Judgment and Decision Making, Vol. 13, No. 2, March 2018, pp. 185–201

Weighted Brier score decompositions for topically heterogenous forecasting tournaments

Edgar C. Merkle* Robert Hartman†

Abstract

Brier score decompositions, including those attributed to Murphy and to Yates, provide popular metrics for estimating forecast performance attributes like calibration and discrimination. However, the decompositions are generally limited to situations where forecasters make successive forecast judgments against the same class of substantive event (e.g., rain vs. no rain). They do not readily translate to common situations where: forecasts are weighted unequally; forecasts can be made against a range of heterogeneous topics and events over varying time horizons; forecasts can be updated over time until an event occurs or an event deadline is reached; or outcome alternatives can vary in number and nature (e.g., ordered vs. unordered outcomes) across forecast questions. In this paper, we propose extensions of the Murphy and Yates decompositions to address these features. The extensions involve new analytic expressions for the decompositions of weighted Brier scores, along with proposed resampling methods. We use data from a recent forecasting tournament to illustrate the methods.

Keywords: Brier score, Murphy decomposition, Yates decomposition, calibration, discrimination, forecasting, probability judgment

1 Introduction

Proper scoring rules comprise one of the most popular classes of metrics for evaluating probabilistic forecasts (e.g., Carvalho, 2016; Gneiting and Raftery, 2007; Merkle & Steyvers, 2013; Winkler & Jose, 2010). The Brier (1950) score, also known as quadratic loss, is particularly popular because it can be decomposed into simple expressions for the forecast properties of calibration and discrimination. In forecasting contexts, calibration roughly refers to the forecaster's reported probabilities matching the long-run proportion of events that resolve in favor of a specific outcome alternative. Discrimination, on the other hand, refers to the forecaster consistently assigning larger (or different) probabilities to realized alternatives, as compared to probabilities of unrealized alternatives.

Approved by the MITRE Corporation for Public Release; Distribution Unlimited. Case Number 17-1687. This research is based on work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via ODNI contract number 2015-14120200002-002. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. ©2017 The MITRE Corporation. ALL RIGHTS RESERVED.

*Department of Psychological Sciences, University of Missouri. Email: merklee@missouri.edu.

†The MITRE Corporation

The Murphy (1973) and Yates (1982) decompositions of the Brier score provide explicit metrics related to a forecaster's calibration and discrimination across a set of resolved questions. They are popular because their components sum to the corresponding mean Brier score and because they can be easily written and computed arithmetically (without, e.g., extra statistical modeling or algorithms). The decompositions are somewhat inflexible, however, because they require (i) all questions to have the same number of consistently-coded outcome alternatives; (ii) question alternatives to be unordered; and (iii) all forecasts to be weighted equally. These requirements would all be fulfilled when we have daily forecasts of whether or not it will rain, or when we have monthly forecasts of whether or not there will be military conflict between a specific pair of countries. In the latter case, for example, the outcome alternatives are always "no conflict" vs. "conflict;" these alternatives have no particular ordering; and any single forecast for a particular pair of countries in a given month would be weighted the same as any other pair of countries in any other month.

These restrictive requirements are frequently violated in applied forecasting activities and competitions (see, e.g., Tetlock, 2005; Mandel & Barnes, 2014). A salient recent example, to which we will refer throughout this paper, is the Aggregative Contingent Estimation (ACE) program. This was a multi-year geopolitical forecasting tournament sponsored by the Intelligence Research Advanced Projects Activity (IARPA; see Tetlock & Gardner, 2016). Over a four-

https://doi.org/10.1017/S1930297500007099 Published online by Cambridge University Press

year period, hundreds of heterogeneous, real-world geopolitical forecasting questions were posed, spanning such topics as regime change, democratic elections, non-state actor violence, the Eurozone crisis, financial market behavior, and infectious disease spread, among others. In addition to their subject matter diversity, tournament competitors were asked to make daily forecast updates of all available questions.

For example, consider forecasting whether or not there would be substantial military conflict between China and N Korea during 2014. Instead of providing a single forecast for this question, competitors supplied a daily forecast until either (i) there was a substantial conflict between the nations, or (ii) the year 2014 ended. Thus, forecast questions varied in their duration, with some opening and closing within a single week, and others staying open for several months. Moreover, although all forecast questions were defined by a discrete set of mutually exclusive and exhaustive outcome alternatives, the questions varied in the number of possible alternatives presented, the labeling of those outcome alternatives (e.g., "yes" vs. "no," Candidate A vs. B vs. C), and in whether those outcomes could be meaningfully ordered. Conventional Brier score decompositions have no way of accommodating these features, so that the modifications described in this paper are required.

2 Brier score decomposition overview

We first review the traditional Murphy and Yates decompositions of the Brier score. These decompositions provide finer detail about forecast attributes, as compared to the mean Brier score, but they also have some inherent limitations. For example, the Murphy decomposition requires rounded ("binned") forecasts, so that the raw data must often be modified. The Murphy decomposition also yields biased estimates of the corresponding large-sample component values (Ferro & Fricker, 2012). More generally, it is often difficult to interpret the magnitudes of individual component values or differences between values. We provide some further detail on these issues below.

2.1 Murphy components

The three Murphy Brier components are uncertainty, calibration, and discrimination. As Murphy (1973) discusses, the uncertainty term reflects the mean Brier score that would be obtained by a base rate judge who reports the base rate of each alternative's occurrence across all resolved questions. This component ranges from 0 to $(M^*-1)/M^*$ (where M^* is the number of alternatives per question), with its value increasing as the base rates become equal across all alternatives. Larger values are not necessarily worse, because they reflect the degree of baseline uncertainty in the forecasting environment, as opposed to the forecaster. However, larger values imply that there is more "room" for forecasters to improve upon a base rate judge.

The second component, calibration, ranges from 0 to 2. The minimum of 0 is achieved when each unique forecast matches the base rate of event occurrences, whereas the maximum of 2 is achieved when the forecaster always assigns a probability of 1 to events that never occur. Because smaller values are better, this component might be more profitably labeled "*mis*calibration" instead of "calibration." Murphy (1973) calls this term "reliability," because it reflects the extent to which the reported probabilities match the relative frequencies of the corresponding alternatives' occurrences.

The final component, discrimination, ranges from 0 to $(M^* - 1)/M^*$, with larger values being better. This component indicates the extent to which forecasts differ for alternatives that occur, as opposed to alternatives that do not occur. In most cases, this metric gauges the extent to which the forecaster reports larger probabilities for alternatives that occur, as compared to alternatives that do not occur.

The Murphy calibration and discrimination terms are correlated, in the sense that better calibration is related to worse discrimination and vice versa (also see Yates, 1982). Further, if two forecasters have the same average Brier score, it is impossible for one of the forecasters to be better than the other on both calibration and discrimination. These results highlight the difficulty of being simultaneously good on both calibration and discrimination.

2.2 Yates components

The focal Yates components include the forecast-outcome covariance, calibration-in-the-large, and excess forecast variance (ΔVar_f). Additional components include uncertainty (which also appears in the Murphy decomposition) and minimum conditional variance. These latter components reflect properties of the forecasting environment, as opposed to the forecaster, so we attend to them less than the other components.

The forecast-outcome covariance metric is somewhat similar to the Murphy discrimination metric, measuring the extent to which forecasts are related to outcome occurrence. This covariance ranges from 0 to $(M^* - 1)/M$, with larger values being better.

The calibration-in-the-large metric is a simple assessment of bias, measuring the squared difference between the average forecast and the base rate of outcomes. The minimum value of 0 is best, with the maximum value of 2 achieved when the forecaster always assigns a forecast of 1 to an outcome that never occurs and a forecast of 0 to an outcome that always occurs. This metric will often be close to zero, because it is measuring squared differences between average probabilities. For example, for binary questions, say that one outcome occurs 40% of the time. If a

https://doi.org/10.1017/S1930297500007099 Published online by Cambridge University Press

forecaster's average probability for this outcome is .5, then his/her calibration-in-the-large value is only 0.02.

Finally, the "excess forecast variance" metric measures the extent to which variability in the observed forecasts is due to noise. Yates points out that a forecaster can improve his/her Brier score by reducing the variance in his/her reported forecasts, but some variance in reported forecasts is necessary due to the "signal" in each question. Thus, Yates derived the minimum variance in the forecasts necessary to obtain the observed forecast-outcome covariance and the observed base rate of outcomes. This minimum variance can be considered the "variability in forecasts due to the question signal," somewhat similarly to the interpretation of R^2 in linear regression contexts. The "excess" variance is then the extent to which the observed variance in forecasts exceeds the minimum necessary variance. Thus, values closer to the minimum of 0 are better, and values closer to the total forecast variance are worse. Because this value explicitly depends on other aspects of the forecasts (in particular, the covariance between forecasts and outcomes), it appears difficult to compare this metric across forecasters.

3 Generalizing the Brier decompositions

Now that we have reviewed the traditional decompositions, we discuss our generalizations to weighted average Brier scores. These generalizations can be used in situations where multiple forecasters report probabilities on the same sets of questions at the same times, with the questions potentially having different numbers of alternatives (possibly ordinal) and different labels for the alternatives. The individual Brier scores can be weighted unequally, which potentially allows for the use of questions whose alternatives have cardinal categories (ranges of continuous values) or questions whose alternatives are conditional on intermediate outcomes (where, e.g., some alternatives may be eliminated during the life of the question). The generalizations described here cannot immediately handle missing data, where forecasters report on different sets of questions, or where forecasters decide when to report on a question. Analysts could potentially impute missing forecasts prior to using the methods described here, or they could turn to model-based methods for evaluating forecasters in the presence of missing forecasts (see, e.g., Merkle et al., 2017).

To handle weighted Brier scores (including ordinal Brier scores), we derive weighted versions of the Murphy and Yates decompositions below, which can be viewed as generalizations of Young (2010). To address diverse questions with different alternatives, we propose averaging the Brier decompositions over the possible ways that alternatives could be grouped together across questions. This averaging is accomplished via a resampling algorithm and is described after the weighted decompositions. For concreteness, we refer to the IARPA ACE example throughout, detailing the use of weighted Brier scores in that tournament.

3.1 ACE Brier scores

In the IARPA ACE program, it was important to provide a simple metric characterizing overall accuracy across the heterogeneous set of forecast questions that defined the tournament. Because competitors were required to submit daily forecast updates on each question, the adopted approach was as follows: for each forecast question, calculate a daily Brier score for each day the question is active; average those daily scores to produce a per-question mean daily Brier score; finally, take the average of those per-question mean daily Brier scores to produce a weighted overall average Brier score. This two-step averaging process was used so that each question counted equally towards the overall Brier score. If we instead averaged all daily Brier scores across all questions in a single step, then questions that were active for longer time periods would be most influential on the overall Brier score. The two-step averaging is thus a weighted average of daily Brier scores, because each daily score's weight depends on the number of days that the associated question was active.

Formally, say that there are J resolved questions, with each question being open for n_j days and having M_j alternatives ($j = 1, \ldots, J$). Then, for a given forecasting system (competitor), we can represent a "Mean Daily Error" (MDE) for question j as

$$\text{MDE}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} \sum_{m=1}^{M_j} (f_{ijm} - d_{jm})^2,$$

where f_{ijm} is the system's probability for alternative m on day i of question j, and d_{jm} is the outcome of alternative m on question j (1 if realized, 0 otherwise). We can then represent a "Mean of Mean Daily Errors" (MMDE) across questions as

$$\text{MMDE} = \frac{1}{J} \sum_{j=1}^{J} \frac{1}{n_j} \sum_{i=1}^{n_j} \sum_{m=1}^{M_j} (f_{ijm} - d_{jm})^2. \qquad (1)$$

The above summation shows that each question's MDE is weighted equally, which means that each daily forecast is weighted *unequally*: daily forecasts associated with a long-term question receive low weight, as compared to daily forecasts associated with a short-term question. Although it has the advantage of simplicity (as compared to model-based alternatives such as, e.g., Bo et al., 2017; Budescu and Chen, 2015; Budescu and Johnson, 2011; Merkle et al., 2016; Satop 2014a, b; Steyvers et al., 2014), this high-level Brier score metric is not readily amenable to traditional Brier decompositions.

https://doi.org/10.1017/S1930297500007099 Published online by Cambridge University Press

Along with the fact that Brier scores are weighted, Equation (1) allows each question to have a different number of alternatives (i.e., question j has M_j alternatives), as was needed in the ACE tournament. However, to use Brier decompositions, we require that all questions have the same number of alternatives. To fulfill this requirement, we can add "phantom alternatives" to questions that do not have the maximum possible number of alternatives. Specifically, let $M^* = \max(M_j)$, $j = 1, \ldots, J$. For each question j with fewer than M^* alternatives, we create new alternatives that increase the number to M^*. These new alternatives always receive forecasts/probabilities of 0, and the new alternatives' outcomes are also coded as 0. These new alternatives have no influence on the Brier score or on MMDE, though they do influence the base rates that go into the uncertainty component of the Brier decompositions. However, because the phantom alternatives are the same across all competitors (we require all systems to respond to the same questions), the uncertainty term will also be the same across all forecasting systems. Individual differences between systems will stem from differences in other Brier components.

From a technical perspective, the phantom alternatives allow us to remove the j subscript from the upper bound of the third summation:

$$\text{MMDE} = \frac{1}{J}\sum_{j=1}^{J}\frac{1}{n_j}\sum_{i=1}^{n_j}\sum_{m=1}^{M^*}(f_{ijm} - d_{jm})^2. \qquad (2)$$

These phantom alternatives can be generally applied to questions with differing numbers of alternatives, though they may introduce excessive noise if there is a large imbalance in the number of alternatives. For example, if a question with 10 alternatives is added to a set of questions with 2 alternatives, then each two-alternative question requires four times as many phantom alternatives as real alternatives. This is less problematic for the ACE data considered here, where questions vary from two to five alternatives.

3.2 Weighted Brier score decomposition

Below, we describe analytic results for a Murphy decomposition of weighted Brier scores, and we also extend the results to the Yates decomposition. These results yield Murphy and Yates components that are specifically tied to MMDE or to other weighted Brier scores.

Just like the original Murphy decomposition, the Murphy decomposition described here requires that forecast probabilities be grouped into a limited number of discrete bins of rounded forecast probability values. For example, probabilities between 0 and .1 might be recoded and grouped as .05, probabilities between .1 and .2 might be recoded as .15, and so on, resulting in 10 discrete probability ranges, from .05 to .95. Although straightforward for binary forecast questions, multinomial questions complicate matters, and we provide further detail on possible approaches in Appendix B (along with sensitivity analyses in the Example section). For now, we simply assume that exhaustive subsets of probability bins have already been defined for Brier score decomposition purposes. We also apply the Yates decomposition to these discrete bins for uniformity, though discretization is not required for the Yates decomposition.

The theorem below assumes that we are computing a weighted average Brier score across N forecasts. Its application to ACE is facilitated by the fact that Equation (2) can be rewritten as

$$\begin{aligned}\text{MMDE} &= \sum_{j=1}^{J}\sum_{i=1}^{n_j}\frac{1}{J \times n_j}\sum_{m=1}^{M^*}(f_{ijm} - d_{jm})^2 \quad (3)\\ &= \sum_{\ell=1}^{N} w_\ell \sum_{m=1}^{M^*}(f_{\ell m} - d_{\ell m})^2, \quad (4)\end{aligned}$$

where $N = \sum_{j=1}^{J} n_j$ is the total number forecasts reported, across all questions. As verbally described earlier, the above equation shows that the MMDE is a sum of weighted Brier scores across all questions and days, where a particular forecast's weight w_ℓ is based on the number of days that its corresponding question was open (n_j) as well as the number of questions J. Note also that, based on the way that MMDE is defined, these weights will always sum to 1, i.e.,

$$\sum_{\ell=1}^{N} w_\ell = 1.$$

We now proceed with the Murphy decomposition theorem.

Theorem. (Weighted Murphy decomposition) *Assume that a forecasting system reports N forecasts, each of which has M^* alternatives and each of which has a particular weight (with the weights across questions summing to 1, corresponding to a weighted average). Further assume that the reported forecasts have been grouped into K subsets, so that all forecasts within subset k ($k = 1, \ldots, K$) are given the same value (f_{k1} $f_{k2} \ldots f_{kM^*}$). Then the forecasting system's weighted average Brier score can be written as*

$$\begin{aligned}\overline{Brier_w} &= \sum_{m=1}^{M^*}\overline{d}_m(1-\overline{d}_m) + \sum_{k=1}^{K} w_k^* \sum_{m=1}^{M^*}(f_{km} - \overline{d}_{km})^2\\ &\quad - \sum_{k=1}^{K} w_k^* \sum_{m=1}^{M^*}(\overline{d}_{km} - \overline{d}_m)^2 \quad (5)\\ &= Unc + Miscalib - Discrim, \quad (6)\end{aligned}$$

https://doi.org/10.1017/S1930297500007099 Published online by Cambridge University Press

where

$$w_k^* = \sum_{\ell \in bin\ k} w_\ell \quad (7)$$

$$\overline{d}_{km} = \frac{1}{w_k^*} \sum_{\ell \in bin\ k} w_\ell d_{\ell m} \quad (8)$$

$$\overline{d}_m = \frac{1}{N} \sum_{j=1}^{N} d_{jm}. \quad (9)$$

For a proof, see Appendix A.

The first term of (5) corresponds to uncertainty, the second term corresponds to (mis)calibration, and the third term corresponds to discrimination. This decomposition is similar to that of Young (2010), except that we explicitly consider questions with more than two alternatives (which, as described later, allows for inclusion of a Brier score for ordered alternatives). We also extend the results to the Yates decomposition, presented below.

Lemma. (Weighted Yates decomposition) *Under the conditions set forth above, the forecasting system's weighted average Brier score can also be written as*

$$\begin{aligned} \overline{Brier_w} &= \sum_{m=1}^{M^*} \overline{d}_m(1-\overline{d}_m) + \sum_{m=1}^{M^*}\sum_{\ell=1}^{N} w_\ell (f_{\ell m} - \overline{f}_m)^2 \\ &\quad + \sum_{m=1}^{M^*} (\overline{f}_m - \overline{d}_m)^2 \\ &\quad - 2\sum_{m=1}^{M^*}\sum_{\ell=1}^{N} w_\ell (f_{\ell m} - \overline{f}_m)(d_{\ell m} - \overline{d}_m) \quad (10) \\ &= \textit{Uncertainty} + \textit{Var}_f + \\ &\quad \textit{Miscal-in-the-large} - 2\textit{Cov}_{fd}. \quad (11) \end{aligned}$$

where

$$\overline{f}_m = \sum_{\ell=1}^{N} w_\ell f_{\ell m}. \quad (12)$$

Further detail on this derivation also appears in Appendix A.

Equation (11) shows the specific components, which mirror the traditional Yates components. We first see the weighted variance of outcomes (Var_d; using weights associated with each daily forecast) and the weighted variance of forecasts (Var_f). The third term is "calibration-in-the-large," the gross measure of miscalibration. We dub it "miscalibration-in-the-large" throughout this paper. Finally, the fourth term gets at discrimination, assessing the extent to which forecasts and outcomes covary with one another.

As mentioned earlier, we can further decompose Var_f into two parts: (i) the minimum possible variance of forecasts, given the observed outcome base rate and covariance between forecasts and outcomes, and (ii) the excess ("noise") variance of forecasts, over and above the minimum. This additional decomposition can be represented by

$$\text{Var}_f = \text{MinVar}_f + \Delta\text{Var}_f, \quad (13)$$

where

$$\text{MinVar}_f = \sum_{m=1}^{M^*} (\overline{f}_{1m} - \overline{f}_{0m})^2 \overline{d}_m (1-\overline{d}_m) \quad (14)$$

$$\Delta\text{Var}_f = \text{Var}_f - \text{MinVar}_f, \quad (15)$$

and $\overline{f}_{1m}$ and $\overline{f}_{0m}$ are the weighted average forecasts when alternative m occurs and does not occur, respectively. The ΔVar_f term can be an informative performance metric, where better forecasters have values closer to 0 (i.e., better forecasters have less excess variance).

A summary of equations for the various weighted Brier components appears in Table 1, where the equations make use of terms from Equations (7) to (9) along with Equation (12). In the sections below, we describe how these components can be further extended to diverse questions.

3.3 Inconsistent alternative labels

The above results require us to compute weighted base rates for each outcome alternative, a computation that implicitly assumes consistently interpretable outcome alternatives over all questions—for example, that all forecast questions concern daily chances of rain or monthly chances of regime change. This requirement does not hold in the IARPA ACE tournament or elsewhere, where the set of question topics and outcome alternatives under consideration is highly heterogeneous. For example, it is intuitive to compute a base rate of rainy days or of months with military conflict over a set of questions focused on a single topic. However, it is less meaningful to compute a base rate of occurrences of "alternative A," where this is simply an arbitrary label for "the first listed outcome alternative for each question," and where the underlying forecast question topics may range from stock index values to disease counts to occurrence or non-occurrence of military events, and so on. Across diverse forecast questions, such a set of "alternative A" outcomes will not share a substantive interpretation to differentiate them from the corresponding "B" and "C" outcome alternatives. Further, depending on exactly which alternatives count as "A" (and as "B," "C," etc.), we will obtain different calibration and discrimination values because each alternative's base rate will change. This is problematic for drawing substantive conclusions about forecasting system performance.

One partial solution here involves coding alternatives in terms of whether or not a "status quo" alternative is maintained (see, e.g., Turner et al., 2014). That is, instead of maintaining the original alternatives, each question could be

https://doi.org/10.1017/S1930297500007099 Published online by Cambridge University Press

Table 1: Weighted Brier components and their expressions.

Component	Expression
Miscalibration	$\sum_{k=1}^{K} w_k^* \sum_{m=1}^{M^*} (f_{km} - \overline{d}_{km})^2$
Discrimination	$\sum_{k=1}^{K} w_k^* \sum_{m=1}^{M^*} (\overline{d}_{km} - \overline{d}_m)^2$
ΔVar_f	$\sum_{m=1}^{M^*} \sum_{\ell=1}^{N} w_\ell (f_{\ell m} - \overline{f}_m)^2 - \sum_{m=1}^{M^*} (\overline{f}_{1m} - \overline{f}_{0m})^2 \overline{d}_m (1 - \overline{d}_m)$
Miscalibration-in-the-large	$\sum_{m=1}^{M^*} (\overline{f}_m - \overline{d}_m)^2$
Cov_{fd}	$\sum_{m=1}^{M^*} \sum_{\ell=1}^{N} w_\ell (f_{\ell m} - \overline{f}_m)(d_{\ell m} - \overline{d}_m)$
Uncertainty	$\sum_{m=1}^{M^*} \overline{d}_m (1 - \overline{d}_m)$

converted to having the two alternatives of "status quo maintained" and "status quo overturned." For example, when forecasting whether or not a conflict will occur between two countries, the status quo is "conflict does not occur." Similarly, in forecasting rainy days, the status quo is typically "no rain" (though this may depend on the geographical location). While helpful for event-oriented questions where one outcome can be framed as the status quo, this approach is not a complete solution to question diversity in applied forecast evaluation. For example, questions whose categories concern continuous quantities do not readily conform to any obvious status quo interpretation, so they would ostensibly need to be dropped from such analyses. To derive discrimination and calibration values for *all forecast questions* that contribute to a system MMDE score, we require a more general decomposition approach.

Our solution here involves averaging over all possible alternative orderings via resampling. At each iteration of the resampling algorithm, we reorder each question's alternatives and apply the weighted Brier decompositions to obtain the Murphy and Yates components. This is similar to randomly choosing a single alternative from each question, then decomposing the forecasts from these randomly-chosen alternatives. Each system's overall component values are then the averages across many iterations. This resampling algorithm maintains the same daily Brier scores across all iterations, allowing us to examine Brier components across the inconsistent alternative labels while holding the Brier score constant. With a "large enough" number of resamples (see sensitivity analyses below), all possible orderings of question alternatives are given equal representation, ensuring that the Brier components converge towards stable values.

These ideas are conceptualized in the following algorithm:

- Loop for B iterations:
  - Randomly determine an ordering of alternatives for each question.
  - Compute the weighted Brier decompositions under the ordering from the first step.
- Compute the average component values over the component distributions produced by the B iterations.

The reordering step could be modified to handle "status quo" questions so that, e.g., all "status quo overturned" alternatives count towards the base rate for one specific alternative. This effectively removes noise from systems' overall component values, leading to fewer possibilities for how alternatives can be ordered. However, the resampling approach also works without status quo information, and it is a judgment call as to whether or not we should use status quo coding. In the example below, we do not use status quo information.

3.4 Ordinal questions

So far, we have derived weighted Brier score decompositions and described a resampling algorithm for handling questions with diverse alternatives. We now handle questions with ordered alternatives. To address the fact that some, but not all, forecast questions in a corpus will include meaningfully ordered outcome alternatives, we can compute an ordered variant of the Brier score involving *cumulative* forecasts (see Jose, Nau and Winkler, 2009). This allows us

https://doi.org/10.1017/S1930297500007099 Published online by Cambridge University Press

to handle ordinal questions using a scoring rule that is very similar to the usual Brier score.

For a specific question j with M_j alternatives, the ordered Brier score is of the form:

$$\frac{1}{M_j - 1} \sum_{m=1}^{M_j - 1} (F_{jm} - D_{jm})^2,$$

where F_{jm} is the cumulative forecast (probability assigned to alternative m and below) and D_{jm} is the "cumulative" outcome (equals 1 if alternative m or below was realized, 0 otherwise) associated with alternative m of question j.

To include these ordinal Brier scores in the decomposition, we treat each of the $(M_j - 1)$ terms in the sum above as new Brier scores. For each of these new Brier scores, we add phantom alternatives to the binary forecasts F_{jm} and outcomes D_{jm}. These cumulative forecast alternatives are reordered in the same way across all ordinal questions, in order to maintain the "partial credit" aspect of the ordered Brier score. The weight for each of these Brier scores is then $1/(J \times n_j \times (M_j - 1))$, with these weights being used in the decomposition. The weights are constructed so that, each ordinal question receives equal weight, as compared to each unordered question.

In practical terms, the ordinal Brier score rewards forecasts where the probability mass is more heavily concentrated around the true outcome alternative. For example, two systems may assign the same probability to the true outcome of a set of questions with ordered alternatives. If one of these systems consistently assigns more of its probability to the alternatives immediately adjacent to the true outcome alternative, then this latter system will generally be identified as having the better Brier components (potentially across all components that the system can influence).

3.5 Summary

In this section, we first derived general Murphy and Yates decompositions for weighted average Brier scores, of which the ACE MMDE is a special case. We then proposed a resampling algorithm for handling heterogeneous questions, whereby the decomposition is applied across multiple reorderings of question alternatives. We showed how ordinal Brier scores could be included in this procedure, and we also addressed two remaining issues: the inclusion of phantom alternatives when not all questions have the same number of alternatives, and the creation of forecast bins for the Murphy decomposition (with further bin detail in Appendix B). We now turn to an application.

4 Example

To illustrate these procedures, we evaluate forecasts from four competing systems in the IARPA ACE tournament. We show how the decompositions provide finer-grained information for system comparison (as opposed to the overall MMDE metric), as well as information about the metrics' uncertainty.

4.1 Method

In this dataset, each of the four forecasting systems reported daily forecasts on each of 76 questions, with questions being open an average of 118 days (leading to 8,968 unique forecasts for each of the four systems). About 70% of the questions had two alternatives, 21% had four alternatives, and the remainder had either three or five alternatives. Additionally, 21% of the questions had ordered alternatives.

We ran the resampling procedure for 100 iterations, where forecasts were binned by rounding to the nearest multiple of .1 (with "round the lowest" handling of multi-category questions, to ensure that each forecast summed to 1; see Appendix B). At each iteration of resampling, each question's alternatives were rearranged and weighted Brier decompositions calculated (with weights corresponding to the MMDE metric described previously). The average components and associated uncertainty intervals were then computed based on the resamples.

4.2 Results

The four systems' original MMDE values were 0.353, 0.320, 0.333, and 0.349. Because lower Brier scores are better than higher Brier scores, we might assign System 2 as the best. However, this overall score does not provide information about specific aspects of performance for which System 2 excelled or exhibited mediocrity; this is where the proposed decompositions become helpful. Table 2 shows the decompositions for the four systems, including the Murphy discrimination and miscalibration metrics; the Yates excess variance ($\Delta\text{Var}_\text{f}$), miscalibration-in-the-large, and covariance metrics; and the uncertainty metric that appears in both decompositions. Finally, the last two rows show the systems' MMDEs, both before and after binning the forecasts.

Examining Table 2, we can see that the rounding had little impact on each system's MMDE, because the two MMDE rows are similar to one another. We further see, e.g., that System 1 has the best (largest) discrimination and lowest excess variance, but it also has the worst miscalibration, covariance, and MMDE. System 2, which had the best MMDE, is not the best on all the Brier components, but it continues to look good: it has the best MMDE, its discrimination is close to that of System 1, and its remaining components are generally close to the best.

To reinforce these results and further compare pairs of systems to one another, we can construct interval estimates of component differences across the resamples. We focus

https://doi.org/10.1017/S1930297500007099 Published online by Cambridge University Press

Table 2: Brier components for four ACE forecasting systems.

	System 1	System 2	System 3	System 4
Discrimination	0.607	0.597	0.573	0.553
Miscalibration	0.198	0.149	0.138	0.136
$\Delta\text{Var}_\text{f}$	0.161	0.199	0.237	0.205
Miscalibration-in-the-large	0.006	0.004	0.004	0.004
Cov_{fd}	0.387	0.472	0.508	0.442
Uncertainty	0.769	0.769	0.769	0.769
MMDE (binned)	0.359	0.321	0.334	0.351
MMDE	0.353	0.32	0.333	0.349

on comparing System 2 to System 3 here, because these two systems had the best MMDE values. The intervals are shown in Table 3, along with the intervals for individual System 2 and System 3 components. While there is overlap between some individual system components (seen in the "System 2" and "System 3" columns), these components are correlated across resamples. Thus, when we examine the intervals of component differences, we may still observe intervals that fail to overlap with zero. In this case, we observe that the intervals for discrimination, excess variance ($\Delta\text{Var}_\text{f}$), and the covariance metric fail to overlap with zero. The discrimination and excess variance metrics favor System 2, whereas the covariance metric favors System 3. While these results generally support our favoring of System 2, they also suggest that System 3 is not too far behind System 2: the discrimination interval nearly overlaps with zero, and System 3's reported forecasts have a stronger covariance with the question outcomes.

4.3 Sensitivity Analyses

Finally, we examine how the number of resamples and the binning procedure influence results. These both represent subjective judgments on the part of the researcher, so it is worthwhile to study the sensitivity of our results to these judgments. We study four possible binning procedures: bin resolutions (forecast rounding) of .05 or .1, crossed with two methods for ensuring that probabilities sum to 1 (modifying the smallest forecast or modifying the furthest forecast, as discussed in Appendix B). For each binning procedure, we obtain 50,000 resamples. We then examine how point estimates vary under each of the four binning procedures, as well as how subsets of the resamples (50, 100, or 500 at a time) vary within any one binning procedure: while larger numbers of resamples are obviously preferable, computation time is an issue for the resampling algorithm, with each iteration taking about 2 seconds on our computers. We focus on System 2 for simplicity, though we also examine differences between Systems 2 and 3 (the top two systems).

Figure 1 shows how the estimated means of System 2 discrimination and miscalibration vary across the factors from the previous paragraph. It is seen for both metrics that, while increased numbers of resamples lead to less variability in the estimates, the "50 resample" results generally agree with the "500 resample" results to two decimal places. Additionally, bin resolutions of .1 lead to lower discrimination and miscalibration metrics as compared to bin resolutions of .05, with the differences being about .02 points in each case. Finally, rounding strategies had a smaller influence on the results, with differences of less than .005 on both metrics. It is unlikely that any of these differences are large enough to influence substantive conclusions, at least for the data examined here (and other datasets may well lead to larger differences).

Figure 2 reinforces these results, illustrating how the mean "System 2 vs System 3" differences in discrimination and miscalibration vary across the same factors. We see that the differences are nearly always greater than 0, with the only occasional exceptions occurring on miscalibration at 50 resamples. This suggests that we would nearly always observe the System 2 means as being larger than the System 3 means, regardless of the specific factors chosen. Interestingly, we observe larger system differences for bin resolutions of .1 as compared to bin resolutions of .05. This suggests that coarse bins impact systems differentially, where systems with worse Brier scores are potentially penalized more heavily than systems with better Brier scores. It is additionally clear that, if computation time is not an issue, 500 resamples provides more precision than smaller numbers of resamples. While none of the differences are large (differences of less than .01 on both metrics), the differences will vary by the characteristics of each dataset. This warrants similar sensitivity analyses when the methods are applied to other datasets. R package *scoring* (Merkle and Steyvers, 2013) is intended to ease such analyses, and the accompanying replication code illustrates how this can be accomplished.

https://doi.org/10.1017/S1930297500007099 Published online by Cambridge University Press

Figure 1: Variability in the estimated mean of System 2's Murphy Brier components (discrimination and miscalibration), by number of resamples (50, 100, or 500), bin resolution (.05 or .1), and strategy for ensuring that the rounded forecasts sum to 1 (round the lowest value or round the farthest value).

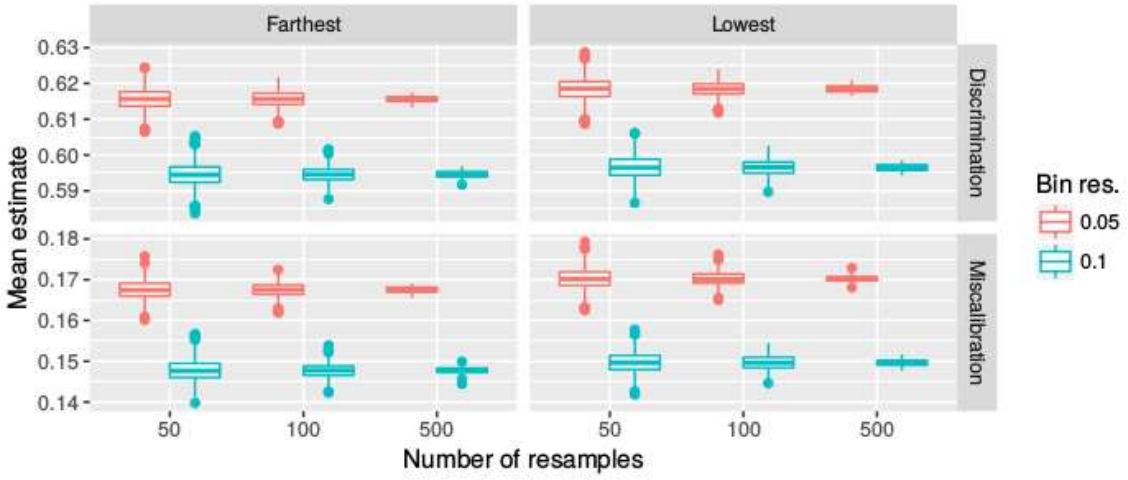
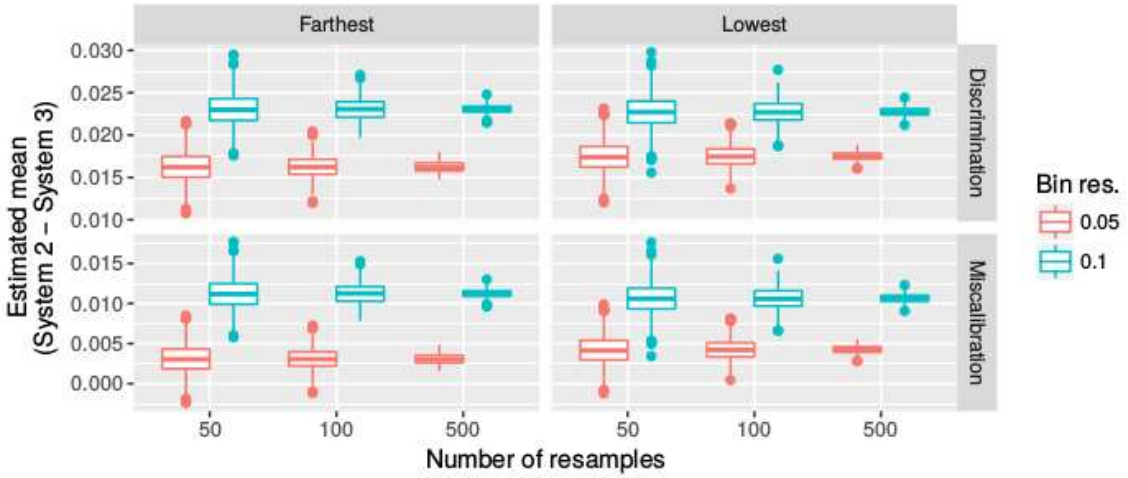


Figure 2: Variability in the estimated mean of Brier component differences between Systems 2 and 3, by number of resamples (50, 100, or 500), bin resolution (.05 or .1), and strategy for ensuring that the rounded forecasts sum to 1 (round the lowest value or round the farthest value).



https://doi.org/10.1017/S1930297500007099 Published online by Cambridge University Press

Table 3: 90% interval estimates of differences between components and of individual components, Systems 2 and 3. Differences are taken as System 2 minus System 3.

	Difference	System 2	System 3
Discrimination	$(0.0034, 0.0503)$	$(0.5638, 0.6348)$	$(0.5459, 0.6056)$
Miscalibration	$(-0.0090, 0.0379)$	$(0.1237, 0.1770)$	$(0.1154, 0.1646)$
$\Delta\text{Var}_\text{f}$	$(-0.0429, -0.0330)$	$(0.1918, 0.2048)$	$(0.2295, 0.2415)$
Miscalibration-in-the-large	$(-0.0041, 0.0032)$	$(0.0007, 0.0081)$	$(0.0010, 0.0089)$
Cov_{fd}	$(-0.0382, -0.0322)$	$(0.4555, 0.4872)$	$(0.4915, 0.5229)$

5 Discussion

The methods proposed in this paper broaden Brier score decompositions, making them applicable to real data for which the original decompositions were not applicable. We extended the original decompositions to handle weighted average Brier scores, of which the Brier metric used in the IARPA ACE tournament is a special case. The weighted Brier decompositions are presented in Equations (5) and (10) and summarized in Table 1; these decompositions work for any weighted average, so they are generally useful to researchers who may wish to use weighted average Brier scores. To handle differing numbers of alternatives for different forecast questions, we added "phantom" alternatives to questions with fewer than the maximum number of alternatives. These phantom alternatives have no impact on the Brier score; they are simply tools to facilitate the decompositions of the Brier score.

Beyond phantom alternatives, we handle inconsistently-labeled alternatives via a resampling algorithm. At each iteration of the algorithm, we reorder each question's outcome alternatives and compute the weighted Brier decomposition. This approach allows us to examine calibration and discrimination metrics across inconsistently-labeled alternatives while holding the Brier score constant. Each system's overall metrics are its stabilized averages across many iterations. Finally, where applicable, we directly incorporate ordinal outcome information in the decomposition by making use of the fact that the ordered Brier score is a series of binary Brier scores applied to cumulative forecasts.

It is important to highlight that, in generalizing the Brier score decompositions in these ways, we are generalizing the very definitions of "calibration" and "discrimination." In traditional Brier decompositions, these concepts are defined directly with respect to a specific substantive event class of interest. For example, in forecasting military conflicts, calibration involves the extent to which a forecaster's reported "conflict" probabilities reflect the proportion of realized conflicts, and discrimination involves the extent to which forecasted conflict probabilities differ across realized "conflict" cases vs. "no conflict" cases. Finally, the base rate refers specifically to the proportion of realized conflicts over all analyzed opportunities for such conflict (i.e., the relative frequency of "positive" cases among all cases). In contrast, when we attempt to characterize more general calibration and discrimination properties of forecasts over heterogeneous question sets, our conceptual definitions also take on a more general, topic-agnostic interpretation. A rough verbal description of the calibration and discrimination metrics we have proposed is as follows:

- Being well-calibrated means that, if we were to randomly select a single alternative from each question, the probabilities assigned to the selected alternatives would match the proportion of those alternatives that are realized.
- Being highly discriminating means that, if we were to randomly select a single alternative from each question, the probabilities assigned to realized alternatives would consistently differ from (be larger than) the probabilities assigned to unrealized alternatives.

The base rate, then, is artificial and is defined as the proportion of randomly selected alternatives that were realized. For example, if we have only binary questions, then this base rate will tend toward 0.5, because each question has two outcomes, one of which is realized and one of which is not. Although it lacks a substantive interpretation, this artificial base rate is held constant across forecasters, because all forecasters are being evaluated against a common set of forecast questions. Additionally, if some subsets of questions all have the same alternatives, we can elect to specially treat those questions so that the base rate of that subset is preserved.

Future work may extend these developments beyond the Brier score, though computation will be more difficult because closed-form, analytic results are unlikely to be available. Some useful prior work here includes Bröcker (2009), who presents results related to decompositions of general proper scoring rules, and Gneiting, Balabdaoui and Raftery (2007), who discuss alternative metrics that are related to the traditional decompositions.

While the developments proposed here all helped us compute Brier decompositions in the IARPA ACE data, each development may also be useful on its own. For example, if

https://doi.org/10.1017/S1930297500007099 Published online by Cambridge University Press

one were computing a weighted average Brier score across homogeneous questions, the analytic derivations here could be used without resampling. Conversely, if one were computing a simple average Brier score across heterogeneous questions, the resampling methods described here could be used without weights. As a result, the methods should be generally useful across a variety of datasets. They aim at providing a principled compromise between simple metrics that cannot accommodate applied data and complex models that may require considerable expertise for their use.

Computational Details

All results were obtained using the R system for statistical computing (R Core Team, 2017), version 3.4.3, with the add-on package *scoring* version 0.6 for Brier decomposition metrics. Data and code for replication of our results is available at https://osf.io/jvhwp/.

6 References

Bo, Y. E., Budescu, D. V., Lewis, C., Tetlock, P. E., and Mellers, B. (2017). An IRT forecasting model: Linking proper scoring rules to item response theory. *Judgment and Decision Making, 12*(2), 90–103.

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

Bröcker, J. (2009). Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society*, 135(643), 1512–1519.

Budescu, D. V. and Chen, E. (2015). Identifying expertise to extract the wisdom of crowds. *Management Science, 61*, 267–280.

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

Carvalho, A. (2016). An overview of applications of proper scoring rules. *Decision Analysis, 13*, 233–242.

Ferro, C. A. T. and Fricker, T. E. (2012). A bias-corrected decomposition of the Brier score. *Quarterly Journal of the Royal Meteorological Society*, 138, 1954–1960.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69*(2), 243–268.

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

Jose, V. R. R., Nau, R. F., and Winkler, R. L. (2009). Sensitivity to distance and baseline distributions in forecast evaluation. *Management Science, 55*(4), 582–590.

Mandel, D. R. and Barnes, A. (2014). Accuracy of forecasts in strategic intelligence. *Proceedings of the National Academy of Sciences*, 111(30), 10984–10989.

Merkle, E. C. and Steyvers, M. (2013). Choosing a strictly proper scoring rule. *Decision Analysis, 10*, 292–304.

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

Merkle, E. C., Steyvers, M., Mellers, B., and Tetlock, P. E. (2017). A neglected dimension of good forecasting judgment: The questions we choose also matter. *International Journal of Forecasting, 33*, 817–832.

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology, 12*(4), 595–600.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Satopää, V. A., Baron, J., Foster, D. P., Mellers, B. A., Tetlock, P. E., and Ungar, L. H. (2014a). Combining multiple probability predictions using a simple logit model. *International Journal of Forecasting, 30*, 344–356.

Satopää, V. A., Jensen, S. T., Mellers, B. A., Tetlock, P. E., and Ungar, L. H. (2014b). Probability aggregation in time-series: Dynamic hierarchical modeling of sparse expert beliefs. *The Annals of Applied Statistics, 8*, 1256–1280.

Steyvers, M., Wallsten, T. S., Merkle, E. C., and Turner, B. M. (2014). Evaluating probabilistic forecasts with bayesian signal detection models. *Risk Analysis, 34*(3), 435–452.

Tetlock, P. E. (2005). *Expert political judgment: How good is it? How can we know?* Princeton University Press.

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

Turner, B. M., Steyvers, M., Merkle, E. C., Budescu, D. V., and Wallsten, T. S. (2014). Forecast aggregation via recalibration. *Machine Learning, 95*(3), 261–289.

Winkler, R. L. and Jose, V. R. R. (2010). Scoring rules. In Cochran, J. J., editor, *Wiley Encyclopedia of Operations Research and Management Sciences*. New York: John Wiley & Sons.

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

Young, R. M. B. (2010). Decomposition of the Brier score for weighted forecast-verification pairs. *Quarterly Journal of the Royal Meteorological Society, 136*, 1364–1370.

A Proofs of weighted Brier score decompositions

In this appendix, we prove the theorem about the Murphy decomposition of the weighted mean Brier score and the lemma about the Yates decomposition. We focus on the decomposition for a single alternative (corresponding to, e.g., the $(0, 1)$ Brier score for a two-alternative question), with immediate generalization to multiple alternatives.

Focusing on the Murphy decomposition theorem, we first write and expand MMDE, where J is the number of questions, n_j is the number of forecasts for question j, and N is the total number of reported forecasts across questions:

$$\begin{aligned}
\text{MMDE} &= \sum_{j=1}^{J}\sum_{i=1}^{n_j}\frac{1}{J\times n_j}(f_{ij}-d_j)^2 \\
&= \sum_{\ell=1}^{N} w_\ell (f_\ell - d_\ell)^2 \\
&= \sum_{\ell} w_\ell f_\ell^2 - 2\sum_{\ell} w_\ell f_\ell d_\ell + \sum_{\ell} w_\ell d_\ell^2 \\
&= \sum_{\ell} w_\ell f_\ell^2 - 2\sum_{\ell} w_\ell f_\ell d_\ell + \sum_{\ell} w_\ell d_\ell,
\end{aligned}$$

with the last equality following because d_ℓ can only equal 0 or 1, so squaring it does not matter.

Now we make use of the fact that forecast probabilities are binned into intervals, so that all the f_ℓ in each bin are equal to f_k. This allows us to expand MMDE as below:

$$\begin{aligned}
&\sum_{k=1}^{K}\left(f_k^2\sum_{\ell\in\text{bin k}} w_\ell - 2f_k\sum_{\ell\in\text{bin k}} w_\ell d_\ell + \sum_{\ell\in\text{bin k}} w_\ell d_\ell\right) \\
&= \sum_{k=1}^{K}\left\{\left(f_k\sqrt{\sum_{\ell\in\text{bin k}} w_\ell} - \frac{\sum_{\ell\in\text{bin k}} w_\ell d_\ell}{\sqrt{\sum_{\ell\in\text{bin k}} w_\ell}}\right)^2 + \sum_{\ell\in\text{bin k}} w_\ell d_\ell - \frac{\left(\sum_{\ell\in\text{bin k}} w_\ell d_\ell\right)^2}{\sum_{\ell\in\text{bin k}} w_\ell}\right\} \\
&= \sum_{k=1}^{K}\left\{\sum_{\ell\in\text{bin k}} w_\ell\left(f_k-\bar{d}_k\right)^2\right\} + \sum_{k=1}^{K}\left\{\sum_{\ell\in\text{bin k}} w_\ell d_\ell - \frac{\left(\sum_{\ell\in\text{bin k}} w_\ell d_\ell\right)^2}{\sum_{\ell\in\text{bin k}} w_\ell}\right\} \\
&= \sum_{k=1}^{K} w_k^*\left(f_k-\bar{d}_k\right)^2 + \sum_{k=1}^{K}\left\{\sum_{\ell\in\text{bin k}} w_\ell d_\ell - \frac{\left(\sum_{\ell\in\text{bin k}} w_\ell d_\ell\right)^2}{\sum_{\ell\in\text{bin k}} w_\ell}\right\}
\end{aligned} \tag{16}$$

where the first equality is obtained by completing the square, where K is the number of forecast bins, and

$$\bar{d}_k = \frac{1}{w_k^*}\sum_{\ell\in\text{bin k}} w_\ell d_\ell \tag{17}$$

$$w_k^* = \sum_{\ell\in\text{bin k}} w_\ell. \tag{18}$$

https://doi.org/10.1017/S1930297500007099 Published online by Cambridge University Press

We now focus on expanding the second term (with curly brackets) from Equation (16):

$$\sum_{k=1}^{K}\left\{\sum_{\ell\in\text{bin k}} w_\ell d_\ell - \frac{\left(\sum_{\ell\in\text{bin k}} w_\ell d_\ell\right)^2}{\sum_{\ell\in\text{bin k}} w_\ell}\right\} = \overline{d} - \sum_{k=1}^{K}\frac{\left(\sum_{\ell\in\text{bin k}} w_\ell d_\ell\right)^2}{\sum_{\ell\in\text{bin k}} w_\ell},$$

where $\overline{d}$ is obtained based on the fact that we are computing a weighted proportion of event occurrences across all days and all questions. When the weights are designed so that each question is weighted equally, this weighted proportion is equal to the simple proportion of event occurrences across all questions. However, we can more generally take

$$\overline{d} = \sum_{\ell=1}^{N} w_\ell d_\ell.$$

Completing the square and rearranging, we then obtain

$$\begin{aligned}
\overline{d} - \sum_{k=1}^{K}\frac{\left(\sum_{\ell\in\text{bin k}} w_\ell d_\ell\right)^2}{\sum_{\ell\in\text{bin k}} w_\ell} &= \overline{d} - 2\overline{d}^2 + 2\overline{d}^2 - \sum_{k=1}^{K}\frac{\left(\sum_{\ell\in\text{bin k}} w_\ell d_\ell\right)^2}{\sum_{\ell\in\text{bin k}} w_\ell} \\
&= \overline{d}(1-\overline{d}) - \left(\sum_{k=1}^{K}\frac{\left(\sum_{\ell\in\text{bin k}} w_\ell d_\ell\right)^2}{\sum_{\ell\in\text{bin k}} w_\ell} - 2\overline{d}^2 + \overline{d}^2\right) \\
&= \overline{d}(1-\overline{d}) - \left(\sum_{k=1}^{K}\frac{\left(\sum_{\ell\in\text{bin k}} w_\ell d_\ell\right)^2}{\sum_{\ell\in\text{bin k}} w_\ell} - 2\overline{d}\sum_{k=1}^{K}\sum_{\ell\in\text{bin k}} w_\ell d_\ell + \overline{d}^2\sum_{k=1}^{K}\sum_{\ell\in\text{bin k}} w_\ell\right) \\
&= \overline{d}(1-\overline{d}) - \sum_{k=1}^{K}\left(\frac{\left(\sum_{\ell\in\text{bin k}} w_\ell d_\ell\right)^2}{\sum_{\ell\in\text{bin k}} w_\ell} - 2\overline{d}\sum_{\ell\in\text{bin k}} w_\ell d_\ell + \overline{d}^2\sum_{\ell\in\text{bin k}} w_\ell\right) \\
&= \overline{d}(1-\overline{d}) - \sum_{k=1}^{K}\left(\sum_{\ell\in\text{bin k}} w_\ell\right)\left(\frac{\left(\sum_{\ell\in\text{bin k}} w_\ell d_\ell\right)^2}{\left(\sum_{\ell\in\text{bin k}} w_\ell\right)^2} - \frac{2\overline{d}\sum_{\ell\in\text{bin k}} w_\ell d_\ell}{\sum_{\ell\in\text{bin k}} w_\ell} + \overline{d}^2\right)
\end{aligned}$$

https://doi.org/10.1017/S1930297500007099 Published online by Cambridge University Press

$$= \overline{d}(1-\overline{d}) - \sum_{k=1}^{K} \left(\sum_{\ell \in \text{bin k}} w_\ell \right) \left(\frac{\left(\sum_{\ell \in \text{bin k}} w_\ell d_\ell \right)}{\sum_{\ell \in \text{bin k}} w_\ell} - \overline{d} \right)^2$$

$$= \overline{d}(1-\overline{d}) - \sum_{k=1}^{K} w_k^* (\overline{d}_k - \overline{d})^2 \tag{19}$$

We can now insert Equation (19) in place of the second term from Equation (16) and rearrange, obtaining

$$\overline{d}(1-\overline{d}) + \sum_{k=1}^{K} w_k^* (f_k - \overline{d}_k)^2 - \sum_{k=1}^{K} w_k^* (\overline{d}_k - \overline{d})^2. \tag{20}$$

□

We now consider the lemma involving the Yates decomposition. Equation (10) of Yates (1982) shows that there is a direct correspondence between the Murphy "calibration" component and the Yates components. This correspondence can be roughly depicted as

$$\text{Miscalib} = \text{Var}_f + (\overline{f} - \overline{d})^2 - 2\text{Cov}_{fd} + \text{Discrim}. \tag{21}$$

Weighted Yates components are facilitated by entering the "weighted" version of Murphy calibration into the above equation. The Yates components then arise from the fact that the terms on the right-hand side all involve summations, so that weights from the Murphy calibration metric can be distributed across each of the components. This leads to the metrics from Equation (10) of the text, as well as the expressions in Table 1. □

B Defining forecast bins/subsets

The Murphy (but not Yates) decomposition requires us to define forecast subsets/bins, such that all forecasts within a bin are treated equivalently. In the case of binary questions, bin definition is simple: we focus on a single alternative, then, e.g., round forecasts to the nearest .05. Thus, we have one bin extending from 0 to .1, another from .1 to .2, and so on up to .9 to 1. Each forecast is coded as its bin midpoint, leading to forecasts of .05, .15, . . ., .95.

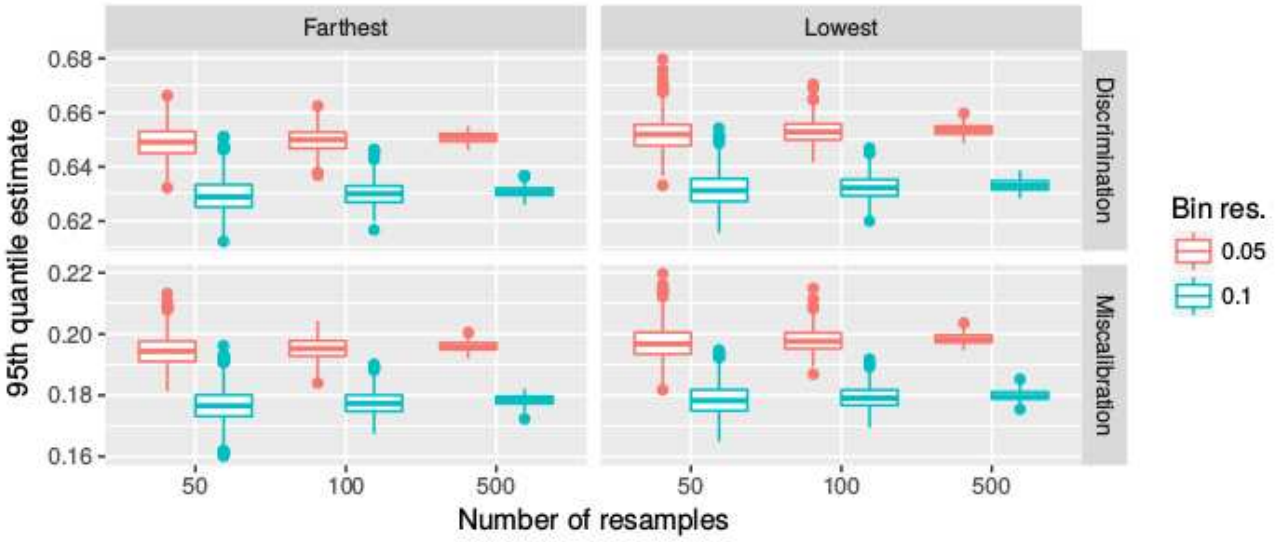
The binning becomes more complicated when we have multinomial questions because we must simultaneously consider all the alternatives. For example, in the case of three alternatives, one bin might be "0 to .1 for Alternative A, 0 to .1 for Alternative B, .8 to .9 for Alternative C," which we could abbreviate as $(0-.1, 0-.1, .8-.9)$. However, if we took the midpoint of each bin, we would obtain the forecast $(.05, .05, .85)$, which does not sum to 1. This causes problems with use of the Brier score, which requires that forecasts sum to 1.

To address this problem, we first round all probabilities to the nearest multiple of .1, .05, or some other number. This rounding automatically creates the bins, because many forecasts will now be equal to one another. There is one caveat here: the rounding can lead to probabilities that do not sum to 1. To make the forecasts sum to 1 (while maintaining rounded probabilities), we focus on two strategies. First, we can find the smallest nonzero forecast and define it to be 1 minus the sum of other rounded forecasts. Second, we can find the individual probability that is furthest from its rounded probability, and define that to be 1 minus the sum of other rounded forecasts (we thank a reviewer for suggesting this second approach).

For example, consider the forecast $(.17, .26, .58)$. When we round, this forecast becomes $(.2, .3, .6)$ and does not sum to 1. Under the first strategy, we would identify .17 as the smallest forecast and set it equal to .1, yielding a rounded forecast of $(.1, .3, .6)$. Under the second strategy, we would identify .26 as being furthest away from a multiple of .1, yielding a rounded forecast of $(.2, .2, .6)$. The rounding is applied to both the Murphy and Yates decompositions for uniformity (though, as mentioned before, the Yates decomposition does not require this). Importantly, due to rounding, no binning solutions will lead to a mean Brier score that is exactly equal to the mean Brier score of the original forecasts; this is true of both the original Murphy decomposition and its extensions here.

https://doi.org/10.1017/S1930297500007099 Published online by Cambridge University Press

Figure 3: Variability in the estimated 95th percentile for the discrimination and miscalibration metrics, by number of resamples (50, 100, or 500), bin resolution (.05 or .1), and strategy for ensuring that rounded forecasts sum to 1 (round the lowest value or round the farthest value).



C Further sensitivity analyses

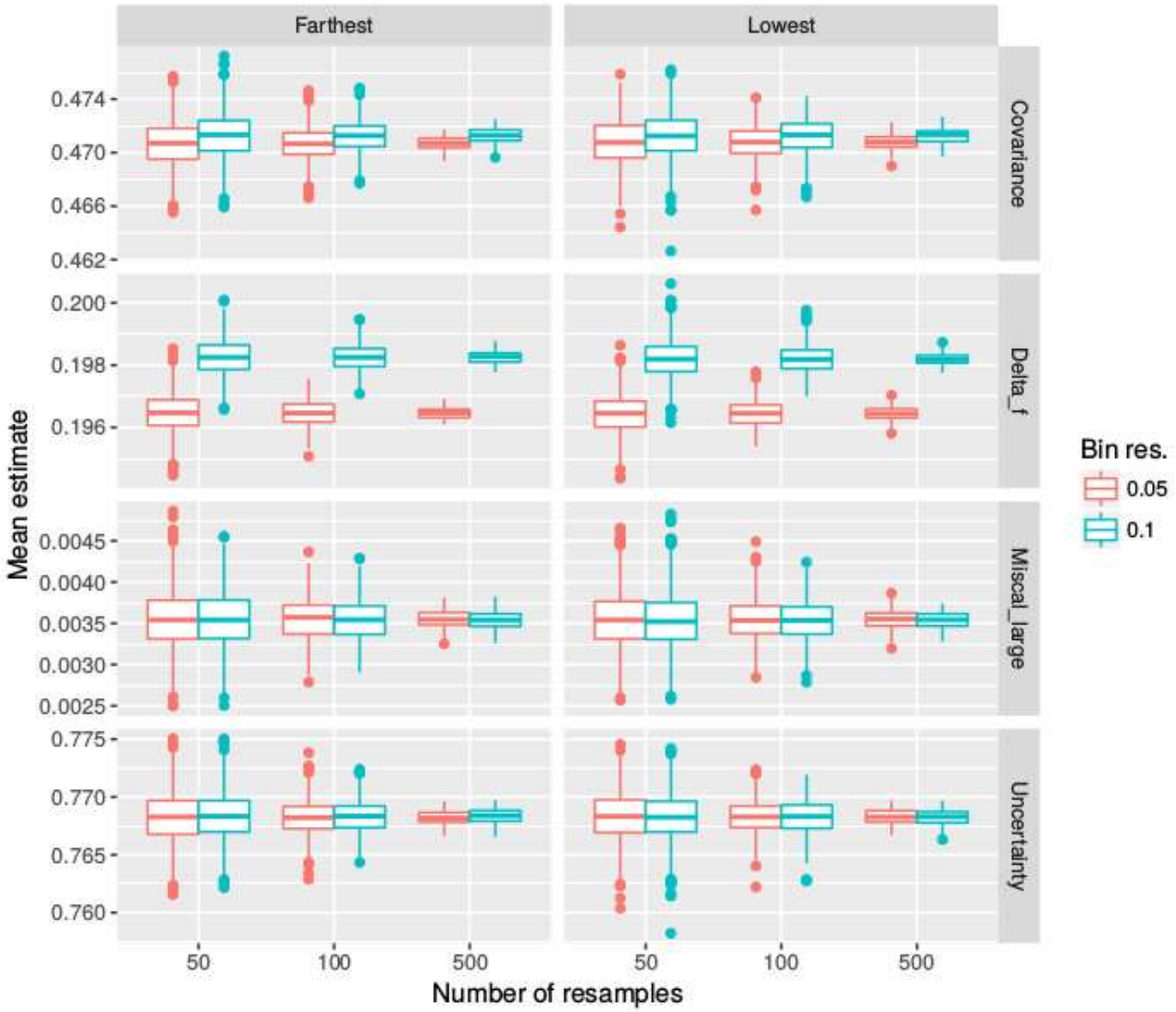
This appendix contains further results of the sensitivity analyses from the Example section, focusing on System 2 (the system with the best overall Brier score). Figure 3 displays resamples of the 95th percentile on both the discrimination and miscalibration metrics. We see that the 95th percentile exhibits more variability across all numbers of resamples, as compared to the means displayed in Figure 1. This is to be expected, because extreme percentiles are more difficult to estimate precisely.

Figure 4 displays resamples of the remaining Brier components' means, including uncertainty, ΔVar_f, miscalibration in the large, and covariance between forecast and outcome (with the latter three stemming from the Yates decomposition). We see that, while precision of the estimates increases with resamples, these components are minimally impacted by the various rounded strategies. The only exception is the ΔVar_f metric, which is lower for the ".05" rounding as compared to the ".1" rounding. Because this metric measures "excess forecast noise," the results correctly suggest that rounding the forecasts to the nearest .1 introduces extra noise in the data, as compared to rounding forecasts to the nearest .05.

Figure 5 displays resamples of the remaining Brier components' mean differences (System 2 minus System 3). We observe that the bin resolutions lead to small differences in the covariance and excess variance metrics, a finding that is similar to the Murphy component results from Figure 2. We also see little influence of rounding method ("round farthest" vs "round lowest"), which is also similar to the Murphy component results.

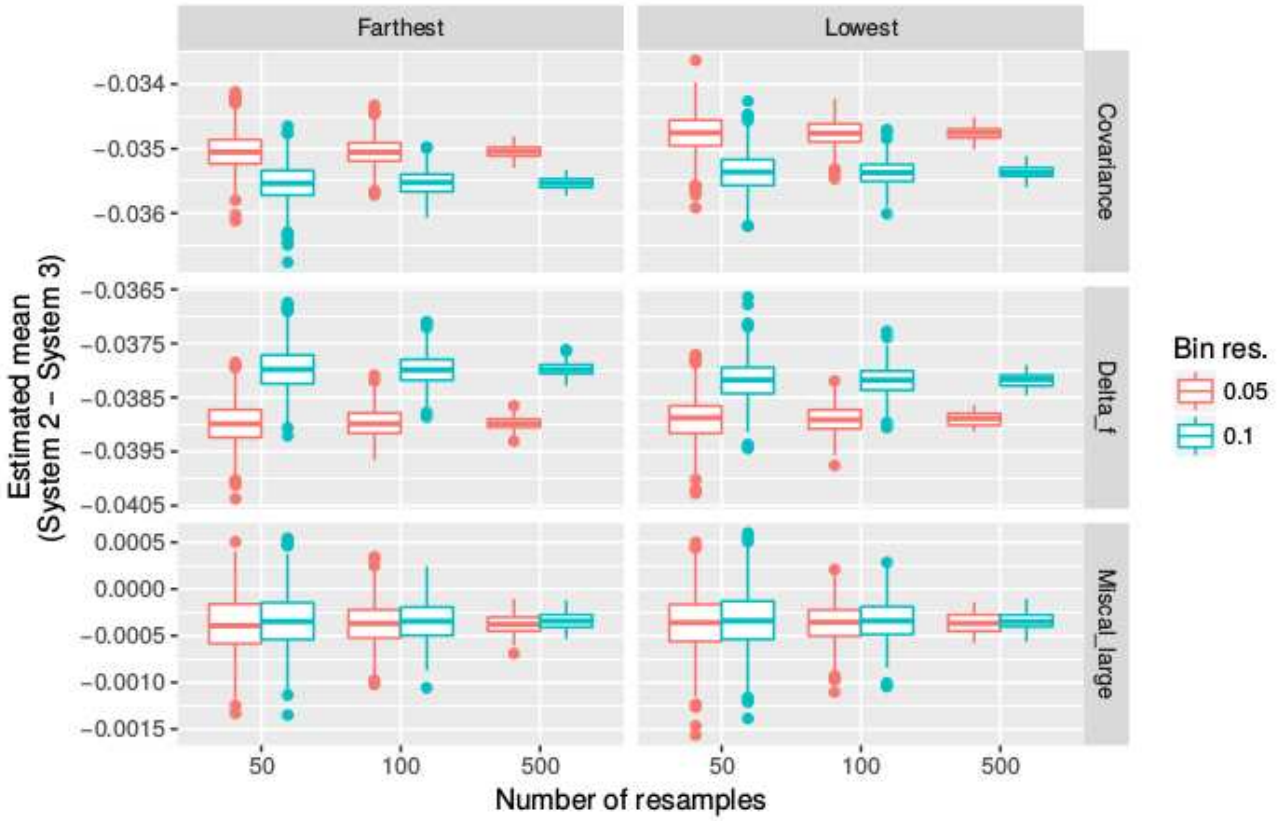
https://doi.org/10.1017/S1930297500007099 Published online by Cambridge University Press

Figure 4: Variability in the estimated means of other Brier components, by number of resamples (50, 100, or 500), bin resolution (.05 or .1), and strategy for ensuring that the rounded forecasts sum to 1 (round the lowest value or round the farthest value).



https://doi.org/10.1017/S1930297500007099 Published online by Cambridge University Press

Figure 5: Variability in the estimated mean differences of other Brier components, by number of resamples (50, 100, or 500), bin resolution (.05 or .1), and strategy for ensuring that the rounded forecasts sum to 1 (round the lowest value or round the farthest value).



https://doi.org/10.1017/S1930297500007099 Published online by Cambridge University Press