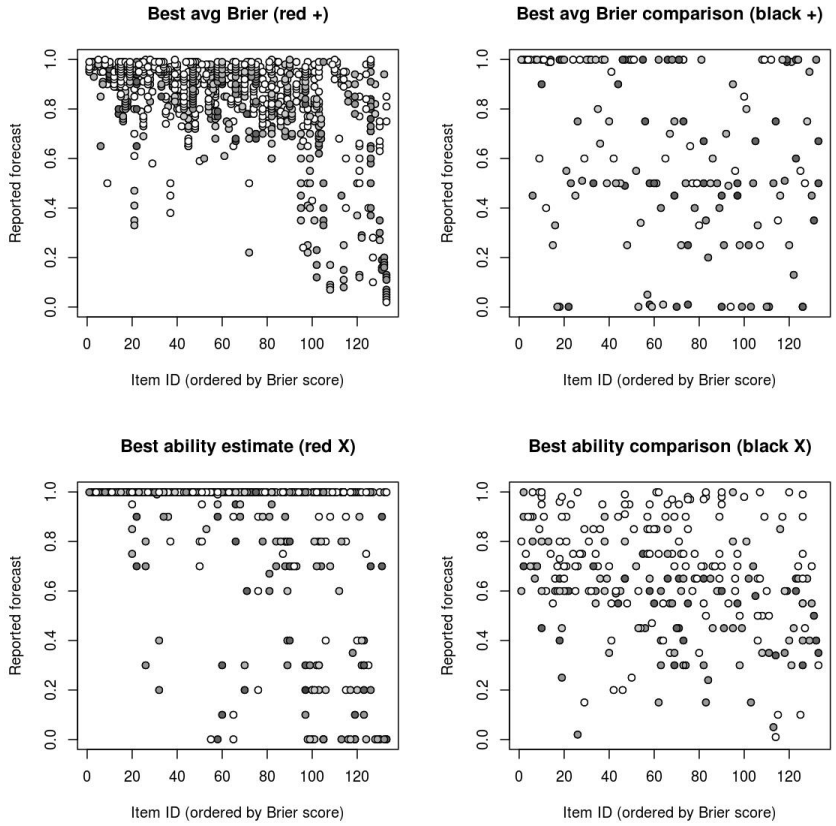

Figure 5. Predicted effects of time on item difficulty, where time is measured by days to resolution (numbers closer to 0 are closer to resolution). Lines depict three different randomly-chosen items.

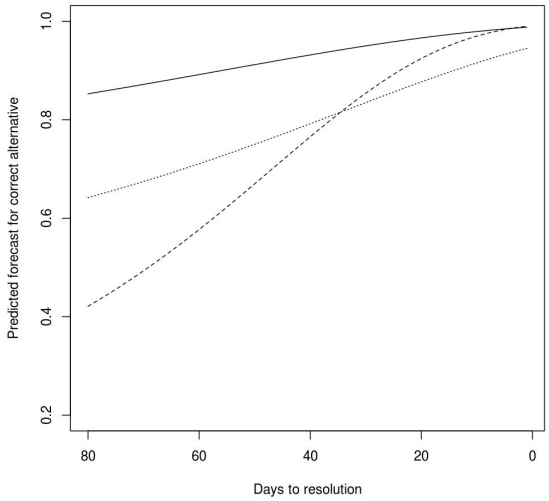

Figure 6. Scatterplots between item parameters and surprise ratings. The γ_1 parameter is related to item easiness at resolution, the α parameter is item discrimination, and the ψ parameter is item residual variability. Pearson correlations are displayed in the upper triangle.

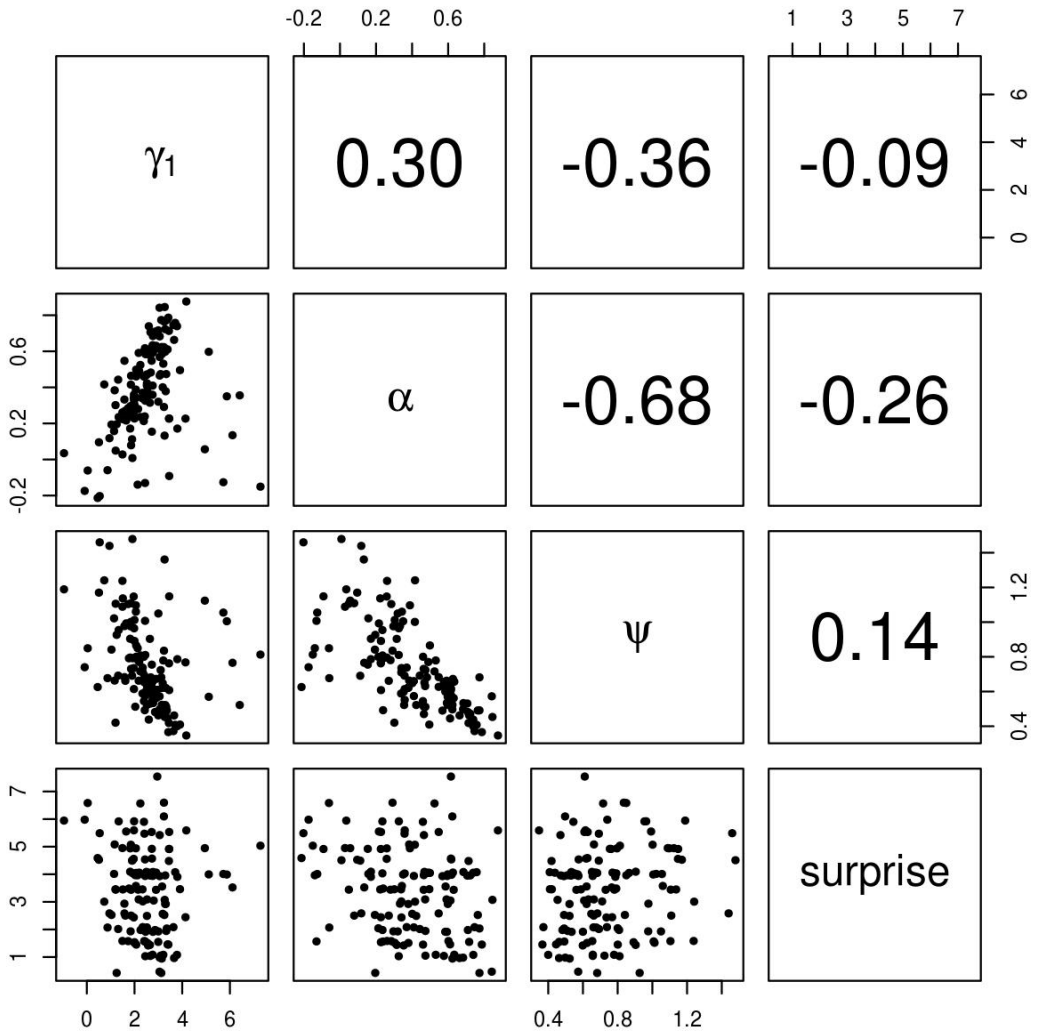

Figure 7. Comparison of forecaster ability estimates from the "continuous forecast" model and from the ordinal model.