DECISION ANALYSIS
Vol. 10, No. 4, December 2013, pp. 292–304
ISSN 1545-8490 (print) | ISSN 1545-8504 (online)

http://dx.doi.org/10.1287/deca.2013.0280

Choosing a Strictly Proper Scoring Rule

Edgar C. Merkle
Department of Psychological Sciences, University of Missouri, Columbia, Missouri 65211, merklee@missouri.edu
Mark Steyvers
Department of Cognitive Sciences, University of California, Irvine, Irvine, California 92697, mark.steyvers@uci.edu

Strictly proper scoring rules, including the Brier score and the logarithmic score, are standard metrics by which probability forecasters are assessed and compared. Researchers often find that one's choice of strictly proper scoring rule has minimal impact on one's conclusions, but this conclusion is typically drawn from a small set of popular rules. In the context of forecasting world events, we use a recently proposed family of proper scoring rules to study the properties of a wide variety of strictly proper rules. The results indicate that conclusions vary greatly across different scoring rules, so that one's choice of scoring rule should be informed by the forecasting domain. We then describe strategies for choosing a scoring rule that meets the needs of the forecast consumer, considering three unique families of proper scoring rules.

Key words: proper scoring rule; forecasting; subjective probability; Brier score
History: Received on March 13, 2013. Accepted by Editor-in-Chief Rakesh Sarin on July 31, 2013, after 2 revisions.

Introduction

In both research and application, there is often the need to assess the correspondence between probabilistic forecasts and event outcomes. There exist a variety of statistical metrics to accomplish this assessment, and analysts typically prefer metrics that are not vulnerable to manipulation; that is, metrics for which an individual cannot gain an advantage by systematically modifying her forecasts. This is because, if a forecaster can improve her scores by modifying the forecasts in light of the metric, then we have no way of knowing when the forecaster is being truthful and when the forecaster is capitalizing on the metric.

Starting with Brier (1950), Murphy (1972), and others, these arguments have led to a body of work on proper scoring rules (for a formal review, see Winkler and Jose 2010). To formally define a proper scoring rule, let f be a probabilistic forecast of a Bernoulli trial d with true success probability p. Proper scoring rules are metrics whose expected values are minimized if $f = p$. Strictly proper scoring rules, a subset of proper scoring rules, are metrics whose expected values are minimized if and only if $f = p$. Although there exists an infinite number of unique, proper scoring rules, researchers typically employ a very small number of strictly proper scoring rules in practice. These include the Brier (quadratic) score, the logarithmic score, and the spherical score. When considering only these few popular rules, researchers often find that one's choice of rule does not impact one's conclusions. This, in turn, may lead researchers to believe that conclusions are robust to choice of scoring rule. For example, Staël von Holstein (1970) states that "different scoring rules lead to essentially the same rankings of the assessors, at least when the ranks are based on average scores" (p. 154), with similar statements being made by Winkler (1971) and O'Hagan et al. (2006). More recently, Bickel (2007) conducted a detailed examination of these three scoring rules and showed that, although rankings resulting from the three rules are highly correlated, specific individuals may lose or gain many spots in the rankings. The change in rankings is most prevalent when the number of potential outcomes is greater than two.

Although many researchers have considered larger families of proper scoring rules (e.g., Gneiting and Raftery 2007; Hand and Vinciotti 2003; Johnstone et al. 2011; Jose et al. 2008, 2009; Winkler 1996), the focus

is often on tailored model estimation instead of tailored forecast evaluation. One exception is Johnstone et al. (2011), who developed tailored scoring rules that match a decision maker's utility for various forecast-outcome combinations. Additionally, Tetlock (2005) considered many adjustments to standard scoring rules in the context of political forecasts (see especially the technical appendix). In general, however, the properties of these alternative scoring rules in practice is underexplored, as are methods for selecting specific scoring rules from these families.

In this paper, we employ families of proper scoring rules (first the beta family, followed by the power and pseudospherical families) to obtain a more comprehensive evaluation of the impact of scoring rule choice on forecaster comparison. We focus on forecasts for binary items, which is the situation where Bickel (2007) and others found the three popular scoring rules (Brier, logarithmic, and spherical) to be most similar. We demonstrate that different strictly proper scoring rules can yield very different substantive conclusions, which implies that researchers should carefully consider the scoring rule that is used to evaluate forecasters. We then provide strategies for choosing a scoring rule that is tailored to a specific forecasting domain.

The remainder of the paper is organized as follows: We first outline the family of proper scoring rules that we employ, along with the ways in which we use the family. Next, we apply the methods to compare forecasts of world events, showing that different strictly proper scoring rules yield considerably different conclusions. In light of this result, we discuss strategies for choosing specific rules from families of proper scoring rules. Finally, we discuss practical implications.

The Beta Family of Proper Scoring Rules

We initially focus on a parametric family of proper scoring rules proposed by Buja et al. (2005). In this section, we describe key results and background that largely follows that of Buja et al.

We begin by considering scoring rules to be loss functions associated with the reported forecasts $\mathbf{f}$. Taking d_i to be the outcome of trial i, f_i to be the associated forecast, and $l(d_i \mid f_i)$ to be the "loss" associated with trial i $(i = 1, \ldots, I)$, we can generally write a scoring rule as

$$L(\mathbf{d} \mid \mathbf{f}) = \frac{1}{I} \sum_{i=1}^{I} l(d_i \mid f_i). \tag{1}$$

Because we are focusing on situations where d_i is the outcome of a Bernoulli trial, we can write the loss as

$$l(d_i \mid f_i) = d_i l(1 \mid 1 - f_i) + (1 - d_i) l(0 \mid f_i), \tag{2}$$

so that $l(0 \mid f_i)$ is increasing in f_i and $l(1 \mid 1 - f_i)$ is increasing in $1 - f_i$. Additionally, both functions should be bounded from below (typically at zero).

For situations where the d_i are Bernoulli, it can be shown that the scoring rule is proper if and only if

$$l(1 \mid 1 - f_i) = \int_{f_i}^{1} (1 - t)\omega(t)\, dt, \tag{3}$$

$$l(0 \mid f_i) = \int_{0}^{f_i} t\omega(t)\, dt, \tag{4}$$

for some function $\omega(t)$ that is nonnegative and finite across all open intervals on $(0, 1)$ (see Schervish 1989, Buja et al. 2005). Additionally, the scoring rule is strictly proper if $\omega(t)$ is nonzero across all open intervals on $(0, 1)$. Thus, we can obtain a variety of (strictly) proper scoring rules by defining $\omega(t)$ in various ways. For example, we obtain the Brier score by defining $\omega(t) = 1$, and we obtain the logarithmic score by defining $\omega(t) = t^{-1}(1 - t)^{-1}$.

Instead of defining a single scoring rule through $\omega(t)$, Buja et al. (2005) developed a family of scoring rules by parameterizing $\omega(t)$. Their *beta family* is defined through the function

$$\omega(t \mid \alpha, \beta) = t^{\alpha - 1}(1 - t)^{\beta - 1}, \quad \alpha > -1,\ \beta > -1, \tag{5}$$

with popular scoring rules being obtained as special cases. For example, taking $\alpha = \beta = 0$, we obtain the logarithmic scoring rule via

$$l(1 \mid 1 - f_i) = \int_{f_i}^{1} (1 - t)t^{-1}(1 - t)^{-1}\, dt \tag{6}$$

$$= \int_{f_i}^{1} t^{-1}\, dt \tag{7}$$

$$= \log(1) - \log(f_i) \tag{8}$$

$$= -\log(f_i) \tag{9}$$

and, similarly, $l(0 \mid f_i) = -\log(1 - f_i)$. The Brier score can be obtained by setting $\alpha = \beta = 1$. Further, as α and β go to ∞ together, we obtain a rule that is equivalent to "misclassification" scoring, defined as

$$l(1 \mid 1 - f_i) = c \text{ if } f_i < 0.5, \ \ 0 \text{ otherwise;} \quad (10)$$

$$l(0 \mid f_i) = c \text{ if } f_i > 0.5, \ \ 0 \text{ otherwise;} \quad (11)$$

for some constant c. As α and β go together to ∞, scoring rules from the beta family assume larger and larger values. The beta family scoring rules can immediately be scaled, however, to attain whatever maximum is desired.

In addition to these popular scoring rules, the beta family yields novel proper scoring rules when $\alpha \neq \beta$. Buja et al. (2005) show that, for $\alpha, \beta > 0$, the rules correspond to situations where the cost of the outcome $d_i = 1$ differs from the cost of the outcome $d_i = 0$. In the misclassification case, this concept is intuitive to understand: if $c \in (0, 1)$ is the cost of a false positive and $1 - c$ is the cost of a false negative, cost-weighted misclassification may be obtained by replacing Equations (10) and (11) with

$$l(1 \mid 1 - f_i) = 1 - c \text{ if } f_i \leq c, \ \ 0 \text{ otherwise;} \quad (12)$$

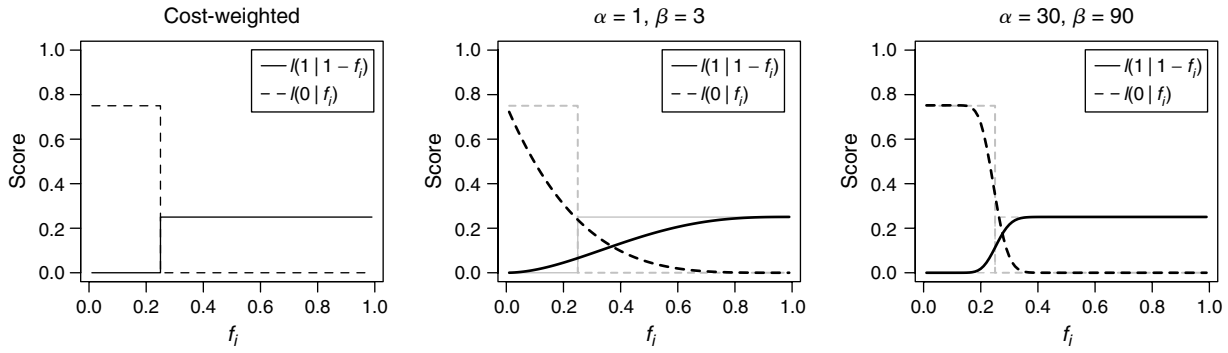
$$l(0 \mid f_i) = c \text{ if } f_i > c, \ \ 0 \text{ otherwise;} \quad (13)$$

where the choice of score at $f_i = c$ is arbitrary. Note that c is involved in both the score and the threshold because we assumed $c \in (0, 1)$ and the false negative cost equals $1 - c$.

The cost-weighted misclassifications are step functions whose values change at the point $f_i = c$, as displayed in the left panel of Figure 1 (taking $c = 0.25$). This figure includes f_i on the x axis, the score associated with f_i on the y axis, and two lines for the two different outcomes associated with d_i. The left panel is a simple visualization of Equations (12) and (13) with $c = 0.25$, with the line whose maximum is 0.25 representing $l(0 \mid f_i)$ and the line whose maximum is 0.75 representing $l(1 \mid 1 - f_i)$. In this graph, it is clear that small values of f_i (< 0.25) are most important: if one reports a small probability and is incorrect, one receives a large score. Conversely, if one reports a large probability and is incorrect, the resulting score is considerably lower.

The beta family's cost-weighted scoring rules are similar to the step functions in the left panel of Figure 1, except that the scores change smoothly across values of f_i. We obtain the false-positive cost c from $\alpha/(\alpha + \beta)$, with the scoring rule being more similar to a step function as α and β increase (while maintaining the same ratio). For values of α and β close to zero, the curves gradually change, essentially reflecting uncertainty in the exact value of c. This result is illustrated in the center panel of Figure 1, which displays curves for $\alpha = 1$, $\beta = 3$. As α, β increase to 30, 90 (right panel), the curves get closer to step functions. In all three panels, low values of f_i have the most impact on the resulting score. In contrast, if we had $\alpha > \beta$, high values of f_i would have the most impact.

Figure 1 Illustrations of Three Proper Scoring Rules in the Beta Family



Notes. The left panel illustrates a cost-weighted misclassification rule, the center panel illustrates the scoring rule with $\alpha = 1, \beta = 3$, and the right panel illustrates the scoring rule with $\alpha = 30, \beta = 90$. It is seen that, as α and β increase while maintaining a constant ratio, the scoring rules become similar to cost-weighted misclassification.

The right panel of Figure 1 also shows that both curves are essentially flat across much of the range of f. Although this scoring rule is technically strictly proper (because the curves are never exactly flat), it yields scores that are practically equal for multiple values of f. For example, if forecaster A always reported a probability of 0.85 and forecaster B always reported a probability of 0.5, the (30, 90) scoring rule would assign them practically equal scores. Thus, we might call this particular rule *practically nonstrict*. We return to this issue later, where we discuss its impact on the variability in the conclusions that one draws. First, however, we use the family to study the scoring rules' properties when applied to real forecasts.

Application: Forecasting World Events

The forecasts considered here arise from the Aggregative Contingent Estimation System (ACES), a Web-based environment that solicited forecasts concerning world events from the general public. The focal data in this paper include forecasts from over 1,000 forecasters on over 200 unique forecasting problems. Importantly, forecasters voluntarily logged in to the website and chose specific problems to forecast, resulting in very sparse data. A more detailed summary of the rationale and data collection procedures can be found in Warnaar et al. (2012).

We use the ACES data to study the stability of forecast evaluations in two general areas: comparison of individual forecasters to a baseline, and comparison of forecasters to one another. The two parameters of the beta family make it straightforward to carry out these comparisons. This is because two sets of forecasts for the same events (provided by human forecasters, statistical models, aggregation methods, etc.) can be evaluated at arbitrary points in the two-dimensional space defined by (α, β), which allows us to evaluate large grids of points within this space.

In our comparisons, we study summary statistics across different scoring rules in the beta family. Our summary statistics are generally based on the rank ordering that is implied by the resulting scores. To study the consistency of rank ordering across J sets of forecasts for the same events (say, $\mathbf{f}_j$, $j = 1, \ldots, J$), we can calculate the rank ordering of the J sets that is implied by one scoring rule. We can then calculate the rank ordering under other scoring rules in the beta family, computing the Spearman correlation between the two sets of rankings:

$$\frac{12\sum_{j=1}^{J}\{[R_{j,\alpha,\beta}-(J+1)/2][S_{j,\delta,\gamma}-(J+1)/2)]\}}{J(J^2-1)}, \quad (14)$$

where $R_{j,\alpha,\beta}$ is the rank of set j under the beta scoring rule with (α, β) and $S_{j,\delta,\gamma}$ is the rank of set j under the beta scoring rule with (δ, γ). In setting $(\delta, \gamma) = (1, 1)$ (which reflects the Brier score), we can use this equation to compare ranks under arbitrary scoring rules to ranks under the Brier score.

We also wish to compare individual forecasters to a baseline forecast. In this situation, we can count the proportion of individuals who "lose to" the baseline (i.e., who have higher scores than the baseline) under specific scoring rules in the beta family. Given values of α and β, this can be written as

$$\sum_{j=1}^{J} 1(L(\mathbf{d} \mid \mathbf{f}_j) > L(\mathbf{d} \mid \mathbf{f}^*)), \quad (15)$$

where $L(\mathbf{d} \mid \mathbf{f}_j)$ is the score associated with forecaster j (see Equation (1)), $L(\mathbf{d} \mid \mathbf{f}^*)$ is the score associated with the baseline $\mathbf{f}^*$, and $1(\ )$ is an indicator function that equals one when the condition is satisfied.

In the next two sections, we use these methods to evaluate forecasts of general world events across the beta family of scoring rules. We use the restriction $\alpha, \beta > 0$ to maintain interpretability: although the beta family extends to $\alpha = -1$, $\beta = -1$, the cost-weighting interpretation breaks down for negative values of α and β.

Comparing Forecasters to One Another

In this section, we use the beta family to generally compare forecasters across a large set of proper scoring rules, studying the extent to which forecaster rankings vary across the rules.

Method. Using the ACES data, we compared 10 forecasters on 21 binary problems that they forecasted (the appendix contains a set of artificial forecasts that mimic the real forecasts). This set of forecasters was selected because they all provided forecasts for 21 common problems. Although the ACES data included thousands of forecasters and hundreds of problems, it was generally difficult to find a set of forecasters providing forecasts for the same

problems. This is because forecasters were free to select forecasting problems, resulting in very sparse data.

For the world events forecasted on the ACES website, it is usually the case that $d_i = 1$ is valued more heavily than $d_i = 0$. This is because the data are coded so that for all i, $d_i = 1$ implies a change from the status quo. That is, $d_i = 1$ includes changes in world leaders and new conflicts between nations, and $d_i = 0$ implies no change from the current state of affairs. Thus, a strictly proper scoring rule with $\alpha \neq \beta$ may be more useful for this particular application. As previously discussed in the context of Figure 1, scoring rules with $\alpha < \beta$ emphasize low-probability forecasts: low-probability forecasts where $d_i = 1$ incur a large penalty, and scores associated with high-probability forecasts do not vary greatly. Conversely, scoring rules with $\alpha > \beta$ emphasize high-probability forecasts: high-probability forecasts associated with $d_i = 0$ incur a large penalty, and the scores associated with low-probability forecasts do not change greatly.

Although scoring rules with $\alpha \neq \beta$ are useful, in practice we may need to choose specific values of α and β for "official" scoring. We provide some discussion of this issue later. For now, however, we focus on the sensitivity of one's conclusions to the choice of scoring rule. This can be assessed via evaluation of the forecasters at multiple values of α and β.

Results. The forecaster rankings implied by the Brier score and by the logarithmic score are displayed in the middle two columns of Table 1. It is observed that these rankings exhibit more variability than may typically be expected, most notably for forecaster 1 (third row). This is because forecaster 1 tended to report extreme forecasts, and the incorrect forecasts were heavily penalized under the logarithmic score. The sensitivity of the logarithmic score to incorrect, extreme forecasts has been characterized as both an advantage (Johnstone 2011) and disadvantage (Selten 1998): advantageous in the sense that it represents an individual who has log utility for wealth in a gambling context (a "Kelly bettor"; see Johnstone 2007 for further discussion), and disadvantageous in the sense that, if one incorrectly makes a forecast of 0 or 1, then one's average score can never recover. Aside from forecaster 1, four other forecasters' rankings changed by four spots across the Brier and logarithmic scores. To study the behavior of the rules in more detail, artificial data that mimic the properties of the real data can be found in the appendix.

Table 1 Forecaster Rankings Implied by the Brier Score, Logarithmic Score, and Another Scoring Rule from the Beta Family

	Scoring rule		
Forecaster number	Brier $\alpha = 1$, $\beta = 1$	Log $\alpha = 0$, $\beta = 0$	Beta $\alpha = 9$, $\beta = 3$
3	1 (0.09)	3 (0.61)	2 (0.05)
9	2 (0.15)	5 (0.78)	4 (0.06)
1	3 (0.16)	10 (1.08)	3 (0.05)
5	4 (0.17)	1 (0.49)	5 (0.07)
6	5 (0.20)	9 (1.00)	6 (0.10)
7	6 (0.20)	2 (0.58)	1 (0.04)
8	7 (0.21)	8 (0.99)	7 (0.11)
2	8 (0.23)	4 (0.70)	8 (0.12)
4	9 (0.31)	7 (0.94)	9 (0.15)
10	10 (0.31)	6 (0.87)	10 (0.16)
Spearman correlations	1.00	0.15 1.00	0.81 0.31 1.00

Note. Numbers in parentheses are the actual scores assigned to each forecaster.

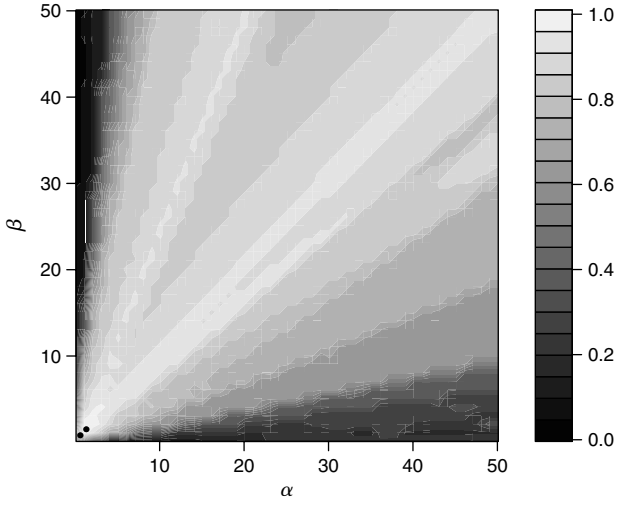
The final column of Table 1 displays the ranking under the scoring rule with $\alpha = 9$, $\beta = 3$, which heavily emphasizes high-probability forecasts. The rankings further differ from those under the Brier and logarithmic scores, resulting in some different conclusions. Most notably, the best forecaster under this beta score was ranked sixth by the Brier score. Additionally, the third-ranked forecaster under this beta score was ranked 10th by the logarithmic score. The Spearman rank-order correlations between the three scoring rules are displayed at the bottom of the table. These statistics show that the rankings are only modestly related, and they also show that large correlations do not necessarily imply complete consistency. For example, rankings under the Brier score and beta score have a relatively large correlation of 0.81. However, as noted previously, some individual forecaster rankings changed considerably across these two rules.

In Table 1, the numbers in parentheses are the average scores that each forecaster received under each rule. These show that the range of scores is compressed under the beta rule with $\alpha = 9$, $\beta = 3$, as compared to the other two rules. This could result in extra

variability in the rankings as compared to the other scores (because they are closer together, so that rankings may switch more often), though it is difficult to use the differences between pairs of scores to draw any specific conclusions. In particular, the scalings of the scoring rules are arbitrary, so that a difference of 0.01 could be very large in one instance and very small in another instance.

To more globally examine the impact of scoring rule on model rankings, Figure 2 displays Spearman correlations between the forecaster ranking implied by the Brier score and the forecaster ranking implied by other scoring rules in the beta family. The x axis represents values of the α parameter, the y axis represents values of the β parameter, and the shading represents the value of the rank-order correlation. As the color moves from white to black, the model ordering from the beta scoring rule becomes less related to the model ordering from the Brier score. The figure shows that, depending on the specific scoring rule used, the rank order of the forecasters can change dramatically. The correlations decrease as we move off the diagonal, especially toward the upper left and lower right corners of the plot. The lower right corner reflects scoring rules for which $\alpha > \beta$, which are rules for which high-probability forecasts are emphasized and low-probability forecasts are de-emphasized. The upper left corner reflects the opposite type of scoring rule. Thus, if we place large value on high- (low-) probability forecasts and small value on low- (high-) probability forecasts, our forecaster assessment will be considerably different than our assessment under the Brier score.

Figure 2 **Contour Plot Displaying Spearman Rank-Order Correlations Between the Ordering Implied by the Brier Score and the Ordering Implied by Beta Family Scoring Rules**



Although choice of scoring rule has a large impact on the results in this section, we do note that only 10 forecasters were involved. The correlations may not change as much with larger numbers of forecasters, though there would also seem to be greater opportunity for changes in rankings. In the next section, we consider a larger number of forecasters.

Comparing Individual Forecasters to the Average

In this section, individual forecasters are compared to the average forecast. The average forecast is not necessarily a good benchmark against which to evaluate individual forecasters. However, the average forecast is often better than the typical forecaster (e.g., Armstrong 2001) or than a randomly-selected forecaster (e.g., Davis-Stober et al. 2013)—results that are generally described as the *wisdom of crowds* (Surowiecki 2005). These results have been demonstrated in a wide variety of applications (e.g., Steyvers et al. 2009, Turner et al. 2013, Yi et al. 2010), and we study here the extent to which the results are robust across sets of proper scoring rules.

Method. Data are from 624 ACES forecasters who forecasted at least eight problems. The eight-problem threshold is arbitrary and is intended to eliminate the variability resulting from forecasters who forecasted a small number of problems. We studied the results under various thresholds from 3 to 16, and they remain similar regardless of the specific threshold chosen.

To study the wisdom of crowds effect, we compare each forecaster to the unweighted average of all items that the forecaster chose to forecast. In other words, we employ criterion (15), where $\mathbf{f}^*$ is the unweighted average forecast and J differs for each forecaster. We first compute (15) with α and β fixed to 1, reflecting the wisdom of crowds effect under the Brier score. We then compute (15) across a large set of scoring rules in the beta family, examining the extent to which the wisdom of crowds effect is robust to choice of scoring rule. As the proportion of individuals "losing to"

the average decreases to 0.5 and beyond, the wisdom of crowds effect disappears. It is of interest to examine the types of rules, if any, that cause the effect to disappear.

Results. Focusing on Brier scores, we found that the average beat 519 of 624 individuals. The 95% confidence interval associated with this proportion (computed via R's binom.test( ) function) is (0.8, 0.86), which may be taken to indicate a wisdom of crowds effect. However, we can also find strictly proper scoring rules for which the effect disappears. For example, under the beta-family scoring rule associated with $\alpha = 0.4$ and $\beta = 3.45$, we find that the average beat only 333 of 624 individuals. The 95% confidence interval associated with this proportion is (0.49, 0.57), which suggests that the wisdom of crowds effect has largely diminished, if not completely disappeared.

The $\alpha = 0.4$, $\beta = 3.45$ scoring rule places emphasis on low-probability forecasts. This is illustrated in Figure 3, which is similar to the earlier Figure 1: the two lines reflect the score that one receives for a forecast f (x axis), depending on whether the outcome is $d = 0$ or $d = 1$. This specific figure looks very similar to the middle panel of Figure 1, where low-probability forecasts associated with $d = 1$ are most heavily penalized. Additionally, high-probability forecasts from about 0.7 to 1 are assigned essentially equal scores. Such a scoring rule may be useful when $d = 1$ is a rare occurrence: regardless of the forecast that one makes on a $d = 0$ trial, one's score does not change by much. However, if one makes a bad forecast on a $d = 1$ trial, one receives a harsh punishment.

Returning to the comparison of individuals to the average forecast, we found that the scoring rule from Figure 3 diminishes the wisdom of crowds effect. This result implies that the average forecast receives harsh punishment more often than the individual forecasters: the average forecast is good at predicting that the status quo will be maintained ($d = 0$), at the cost of some bad forecasts associated with the overturning of the status quo ($d = 1$). Conversely, individuals forecast the overturning of the status quo more often, which in turn avoids the large penalties under the $\alpha = 0.4$, $\beta = 3.45$ rule.

We consider an expanded set of scoring rules in Figure 4, displaying the proportion of individuals beaten by the average under each rule in the set. The x axis corresponds to values of the α parameter, the y axis corresponds to the proportion of individuals beaten by the average, and separate lines correspond to values of the β parameter. It is seen that, for values of β that are large relative to α, the average beats fewer than half of the individuals. It is additionally seen that, in cases where $\alpha = \beta$, the proportion remains stable at just above 0.8. This is especially notable because the logarithmic score ($\alpha = 0, \beta = 0$) and Brier score ($\alpha = 1, \beta = 1$) are included among these cases.

We do not argue that the $\alpha = 0.4$, $\beta = 3.45$ scoring rule (or others considered in Figure 4) is the most

Figure 3 **Plot of the $\alpha = 0.4$, $\beta = 3.45$ Scoring Rule**

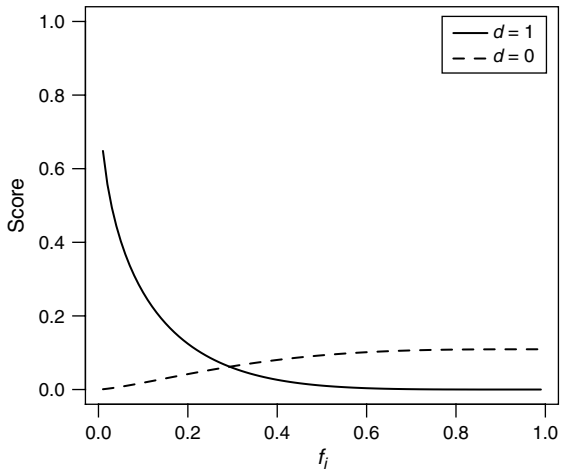
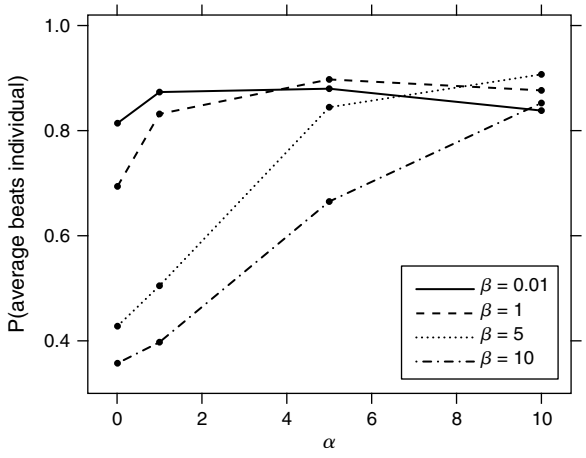


Figure 4 **The Proportion of Individuals Beaten by the Average Under an Expanded Set of Scoring Rules in the Beta Family**



sensible one to use in practice. Additionally, the fact that the lines in Figure 3 look flat over some intervals of f implies that this particular scoring rule is practically nonstrict. That is, although the scoring rule has a unique minimum (achieved by forecasting the true probability of event occurrence), there exist multiple values of f_i for which the resulting score is practically equivalent to the minimum. In practice, one may wish to define a threshold that separates "practically nonstrict" scoring rules from other, strictly-proper scoring rules. If we only consider the latter subset, then the wisdom of crowds effect may not be diminished to such an extent. More generally, the results in this section illustrate the fact that strictly-proper scoring rules do not place strong constraints on the conclusions drawn.

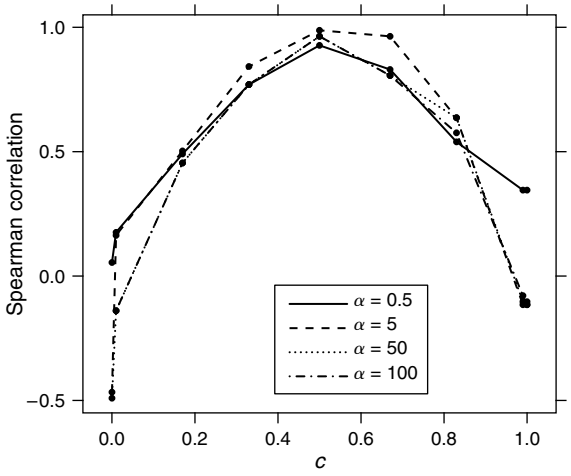
Choice of Scoring Rule

Given the variability observed in the previous sections, the reader is likely to wonder how a specific scoring rule could be chosen for a specific forecasting domain. Focusing on the beta family of scoring rules, the most intuitive choice may involve a focus on the cost $c = \alpha/(\alpha + \beta)$. Note that, in the context of evaluation, c represents the decision-maker's cost, as opposed to the forecaster's cost. The forecasters themselves may often have a different view of the costs, either through their own beliefs or through the scoring rule that was presented to them.

In classification contexts (i.e., when forecasts can only equal zero or one), we mentioned earlier that c reflects the cost of a false positive, and $1 - c$ the cost of a false negative (see Equations (10) and (11)). In probabilistic forecasting contexts, c reflects the relative emphasis on high-probability forecasts, as opposed to low-probability forecasts: a value of $c = 0.5$ reflects equal emphasis on low- and high-probability forecasts, and values of c greater (less) than 0.5 reflect emphasis on high- (low-)probability forecasts. Although c does not determine a specific scoring rule in the beta family (one must also fix either α or β), it often accounts for much of the variability that is observed across scoring rules.

To demonstrate the impact of c on scoring-rule variability, we revisit the two examples described earlier in the paper. For each example, we examine the variability in scoring rules for fixed values of $c \in (0, 1)$. Focusing on the comparison of forecasters to one another, Figure 5 displays Spearman correlations between forecaster ranking under the Brier score and forecaster ranking under other rules in the beta family. The x axis is c, the y axis is the Spearman correlation, and lines reflect values of α (along with α and β, a beta family scoring rule can be uniquely determined by c and α). It is observed that, for a fixed value of c, the Spearman correlations tend to be very similar to one another. This implies that, if one fixes c a priori, the results are relatively insensitive to the specific scoring rule chosen. As c becomes more extreme, the results exhibit more variability.

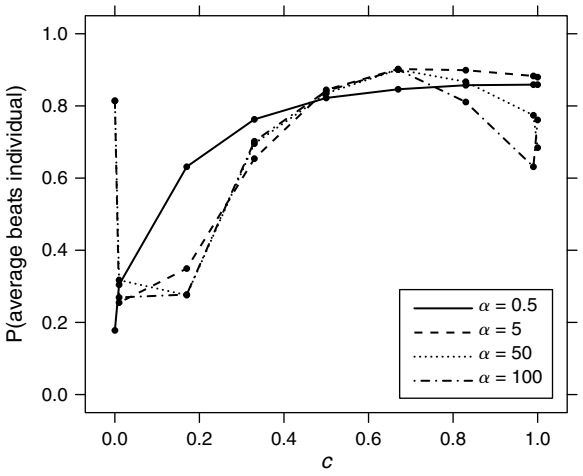
Figure 5 **Spearman Correlation Between Forecaster Rankings Under the Brier Score and Forecaster Rankings Under Other Scores in the Beta Family**



Focusing on the comparison of forecasters to the average, Figure 6 displays the proportion of individuals that are beaten by the average for various scoring rules in the beta family. This figure is similar to Figure 4. It is again observed that, for fixed values of $c \in (0.2, 0.8)$, the proportion of individuals beaten by the average is similar across scoring rules. Additionally, there is greater variability in the correlations for extreme values of c.

These results suggest a general strategy for choosing a scoring rule in the beta family. First, one chooses a value c that reflects the relative cost of false positives and false negatives in a misclassification context. If c is, say, between 0.2 and 0.8, then the specific scoring rule that is ultimately chosen may not exert a large influence over one's conclusions. If c is more

Figure 6 **The Proportion of Individuals Beaten by the Average Under an Expanded Set of Scoring Rules in the Beta Family**



extreme, however, then choice of scoring rule may still exhibit large variability. To choose a specific scoring rule for fixed c, we propose restricting $\alpha+\beta>0$, then experimenting with a small number of specific $\alpha+\beta$ values. These values can be roughly conceptualized as the certainty associated with choice of c (see Buja et al. 2005 for further discussion): values of $\alpha+\beta$ close to zero imply low certainty in c, whereas increasing values of $\alpha+\beta$ imply increasing certainty. Additionally, to give some perspective, our experience indicates that values of $\alpha+\beta$ greater than, say 100, tend to result in similar conclusions.

Although useful, this strategy is not foolproof. For example, there are likely to be some situations where c is close to 0.5 yet scoring rules exhibit varying conclusions. Additionally, researchers may be interested in data summaries other than those used in this paper (which were Spearman correlations and proportions of forecasters surpassing a threshold). To choose a scoring rule in these situations, researchers may fix c, generate artificial data that mimic the domain of interest, and study variability in the measure of interest across multiple scoring rules. To ease such an examination, we have written an R package (scoring) that contains our implementation of the beta family (along with the power and pseudospherical families described later). This package is on the Comprehensive R Archive Network and can be downloaded and installed in the usual way.

Comparison to Other Families

The beta family is not the only one that could be used to study sensitivity to choice of scoring rule. Other notable families include the power family and pseudospherical family, which are one-parameter families that encompass a wide variety of proper scoring rules. For two-alternative situations such as those considered here, the families may be written as (Jose et al. 2008, 2009):

$$l_{\text{pow}}(d_i \mid f_i) = -\left(\frac{r_i^{\gamma-1}-1}{\gamma-1} - \frac{[r_i^{\gamma}+(1-r_i)^{\gamma}-1]}{\gamma}\right), \quad (16)$$

$$l_{\text{sph}}(d_i \mid f_i) = -\frac{1}{\gamma-1}\left[\left(\frac{r_i}{(r_i^{\gamma}+(1-r_i)^{\gamma})^{1/\gamma}}\right)^{\gamma-1}-1\right], \quad (17)$$

where $r_i = d_i f_i + (1-d_i)(1-f_i)$ (which is just the forecast associated with the outcome that occurred) and $\gamma>1$. As γ tends to one, both families converge to the logarithmic scoring rule. For $\gamma=2$, we obtain the Brier score from the power family and the spherical score from the pseudospherical family.

We conjecture that scoring rules from the beta family are more likely to exhibit varying conclusions than are the scoring rules within either family above, because the beta family has two parameters and a more complex functional form. The abovementioned families have been extended, however, to situations where one wishes to evaluate forecasts with respect to a baseline (or prior) forecast (Jose et al. 2009). For two alternative forecasts, we take $q_i = d_i b + (1-d_i)(1-b)$, where b is the baseline forecast associated with $d_i=1$. The power and pseudospherical families with baseline b may then be written as:

$$l_{\text{bpow}}(d_i \mid f_i) = -\left(\frac{(r_i/q_i)^{\gamma-1}-1}{\gamma-1} - \frac{[r_i^{\gamma}/q_i^{\gamma-1}+(1-r_i)^{\gamma}/(1-q_i)^{\gamma-1}-1]}{\gamma}\right), \quad (18)$$

$$l_{\text{bsph}}(d_i \mid f_i) = -\frac{1}{\gamma-1}\left[\left(\frac{r_i/q_i}{(r_i^{\gamma}/q_i^{\gamma-1}+(1-r_i)^{\gamma}/(1-q_i)^{\gamma-1})^{1/\gamma}}\right)^{\gamma-1}-1\right]. \quad (19)$$

To study variability in conclusions across scoring rules, these families could potentially be used in a

manner similar to that of the beta family: we can first fix b at a baseline forecast of interest, just as we fixed c in the beta family. The γ parameter of these families is more difficult to set, because it does not have a simple interpretation (just as the beta family's $\alpha + \beta$ did not have a simple interpretation). However, we can still examine variability across values of γ.

To compare the use of these families to that of the beta family, we replicated Figure 5 using the two new families. In these replications, we allowed b to vary from 0 to 1, examining the Spearman correlation between rankings under the Brier score and rankings under rules in the families from (18) and (19). Results are shown in Figures 7 and 8. It is seen that the pseudospherical family results (Figure 8) are similar to the beta family results: once one fixes the parameter b, there is less variability in scores for different values of γ. However, the range of correlations is smaller (and closer to one) than the range of correlations under the beta family.

The power family results (Figure 7) differ from the other families, however. Under this family, the γ parameter has a larger impact on the correlation than the b parameter. This is because, as γ gets large, only extreme forecasts influence the rankings (nonextreme forecasts are all assigned the same score, regardless of the outcome). Additionally, at $\gamma = 2$, the parameter b has no impact on forecaster rankings. This is because,

Figure 7 **Spearman Correlation Between Forecaster Rankings Under the Brier Score and Forecaster Rankings Under Scores in the Baseline Power Family**

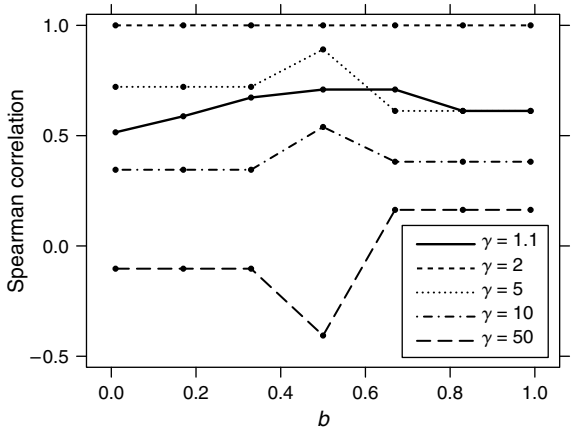
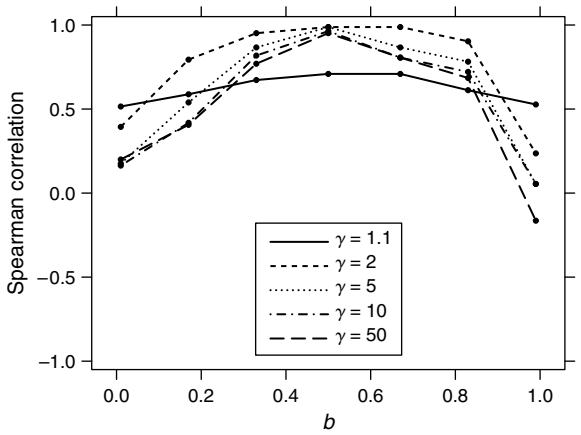


Figure 8 **Spearman Correlation Between Forecaster Rankings Under the Brier Score and Forecaster Rankings Under Scores in the Baseline Pseudospherical Family**



for $\gamma = 2$, Equation (18) reduces to

$$\frac{r_i^2 - 2r_i}{2q_i(1-q_i)} + \frac{1}{2(1-q_i)} + \frac{1}{2},$$

where the denominator of the first term is the same regardless of the outcome d_i, and the second term is constant across forecasters.

Because the power family's b parameter has little impact at large values of γ, one must adopt a modified strategy for choosing a scoring rule from the family. We suggest first setting b because it is more intuitive, as was done for the other families. In setting γ, then, one must decide whether the scoring rule should be sensitive to nonextreme forecasts (e.g., for $b = 0.5$, whether a forecast of 0.4 should receive a different score from a forecast of 0.6). If the scoring rule should be sensitive to these forecasts, then smaller values of γ (say, less than 20) are necessary. To choose a specific value of γ, it is probably necessary to create plots of specific scoring rules in a manner similar to Figure 3. These plots can be easily created using the scoring package that we described previously.

In addition to the power and pseudospherical families, Johnstone et al. (2011) propose a family of proper scoring rules that are tailored to decision-makers' utility functions. For example, in a betting context, a decision maker may be assumed to have a utility function associated with gains and losses in wealth. The Johnstone et al. family can be used to obtain

a proper scoring rule that reflects this utility function. Importantly, this family presumes that the decision maker can declare her utility associated with the actions she may take based on a forecast f. As the authors state, such a declaration is difficult or impossible in complex forecasting scenarios or in situations where multiple decision makers use the same forecast f. Thus, for the forecasts of world events considered here, this family would not be useful without strong assumptions about the forecasts' consumption.

Conclusions

As mentioned in the introduction, previous researchers have stated that different proper scoring rules lead to similar rankings of the assessors, at least when the rankings are based on average scores. Similarly, we find that different scoring rules in the beta family lead to similar rankings of forecasting methods, so long as our scoring rules are such that c remains approximately constant (focusing on the beta family). However, it is possible to find practically different rankings under beta family rules where $\alpha \neq \beta$; these are scoring rules that generally lie off the diagonal of graphs such as Figure 2 and that have differing cost parameters c. This, in turn, implies that it is insufficient to use a scoring rule simply because it is strictly proper; instead, it is beneficial to consider the specific way in which the scoring rule rewards and penalizes forecasts.

The beta family, or other two-parameter families such as the power or pseudospherical with baseline, can generally help analysts choose a scoring rule that suits their needs. This can be accomplished by first fixing a parameter that is interpreted as a cost of false positives (in the beta family case) or as the baseline forecast (in the power and pseudospherical cases). The second parameter can then be chosen by plotting the resulting scoring rule under multiple potential values, in a manner similar to Figure 3. Additionally, in the beta and pseudospherical cases, the analyst may be comforted by the fact that this second parameter has a smaller impact on forecaster rankings than the first parameter. Finally, in the absence of the need to choose a single scoring rule, one can easily visualize results across sets of scoring rules, as was displayed in Figures 2 and 4. These comparisons can provide the analyst with information about the types of forecasts that are (in)accurate and about the extent to which conclusions are robust.

This paper further shows that, although proper scoring rules encourage honest reporting from the forecaster, they place much less constraint on the individual who chooses the scoring rule. In particular, (i) choice of strictly proper scoring rule can have a large impact on one's results and conclusions, (ii) families of scoring rules can be used to evaluate forecasts more holistically, and (iii) it is possible to choose specific scoring rules from these families in a relatively intuitive fashion. Thus, forecast evaluators should routinely consider choosing scoring rules from these families that are tailored to the domain, as opposed to relying on popular, default scoring rules.

Computational Details

Results were obtained using the R system for statistical computing (R Development Core Team 2013), version 3.0.1. R is freely available under the General Public License 2 from the Comprehensive R Archive Network (CRAN) at http://CRAN.R-project.org/. To evaluate forecasts under the beta, power, and pseudospherical families of scoring rules (for binary outcomes), the R package scoring is also freely available under the General Public License 2 from CRAN.

Acknowledgments

The authors thank the associate editor and three anonymous reviewers for their insightful comments. 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 IARPA, DoI/NBC, or the U.S. government. This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI/NBC) [Contract Number D11PC20059]. The U.S. government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The work was also supported by a grant from the University of Missouri Research Board.

Appendix

Although we are unable to share the original data, Table A.1 includes artificial forecasts from the section titled "Comparing Forecasters." These forecasts mimic the distributional properties of the original forecasts. Forecasts of exactly 0

Table A.1 Artificial Forecasts Mimicking the Distributional Properties of the Forecasts from the "Comparing Forecasters" Section

	Forecaster ID										
Item	1	2	3	4	5	6	7	8	9	10	Outcome
1	0	0.44	0	0.26	0	0.43	0.55	0.09	0.14	0.09	0
2	0	0.41	0.3	0.74	0.41	0.59	0.51	0.04	0.02	0.48	0
3	0	0.99	0.21	0.94	0.26	0.03	0.19	1	0.04	0.5	0
4	0	0.41	0.35	0.89	0.89	0.79	0.52	0.91	0.04	0.59	1
5	0	0	0	0	0.1	0.3	0.4	0.1	0	0	0
6	0.11	0.22	1	0.05	0.14	0.3	0.64	0.02	0.09	0.11	1
7	0	0	0.3	0.85	0.05	0.06	0.18	0	0.17	0.41	0
8	0	0.71	0.19	0.17	0.05	0	0.45	0.19	0.25	0.41	0
9	0.1	0.04	0	0.01	0.35	0.03	0.54	0.07	0.19	0.49	0
10	0.75	0.84	0.39	0.97	0.75	1	0.46	0	1	0.76	0
11	0	0.57	0	0.01	0.09	0.09	0.45	0	0.19	0.16	0
12	0	0.03	0	0.22	0.84	0	0.5	0.01	0	0.76	0
13	0.26	0.75	0	0.24	0.01	0.06	0.4	0.05	0.18	0.75	0
14	0.09	0.02	0.21	0.75	0.16	0.02	0.34	0.03	0.21	0.91	0
15	0	0.01	0	0.53	0.04	0.01	0.25	0.06	0.24	0.84	0
16	0	0.05	0.14	0.4	0.14	0.01	0.49	0	0	0.09	0
17	0	0.03	0.11	0.37	0.3	0	0.5	0.06	0	0.8	0
18	0	0.05	0.09	0.25	0.54	0	0.46	0.74	0	0.14	0
19	0	0.47	0.85	0.18	0.75	0.01	0.44	0.08	1	0.75	1
20	0	0.02	1	0.58	0.76	0.97	0.53	0.95	0	0.76	0
21	0	0.06	0.09	0.28	0.19	0.01	0.48	0.06	0	0.14	0

and 1 were coded as 0.0001 and 0.9999, respectively, so that no forecaster could obtain a score of infinity.

References

Armstrong JS (2001) *Principles of Forecasting* (Kluwer Academic, Norwell, MA).

Bickel JE (2007) Some comparisons among quadratic, spherical, and logarithmic scoring rules. *Decision Anal.* 4:49–65.

Brier GW (1950) Verification of forecasts expressed in terms of probability. *Monthly Weather Rev.* 78:1–3.

Buja A, Stuetzle W, Shen Y (2005) Loss functions for binary class probability estimation and classification: Structure and applications. Accessed May 2, 2012, http://stat.wharton.upenn.edu/buja/PAPERS/.

Davis-Stober CP, Budescu DV, Dana J, Broomell SB (2013) When is a crowd wise? *Decision*. Forthcoming.

Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* 102:359–378.

Hand DJ, Vinciotti V (2003) Local versus global models for classification problems: Fitting models where it matters. *Amer. Statistician* 57:124–131.

Johnstone DJ (2007) The parimutuel Kelly probability scoring rule. *Decision Anal.* 4:66–75.

Johnstone DJ (2011) Economic interpretation of probabilities estimated by maximum likelihood or score. *Management Sci.* 57:308–314.

Johnstone DJ, Jose VRR, Winkler RL (2011) Tailored scoring rules for probabilities. *Decision Anal.* 8:256–268.

Jose VRR, Nau RF, Winkler RL (2008) Scoring rules, generalized entropy, and utility maximization. *Oper. Res.* 56:1146–1157.

Jose VRR, Nau RF, Winkler RL (2009) Sensitivity to distance and baseline distributions in forecast evaluation. *Management Sci.* 55:582–590.

Murphy AH (1972) Scalar and vector partitions of the probability score: Part I. Two-state situation. *J. Appl. Meteorology* 11:273–282.

O'Hagan A, Buck CE, Daneshkhah A, Eiser JR, Garthwaite PH, Jenkinson DJ, Oakley JE, Rakow T (2006) *Uncertain Judgements: Eliciting Experts' Probabilities* (Wiley, Hoboken, NJ).

R Development Core Team (2013) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

Schervish MJ (1989) A general method for comparing probability assessors. *Ann. Statist.* 17:1856–1879.

Selten R (1998) Axiomatic characterization of the quadratic scoring rule. *Experiment. Econom.* 1:43–62.

Staël von Holstein CAS (1970) Measurement of subjective probability. *Acta Psych.* 34:146–159.

Steyvers M, Lee MD, Miller BJ (2009) Wisdom of crowds in the recollection of order information. Bengio Y, Schuurmans D, Lafferty J, Williams CKI, Culotta A, eds. *Advances in Neural Information Processing Systems*, Vol. 22 (MIT Press, Cambridge, MA), 1785–1793.

Surowiecki J (2005) *The Wisdom of Crowds* (Anchor, New York).

Tetlock PE (2005) *Expert Political Judgment: How Good Is It? How Can We Know*? (Princeton University Press, Princeton, NJ).

Turner BM, Steyvers M, Merkle EC, Budescu DV, Wallsten TS (2013) Forecast aggregation via recalibration. *Machine Learn*. Forthcoming.

Warnaar DB, Merkle EC, Steyvers M, Wallsten TS, Stone ER, Budescu DV, Yates JF, et al. The aggregative contingent estimation system: Selecting, rewarding, and training experts in a wisdom of crowds approach to forecasting. Pantofaru C, Chernova S, Sorokin A, eds. *Proc. 2012 Assoc. Advancement of Artificial Intelligence Spring Sympos. Ser. (AAAI Tech. Rep. SS*-12-06*)* (AAAI Press, Palo Alto, CA), 75–76.

Winkler RL (1971) Probabilistic prediction: Some experimental results. *J. Amer. Statist. Assoc.* 66:675–685.

Winkler RL (1996) Scoring rules and the evaluation of probabilities. *Test* 5:1–26.

Winkler RL, Jose VRR (2010) Scoring rules. Cochran JJ, ed. *Wiley Encyclopedia of Operations Research and Management Science* (John Wiley & Sons, New York).

Yi SKM, Steyvers M, Lee MD, Dry M (2010) Wisdom of crowds in minimum spanning tree problems. Ohlsson S, Catrambone R, eds. *Proc. 32nd Annual Conf. Cognitive Sci. Soc.* (Lawrence Erlbaum Associates, Inc., Mahwah, NJ), 1840–1845.

Edgar C. Merkle is assistant professor in the Department of Psychological Sciences at the University of Missouri. He received a Ph.D. in quantitative psychology and an MS in statistics, both from The Ohio State University. His research interests include latent variable models, subjective probability and forecasts, and statistical computing. He has authored numerous journal articles within these areas.

Mark Steyvers is a professor of cognitive science at UC Irvine and is affiliated with the Computer Science Department as well as the Center for Machine Learning and Intelligent Systems. He is currently the president of the Society of Mathematical Psychology. He received his Ph.D. from Indiana University and was a postdoctoral fellow at Stanford University. His recent work is on developing Bayesian models for aggregating human judgments, including probability estimates, rankings, and problem-solving behavior. He also works on probabilistic topic modeling and text mining. In this research area, he has shown how to develop topic models to automatically extract high-level semantic representations from text documents. For his computational modeling work, Dr. Steyvers received the New Investigator Award from the American Psychological Association as well as the Society of Experimental Psychologists.