

Probability Elicitation, Scoring Rules, and Competition among Forecasters

Kenneth C. Lichtendahl, Jr.
Darden School of Business
University of Virginia
Charlottesville, VA 22906-6550
lichtendahlc@darden.virginia.edu

Robert L. Winkler
Fuqua School of Business
Duke University
Durham, NC 27708-0120
rwinkler@duke.edu

Abstract

Probability forecasters who are rewarded via a proper scoring rule may care not only about the score, but also about their performance relative to other forecasters. We model this type of preference and show that a competitive forecaster who wants to do better than another forecaster typically should report more extreme probabilities, exaggerating toward zero or one. We consider a competitive forecaster's best response to truthful reporting and also investigate equilibrium reporting functions in the case where another forecaster also cares about relative performance. We show how a decision maker can revise probabilities of an event after receiving reported probabilities from competitive forecasters and note that the strategy of exaggerating probabilities can make well-calibrated forecasters (and a decision maker who takes their reported probabilities at face value) appear to be overconfident. However, a decision maker who adjusts appropriately for the misrepresentation of probabilities by one or more forecasters can still be well-calibrated. Finally, to try to overcome the forecasters' competitive instincts and induce cooperative behavior, we develop the notion of joint scoring rules based on business sharing and show that these scoring rules are strictly proper.

Key Words: probability elicitation; scoring rules; forecasting competitions; probability forecasts; truthful revelation; overconfidence bias

February 27, 2007

1. Introduction

When Bayesian statistics and decision analysis are used to model inferential and decision-making problems, uncertainty is represented in probabilistic form. Often multiple experts are consulted and their probability forecasts for events of interest are elicited. Scoring rules, which involve the computation of a score based on the probability forecast and on the event that actually occurs, can be used in an ex post sense to evaluate the probabilities. For example, the Brier score (Brier 1950) is an early example of a scoring rule developed to evaluate probabilistic weather forecasts. Scoring rules can also be used in an ex ante sense, as part of the elicitation process, to provide an incentive for careful and honest forecasting (e.g., de Finetti 1962, Savage 1971). For a general review of scoring rules, see Winkler (1996).

The ex ante incentive aspect of scoring rules is based on the assumption that a forecaster's objective is to maximize expected score. This implies that the forecaster's utility is linear in the score, which is the case if, for example, the forecaster is given a monetary reward proportional to the score and is risk-neutral with respect to money. Of particular interest, then, are strictly proper scoring rules, for which the forecaster can maximize her expected score only by reporting her probabilities truthfully. Nonlinear utility can lead to nontruthful reporting under strictly proper scoring rules (Winkler 1969, Winkler and Murphy 1970). Moreover, further complications arise if the forecaster has any stakes in the event other than through the score received from the scoring rule (Kadane and Winkler 1988).

In this paper, we consider a particular type of "competitive" stakes that could affect the forecaster's reported probabilities. Suppose, for example, that a mutual fund manager hires two economists with expertise concerning a particular country to provide the probability that the country's central bank will increase interest rates or the probability that the central bank will devalue the country's currency before the end of the year. Each analyst has some incentive to provide "accurate" probabilities, and the manager uses a scoring rule to evaluate the accuracy of the analysts' probabilities relative to the truth at the end of the year. But the analysts may also feel that how they perform relative to each other is important in terms of being hired again by the mutual fund manager and in terms of reputation building more generally, so there is some utility associated with doing better than the other analyst as well as with

having accurate probabilities. Moreover, the presence of another forecaster can induce the spirit of a competition among the forecasters, whether such a competition is formally organized (e.g., via a prize given to the forecaster with the better score), implicit (e.g., if the forecasters believe that relative performance will be important in future decisions regarding promotions or layoffs), or simply driven by the forecasters' innate competitiveness.

Such competitiveness is likely to be encountered in a variety of situations. The systems underlying much of our global economy are based on competition, as are many of our leisure activities (e.g., spectator sports), and disparities in income have increased in recent years, notably at the upper end of the income distribution. Frank and Cook (1995) use the term "winner-take-all-society" and claim (p. 7) "that the phenomenon has spread so widely and that so many of the top prizes have become so spectacular." Within the realm of forecasting, Tetlock (2005), who obtained probability forecasts from experts on political and economic issues and used scoring rules, discusses the notion that experts can be seduced by fame, fortune, and power. In response to one expert's comment, "I fight to preserve my reputation in a cutthroat adversarial culture," Tetlock notes (p. 186), "In his world, only the overconfident survive and only the truly arrogant thrive" and points out that another expert "stressed the need – if one wants to be remembered – to make bold claims that run against the grain of the conventional wisdom." We suggest that this sort of phenomenon is not uncommon in many arenas, such as the investment advice business (where "gurus" who make lists of "top stock pickers" in publications such as *The Wall Street Journal* might anticipate fame, fortune, and power), forecasting sporting events, and forecasting the next "hot trend."

Previous research on scoring rules has focused on individual forecasters and has not taken possible competition among probability forecasters into consideration, with one notable exception. Kilgour and Gerchak (2004) introduce the notion of competitive scoring rules, in which payments based on scoring rules depend on the scores of other forecasters as well as a forecaster's own score. This formalizes the competitive nature of the forecasting situation, assuming risk-neutrality and channeling everything through the relative nature of the payments. In contrast, we separate out the forecaster's

preferences regarding relative rank among the forecasters (which may be related to implicit rewards such as improved future opportunities as well as any intrinsic satisfaction or dissatisfaction associated with beating or being beaten by other forecasters) from any payment related to the forecaster's score itself. An elicitation method developed by Prelec (2004) involves eliciting not just a forecaster's probabilities for the events of interest, but also the forecaster's expected distribution of probabilities given by other forecasters. The resulting score depends on the probabilities and the events that occur, as do typical scoring rules, and on the expected and actual distributions of probabilities. However, there is no competitive aspect to the process, although thinking about other forecasters' probabilities could stimulate a feeling of competitiveness. In a different setting related to rank-order contests, Ottaviani and Sorensen (2005) consider point forecasts instead of probability forecasts and assume that only relative performance matters, although they briefly mention the situation we address here: "The statistical literature on scoring rules and probability forecasting is normative in nature, but has focused on the case of single forecasters, disregarding the possible strategic elements arising with multiple forecasters" (pp. 20-21).

The paper is organized as follows. In §2, we consider the situation where two forecasters report probabilities for an event E . We present a model in which a forecaster's utility is additive in relative performance and score and the forecaster's beliefs about the probability of another forecaster are represented probabilistically. In §3, we investigate a competitive forecaster's optimal reporting strategies in response to truthful reporting by another forecaster. The forecaster typically should report more extreme probabilities, exaggerating toward zero or one. We then consider the case where both forecasters are competitive in §4, modeling the forecasting competition as a game of incomplete information, analyzing the forecasters' decisions simultaneously, and finding equilibrium reporting functions. As in §3, the forecasters should exaggerate their forecasts toward zero or one. In §5, we look at the situation from the perspective of a decision maker revising probabilities after receiving forecasts and note potential connections with observed overconfidence on the part of forecasters. Then we show how the decision maker might attempt to overcome competitive tendencies by inducing cooperative behavior through a business-sharing game in §6. Finally, we briefly summarize our results in §7.

2. A Model with Competitive Forecasters

Suppose that we are interested in an event E . We consult two forecasters and elicit the probability of E from each of them, rewarding forecaster i ($i = 1, 2$) via a score S_i from a strictly proper scoring rule S that yields $s_E(\beta_i)$ if E occurs and $s_{E'}(\beta_i)$ if its complement E' occurs, where β_i is forecaster i 's reported probability of E . Each forecaster is aware of the scoring rule that is used. Letting p_i represent forecaster i 's probability for E (i.e., i 's beliefs about the chance that E will occur), we represent i 's reporting strategy in the form $\beta_i = b_i(p_i)$. The scoring rule is normalized so that $0 \leq S_i \leq 1$, where a higher score is better. For our examples, we will use a normalized quadratic score with

$$s_E(\beta_i) = 1 - (1 - \beta_i)^2 \text{ and } s_{E'}(\beta_i) = 1 - \beta_i^2, \quad (1)$$

but any other normalized strictly proper scoring rule could equally well be used.

Forecasters' Utilities. We assume that forecaster i 's preferences can be represented by a utility function that is additive in relative performance R_i and score S_i ,

$$u_i(R_i, S_i) = w_i R_i + (1 - w_i) S_i, \quad (2)$$

where $0 \leq w_i \leq 1$ and $R_i = 0(0.5)1$ if $S_i < (=) > S_j$, $j \neq i$ (i.e., if i 's score is worse than, equal to, or better than the score of the other forecaster). Here w_i represents forecaster i 's preference tradeoff between relative performance and score. Note that the relative performance term in (2) does not depend on the specific scoring rule that is used, since $s_E(s_{E'})$ is strictly increasing (decreasing) for any strictly proper scoring rule (Schervish 1989). Thus, if $E(E')$ occurs, the forecaster with the higher (lower) β_i has the better performance, with $R_i = 1$. Only the second term on the right-hand side of (2) depends on the specific scoring rule. Note also that $R_i = 0.5$ in the case of a tie, which is consistent with the usual tie-breaking rule that gives each forecaster equal probability of "winning" or "losing."

The utility function in (2) has a natural discontinuity through the relative performance term, which can be justified by considerations such as career concerns (i.e., promotion to a better job or keeping one's current job). As athletes often say after a close win, "What matters isn't how much we won by, it's

that we won.” Contests with significant “prizes” have such discontinuities, not only because of the immediate rewards but also because of anticipated future benefits (Gaba, Tsetlin, and Winkler 2004). Savage (1971, pp. 798-799) notes that the scoring-rule incentive to report honestly “may not be fully in harmony with the expert’s incentive to appear worthy, as opposed to simply being worthy, of retention and promotion – a complication that affects ... all applications of proper scoring rules to a professional expert.” In any event, the absolute performance term is a continuous function of β_i , and the impact of the discontinuity in the relative performance term depends on the size of w_i , which in turn reflects the importance of the competitive nature of the situation to the forecaster.

Forecasters’ Conditional Beliefs. Forecaster i ($i = 1, 2$) knows her own probability p_i . She does not know the other forecaster’s probability but has a distribution for that probability. Suppose that forecaster i ’s conditional beliefs about p_j ($i = 1, 2, j \neq i$) given p_i are represented by continuous cumulative distribution functions (cdfs) $F_{p_j|E, p_i}$ and $F_{p_j|E', p_i}$, with corresponding density functions (pdfs) $f_{p_j|E, p_i}$ and $f_{p_j|E', p_i}$. For example, forecaster i might consider p_j to be conditionally independent of p_i , given E or E' , with beta distributions

$$f_{p_j|E, p_i} = f_{Be}(p_j | c, d) \text{ and } f_{p_j|E', p_i} = f_{Be}(p_j | c', d'), \quad (3)$$

where $c, d, c', d' > 0$ and $f_{Be}(p | y, z) \propto p^{y-1}(1-p)^{z-1}$ for $0 \leq p \leq 1$. Morris (1974) used beta distributions to represent beliefs about experts’ probabilities in a Bayesian expert combination model. Note that the distributions in (3) are distributions for the other forecaster’s probabilities (her beliefs), not distributions for her *reported* probabilities, which can differ from her beliefs.

3. A Competitive Forecaster’s Best Response to Truthful Reporting

When forecaster 2 doesn’t care about relative performance, $w_2 = 0$ and forecaster 2’s utility simply equals her score. As a result, her beliefs about forecaster 1 are not relevant, and her strategy is identical to what it would be if she were the only forecaster. Since the scoring rule is strictly proper, her optimal strategy is truthful reporting: $\beta_2 = p_2$. Note that this situation would also apply if forecaster 2 is

a model rather than a human. In this section we analyze forecaster 1's optimal response to truthful reporting, taking into account forecaster 1's utility function and beliefs about p_2 . Then, in §4, we will drop the assumption of truthful reporting for forecaster 2 and shift to a game-theoretic approach in order to analyze the reporting decisions of the forecasters simultaneously.

Note that the continuous cdfs and the assumption that $w_2 = 0$ imply that forecaster 1 believes that a tie has probability zero. Proposition 3.1 shows that if forecaster 1 cares about relative performance, the best response to truthful reporting is not necessarily to report truthfully in return. Proofs of all propositions are given in the Appendix.

Proposition 3.1. If $0 < w_1 < 1$, $0 < p_1 < 1$, and forecaster 1 believes that forecaster 2 will report truthfully, truthful reporting satisfies the first-order condition to be forecaster 1's best response if and only if

$$f_{p_2|E, p_1}(p_1) / f_{p_2|E, p_1}(p_1) = p_1 / (1 - p_1). \quad (4)$$

The point of Proposition 3.1 is to show that a very specific condition is necessary for truthful reporting to be a best response to the other forecaster's truthful reporting. It is possible that (4) will hold in some situations, but it is quite restrictive and not intuitively compelling. For instance, the conditional beliefs $f_{p_2|E, p_1}(p_2) = 2(1 - p_2)$ and $f_{p_2|E, p_1}(p_2) = 2p_2$ satisfy (4) but are the reverse of what we'd expect to hear from an expert (e.g., conditional on E , lower values of p_2 are more likely). We expect that truthful reporting is seldom forecaster 1's best response.

Example 3.1. Consider the normalized quadratic scoring rule in (1), and suppose that forecaster 1's conditional beliefs about p_2 are beta with $(c, d, c', d') = (2, 1, 1, 2)$. The first-order condition for maximizing expected utility is $\beta_1 = (p_1 - w_1) / (1 - 2w_1)$. If p_1 and w_1 are such that this $\beta_1 \notin [0, 1]$ or the second-order condition ($2w_1 - 1 \leq 0$) does not hold, then the optimal β_1 is zero or one. Thus, when $w_1 < 0.5$,

$$\beta_1 = \begin{cases} 0 & \text{if } 0 \leq p_1 < p_1^* \\ \alpha_1 p_1 + \alpha_0 & \text{if } p_1^* \leq p_1 \leq 1 - p_1^* \\ 1 & \text{if } 1 - p_1^* < p_1 \leq 1, \end{cases} \quad (5)$$

where $\alpha_1 = 1/(1 - 2w_1)$, $\alpha_0 = (1 - \alpha_1)/2 = -w_1/(1 - 2w_1)$, and $p_1^* = -\alpha_0/\alpha_1 = w_1$. When $w_1 \geq 0.5$, extreme reporting is a best response:

$$\beta_1 = \begin{cases} 0 & \text{if } p_1 < 0.5 \\ 0 \text{ or } 1 & \text{if } p_1 = 0.5 \\ 1 & \text{if } p_1 > 0.5. \end{cases} \quad (6)$$

Thus, as long as $w_1 > 0$, the reported probability from (5) and (6) is often equal to zero or one and is always closer to zero or one than p_1 unless $p_1 = 0.5$ or p_1 is itself zero or one.

Figure 1a shows $\beta_1 = b_1(p_1)$ from (5) for $w_1 = 0.3$. As $w_1 \rightarrow 0$, β_1 approaches truthful reporting, as would be expected. As $w_1 \rightarrow 0.5$, the horizontal segments with $\beta_1 = 0$ or 1 are longer and the line from $(p_1^*, 0)$ to $(1 - p_1^*, 1)$ is steeper, approaching the extreme reporting of (6), which is shown in Figure 1b. These results suggest that caring more about relative performance, as reflected by a higher weight w_1 , should lead to more extreme probabilities.

The intuitive explanation for the sort of strategy illustrated by Example 3.1 (and the examples to come in §4) is that forecaster 1 is willing to give up something in terms of expected score by deviating from her beliefs and reporting a higher probability for the more likely event in order to have a better chance of winning. She is making a tradeoff between the two terms in her utility function, and the larger w_1 is, the more she is willing to trade off on the score to increase her chances on the relative performance term.

4. A Competitive Forecaster's Best Response to a Strategic Competitor

In §3, we modeled the forecasting situation from the viewpoint of forecaster 1, taking into account forecaster 1's beliefs about the probability and the utility function of forecaster 2. Said differently, we used decision analysis to model forecaster 1's reporting decision. Now we turn to a model

that uses game theory (or what might be called mutually consistent decision analysis) to analyze both forecasters' decisions simultaneously, seeking a Bayesian Nash equilibrium (Fudenberg and Tirole 1991). As we have seen in §3, truthful reporting is typically not a best response to truthful reporting if nonzero weight is put on relative performance, so it is generally not a symmetric equilibrium unless $w_1 = w_2 = 0$. We focus on a symmetric two-forecaster situation in which each forecaster has the same utility function and has the same beliefs about the other forecaster's probability:

Utility: Each forecaster's utility function is of the form (2) with $w_i = w$, $i = 1, 2$.

Scoring Rule: The scoring rule is strictly proper.

Conditional Beliefs: The cdfs $F_{p_j|E, p_i}$ and $F_{p_i|E, p_j}$ of each forecaster for the other forecaster's private information are continuous in p_j for $0 \leq p_j \leq 1$, $i = 1, 2$, $j \neq i$, and symmetric in the sense that

$$F_{p_2|E, p_1}(p_2) = F_{p_1|E, p_2}(p_1) \text{ and } F_{p_2|E, p_1}(p_2) = F_{p_1|E, p_2}(p_1) \text{ for } 0 \leq p_1, p_2 \leq 1.$$

The justification for the symmetry of conditional beliefs is one of exchangeability: The forecasters are experts and as a first cut might be viewed (and might view each other) as roughly exchangeable. Proposition 4.1 and Example 4.1 invoke symmetry of utilities and conditional beliefs, but these assumptions are relaxed in Examples 4.2 and 4.3, respectively.

Proposition 4.1. In this forecasting game, any strictly increasing, differentiable piece of a nondecreasing Bayesian Nash equilibrium reporting function $b(p_i)$ must satisfy the following differential equation:

$$w[p_i f_{p_j|E, p_i}(p_i) - (1 - p_i) f_{p_j|E, p_i}(p_i)] + (1 - w)[p_i s_E'(b(p_i)) + (1 - p_i) s_E'(b(p_i))] b'(p_i) = 0. \quad (7)$$

Example 4.1. Suppose that the scoring rule is quadratic, as given in (1), and the forecasters' beliefs about their opponent's beliefs can be represented by beta distributions, as in (3). Applying (7) yields the ordinary, nonlinear, first-order differential equation

$$b'(p_i) = \frac{w[p_i f_{Be}(p_i | c, d) - (1 - p_i) f_{Be}(p_i | c', d')]}{2(1 - w)[b(p)_i - p_i]}. \quad (8)$$

The solutions of (8) for $w = 0.3$ and $(c, d, c', d') = (2, 1, 1, 2)$ and $(3, 1, 2, 2)$ are shown in Figure 2. The

strictly increasing pieces of Bayesian Nash equilibria are related to these solutions because they are consistent with beta common priors, which will be discussed in §5. If $(c, d, c', d') = (2, 1, 1, 2)$, the equilibrium reporting function (which we call a \hat{p} equilibrium) when $w < 2/3$ is of the form

$$b(p_i) = \begin{cases} 0 & \text{if } 0 \leq p_i < \hat{p} \\ \alpha_1 p_i + \alpha_0 & \text{if } \hat{p} \leq p_i \leq 1 - \hat{p} \\ 1 & \text{if } 1 - \hat{p} < p_i \leq 1 \end{cases} \quad (9)$$

where $\alpha_1 = (1 + \sqrt{1 + 8[w/(1-w)]})/2$, $\alpha_2 = (1 - \alpha_1)/2$, and $\hat{p} = 0.2893$ is the root of a quadratic equation.

When $w \geq 2/3$, we get the extreme reporting in (6). The reporting function from (9) for $w = 0.3$ is shown in Figure 3 along with the best response to truthful reporting in this case. The general nature of the two strategies is similar in the sense that they both imply that $b(p_i)$ should be closer to zero or one than p_i , with the equilibrium strategy yielding slightly less extreme reporting and “jumps” from 0 to $\hat{\beta} = b(\hat{p}) = 0.173$ at \hat{p} and from $1 - \hat{\beta} = 0.827$ to 1 at $1 - \hat{p}$. The “jumps” at \hat{p} and $1 - \hat{p}$ arise from comparisons of the expected utilities of reporting along the solution to (8) and reporting at the extremes.

Although the symmetric model seems reasonable when we are dealing with true experts, it is useful to have some idea of what might happen when we don’t have symmetry. As in the analysis of asymmetric auctions (Marshall et al. 1994, Lebrun 1999), extending the analysis to the asymmetric case is straightforward in principle but can be tedious numerically, entailing the solution of systems of differential equations. However, Examples 4.2 and 4.3 illustrate two types of asymmetry. In each case, the asymmetry leads to a system of two differential equations (one for each forecaster) similar to (8) and to different reporting functions for the two forecasters.

Example 4.2. Here we consider the conditions of Example 4.1 but allow $w_1 \neq w_2$, reflecting the notion that the forecasters may differ in the importance they place on the competitive aspect associated with relative performance. The resulting equilibrium when $w_1 = 0.25$, $w_2 = 0.5$, and $(c, d, c', d') = (2, 1, 1, 2)$ is shown in Figure 4a. The increasing portions of the reporting functions are now nonlinear. As expected, forecaster 2, who puts more weight on the relative performance term, gives more extreme reports for

intermediate values of p (i.e., values of p such that $\hat{p}_1 \leq p \leq 1 - \hat{p}_1$). Note that the two forecasters jump to extreme forecasts of 0 or 1 at different probabilities, since $\hat{p}_1 = 0.274$ and $\hat{p}_2 = 0.404$, but for each forecaster the magnitude of both “jumps” is $\hat{\beta}_i = b_i(\hat{p}_i) = 0.209$.

Example 4.3. Now we modify the beliefs in Example 4.1 instead of the weights. Figure 4b shows the equilibrium when beliefs about p_1 and p_2 are beta with $(c, d, c', d') = (1.25, 0.25, 0.25, 1.25)$ and $(2, 1, 1, 2)$, respectively, and $w_1 = w_2 = 0.25$. Forecaster 1 has more “expertise” in the sense that the conditional distributions of p_1 almost dominate those of p_2 and have means of 0.83 given E and 0.17 given E' (as compared with 0.67 and 0.33 for p_2) with identical variances for all four distributions. The increasing portions of the reporting functions cross, with forecaster 1 giving more extreme reports than forecaster 2 for p near 0.5 and less extreme reports when $\hat{p}_1 = 0.261 \leq p < 0.293 = \hat{p}_2$ or $0.707 < p \leq 0.739$. The two reporting functions, however, are quite close to each other and are also quite similar to those with symmetric beta $(2, 1, 1, 2)$ beliefs. It is hard to generalize from examples, but the asymmetric weights in Example 4.2 lead to greater differences in reporting functions than the asymmetric beliefs in Example 4.3.

In our model of competing forecasters, the utility functions and conditional beliefs are common knowledge, which is standard practice in game theory to enable us to find equilibria. Such common knowledge is a strong assumption, but it is a “convenient fiction” that gives us an idea of what can happen if forecasters’ have some idea of each others’ abilities and preferences. Of course, each forecaster also has private information: her own p_i , which is not known to the other forecaster. This is similar to models of auctions, where bidders have private estimates of value, distributions of others’ estimates (but not the others’ estimates themselves) are common knowledge, bids are submitted, and the high bid wins.

5. Revising after Seeing Reports from Competitive Forecasters

Suppose that a decision maker uses a proper scoring rule to elicit probabilities from forecasters and revises his own probability based on the forecasters’ reports. To allow for this, we extend the model of forecasters’ beliefs to include the decision maker’s initial probability p_0 for E and consider a common

prior with joint cdf

$$F(p_1, p_2, x) = \begin{cases} 0 & \text{if } x < 0 \\ p_0 F_{p_1, p_2 | E}(p_1, p_2) & \text{if } 0 \leq x < 1 \\ (1 - p_0) F_{p_1, p_2 | E'}(p_1, p_2) & \text{if } x \geq 1 \end{cases} \quad (10)$$

for $0 \leq p_1, p_2 \leq 1$, where $x = 1$ if E occurs and 0 otherwise, and $P(E | p_i) = p_i$ for $i = 1, 2$. For example, we will use a “beta common prior” for exchangeable forecasters denoted $BeCP(c, d, c', d')$, having beliefs as in (3) and joint cdf

$$F(p_1, p_2, x) = \begin{cases} 0 & \text{if } x < 0 \\ p_0 F_{Be}(p_1 | c, d) F_{Be}(p_2 | c, d) & \text{if } 0 \leq x < 1 \text{ for } 0 \leq p_1, p_2 \leq 1 \\ (1 - p_0) F_{Be}(p_1 | c', d') F_{Be}(p_2 | c', d') & \text{if } x \geq 1. \end{cases} \quad (11)$$

If the decision maker consults only forecaster i and believes that the forecaster reports truthfully (e.g., $w_i = 0$), then the decision maker’s revised probability is

$$P(E | \beta_i) = P(E | p_i) = \frac{p_0 f_{p_i | E}(p_i)}{p_0 f_{p_i | E}(p_i) + (1 - p_0) f_{p_i | E'}(p_i)}. \quad (12)$$

From the decision maker’s perspective, forecaster i is an expert in the sense of Dawid, DeGroot, and Mortera (1995) because $P(E | p_i) = p_i$. Under $BeCP(c, d, c', d')$, $P(E | p_i) = p_i$ for $i = 1, 2$ if and only if $c' = c - 1$, $d' = d + 1$, and $p_0 = (c - 1)/(c + d - 1)$.

When reports from forecasters 1 and 2 are received and they both report truthfully,

$$P(E | \beta_1, \beta_2) = P(E | p_1, p_2) = \frac{p_0 f_{p_1, p_2 | E}(p_1, p_2)}{p_0 f_{p_1, p_2 | E}(p_1, p_2) + (1 - p_0) f_{p_1, p_2 | E'}(p_1, p_2)}. \quad (13)$$

As demonstrated in §4, competitive forecasters may not report truthfully, in which case the likelihoods in (12) and (13) are functions of the β_i instead of the p_i , thereby involving the decision maker’s beliefs about the nature of the forecasters’ misrepresentation of their probabilities as well as about the probabilities themselves.

Example 5.1. Consider the normalized quadratic scoring rule in (1) and a beta common prior

$BeCP(2, 1, 1, 2)$ with $p_0 = 0.5$. Under these conditions, if the decision maker believes that $w_i = 0$, then

$P(E | \beta_i) = P(E | p_i) = p_i$. However, if the decision maker believes that $w_1 = w_2 = 0.3$, then $P(E | \beta_i)$ is as given in Figure 5a. For $\hat{\beta} \leq \beta_i \leq 1 - \hat{\beta}$, where $\hat{\beta} = b(\hat{p}) = 0.173$, $P(E | \beta_i)$ is simply the linear portion of Figure 3 with the axes reversed. $P(E | \beta_i = 0)$ and $P(E | \beta_i = 1)$ reflect $0 \leq p_i < \hat{p}$ and $1 - \hat{p} < p_i \leq 1$, respectively. The decision maker's revised probability $P(E | \beta_1, \beta_2)$ after seeing both reports is given in Figure 5b. This is a smooth surface when $\hat{\beta} \leq \beta_i \leq 1 - \hat{\beta}$ for $i = 1, 2$ and is otherwise on the edge of the unit cube.

Example 5.1 illustrates how a decision maker can revise his probability of an event E after seeing probabilities reported by one or more experts. If forecasters care about their relative performance, the revision process is more complex, involving considerations of how this concern for relative performance will cause forecasters to misrepresent their probabilities. When the concern takes the form of positive increments in utility from outperforming other forecasters, the misrepresentation will generally be an exaggeration of probabilities toward 0 or 1, in which case the decision maker's revision process reacts to that by yielding revised probabilities that are closer to 0.5 than the probabilities reported by the experts.

These results have implications for the calibration of reported forecasts. If a forecaster is well-calibrated and reports truthfully, we expect her reports to be well-calibrated empirically. If she exaggerates her reported probabilities toward 0 or 1, however, we would expect her calibration curve to look more like Figure 5a, which resembles a calibration curve exhibiting overconfidence. This suggests that observed overconfidence on the part of forecasters (e.g., O'Hagan et al. 2006) need not always reflect a *cognitive* bias (the most common attribution) or a *statistical* phenomenon such as the regression-to-the-mean effect (Pfeifer 1994), but could be caused by a *motivational* bias. If forecasters' probabilities are well-calibrated but they care about their relative performance, then it is perfectly rational under the expected utility model for them to report exaggerated probabilities, in which case they will appear to be overconfident. Even if no other forecasters are reporting probabilities, a forecaster might exaggerate her reported probabilities because she is trying to reach an aspiration level (in a sense, competing with herself) or trying to impress a decision maker in the hope of being hired to give future forecasts.

In turn, the results have implications for the calibration of the decision maker's revised probabilities. If reports from well-calibrated forecasters who care about relative performance and exaggerate their reported probabilities are incorrectly assumed by the decision maker to be truthful, then the decision maker's revised probabilities will be overconfident instead of well-calibrated. On the other hand, proper allowance for the forecasters' exaggeration can correct the overconfidence and eliminate miscalibration. Of course, if the decision maker believes that the forecasters are overconfident to begin with due to cognitive bias and then also exaggerate, the decision maker should incorporate his beliefs about the forecasters' overconfidence into his model. Moreover, the forecasters themselves might have similar views about other forecasters, thinking they will exhibit an overconfidence bias; such views can be reflected in the conditional beliefs.

6. Inducing Cooperative Behavior: Joint Scoring Rules

As we have seen, competition among forecasters can lead to nontruthful reporting that can introduce distortion into the reporting process, thus necessitating more complicated modeling for a decision maker trying to revise his probabilities after seeing forecasters' reports. One response is to attempt to overcome the forecasters' "competitive stakes" by inducing cooperative behavior. Scoring rules need not be imposed directly. They may already be endogenous to some degree, stemming from the forecasters' realization that good outcomes for the decision maker can be good for the forecasters as well. To sharpen the effect of such cooperative stakes and to reduce the effect of other stakes (such as the competitive stakes considered in previous sections), we formalize a business-sharing game. This is in the spirit of McCarthy (1956) and Savage (1971), although they considered only a single forecaster. Here we have a dynamic game of incomplete information because the forecasters will give their reported probabilities and the decision maker will then choose an action. Therefore, the appropriate equilibrium concept is a perfect Bayesian equilibrium (Fudenberg and Tirole 1991).

Suppose that the decision maker is choosing between two actions, a_1 and a_2 . For example, these might correspond to "go" or "no go" on a project (investing in a project, deciding whether to undergo

surgery, etc.). The consequences depend on the action that is taken and on whether or not E occurs. The decision maker consults two forecasters and sets up a business-sharing game with the following assumptions:

Utility: The decision maker and the forecasters have a common understanding of the consequences and have equivalent normalized utilities with $\max\{u(a_1, E), u(a_2, E')\} = 1$ and $\min\{u(a_1, E'), u(a_2, E)\} = 0$, where $u(a_1, E) > u(a_2, E)$ and $u(a_2, E') > u(a_1, E')$ without loss of generality.

Beliefs: The decision maker and the forecasters have a common prior of the form (10) with continuous $F_{p_1, p_2|E}(p_1, p_2) = F_{p_1|E}(p_1)F_{p_2|E}(p_2)$ and $F_{p_1, p_2|E'}(p_1, p_2) = F_{p_1|E'}(p_1)F_{p_2|E'}(p_2)$.

Proposition 6.1. In this business-sharing game, truthful reporting by each forecaster, probability revision by the decision maker assuming such truthful reporting, and the choice of an optimal action based on the revised probabilities constitute a perfect Bayesian equilibrium. Furthermore, this equilibrium's payoffs to the forecasters can be implemented as a joint scoring rule $S = \{S_1, S_2\}$ with strictly proper score S_i for forecaster i :

$$s_{i,E}(\beta_i) = u(a_1, E) - F_{p_j|E} \left(\frac{1 - \beta_i}{1 + (k^* - 1)\beta_i} \right) [u(a_1, E) - u(a_2, E)] \quad (14)$$

and $s_{i,E'}(\beta_i) = u(a_1, E') - F_{p_j|E'} \left(\frac{1 - \beta_i}{1 + (k^* - 1)\beta_i} \right) [u(a_1, E') - u(a_2, E')]$,

where $j \neq i$ and $k^* = (1 - p_0)[u(a_1, E) - u(a_2, E)] / \{p_0[u(a_2, E') - u(a_1, E')]\}$.

The fact that truthful reporting is an equilibrium in this situation is not particularly surprising because the forecasters' rewards depend on the decision, which in turn depends on the reported probabilities. What is somewhat surprising, however, is that this equilibrium is implementable through a set of proper scoring rules for the forecasters. Since the generation of these proper scoring rules is related to each forecaster's beliefs about the other forecaster's probability, we call this set of interdependent scoring rules a *joint* scoring rule, and this notion extends to more than two forecasters and more events. Note that the forecasters' beliefs about the other forecaster's probability need not be identical, so the

forecasters need not face the same scoring rule. As the following example illustrates, commonly used scoring rules such as the quadratic rule can be generated from the business-sharing game.

Example 6.1. Suppose that $u(a_1, E) = 1$, $u(a_1, E') = 0$, $u(a_2, E) = u(a_2, E') = 0.5$, and the beliefs are represented by a beta common prior $BeCP(2,1,1,2)$ with $p_0 = 0.5$. From Proposition 6.1, the joint scoring rule is quadratic:

$$s_{i,E}(\beta_i) = 1 - [(1 - \beta_i)^2 / 2] \text{ and } s_{i,E'}(\beta_i) = (1 - \beta_i^2) / 2 \quad (15)$$

for $i = 1, 2$. The scores $s_{i,E}$ and $s_{i,E'}$ and the expected score under truthful reporting are shown as functions of β_i in Figure 6a. This strictly proper quadratic score is different from the normalized quadratic rule in (1), which itself is generated if $u(a_1, E) = u(a_2, E') = 1$ and $u(a_1, E') = u(a_2, E) = 0$. The scores $s_{i,E}$ and $s_{i,E'}$ and the expected score under truthful reporting for this case are shown in Figure 6b. Interestingly, although both situations generate strictly proper quadratic scores, the nature of the scores and expected scores are quite different. In the first case (Figure 6a), the consequence from a_1 is a utility of either 0 or 1, whereas a_2 yields a sure utility of 0.5. As a result, the expected score under truthful reporting is not symmetric about 0.5; it increases from 0.5 at $\beta_i = 0$ to 1 at $\beta_i = 1$. In the second case (Figure 6b), each action yields a utility of either 0 or 1, leading to an expected score function that is minimized at $\beta_i = 0.5$ and increases symmetrically to 1 at $\beta_i = 0$ or 1. The difference between the symmetric and asymmetric expected score functions is driven by the nature of the underlying decision problems.

7. Summary and Discussion

In decision analysis and risk analysis, probability forecasts are elicited from forecasters and scoring rules can be used as part of the elicitation process to provide an incentive for truthful reporting. Forecasters often may be competitive in the sense of caring not just about their own scores, but also about doing better than other forecasters. We develop a model of this situation and show that competitive forecasters should exaggerate their reported probabilities toward 0 or 1, sometimes going so far as to give

extreme forecasts that are *always* 0 or 1. Forecasters who are more competitive (those placing greater weight on relative performance) tend to exaggerate more. These results are quite robust, holding for a competitive forecaster competing with a forecaster who is assumed to report truthfully (possibly a model as well as a human), for one competing with a forecaster who is competitive herself, and for the symmetric case of exchangeable forecasters as well as the contrasting asymmetric case. In the interest of parsimony, our propositions are limited to a single event and its complement, and our illustrative examples assume quadratic scoring rules and beliefs about forecasters' probabilities represented by beta distributions. However, similar results hold for more events, more forecasters, other scoring rules, and other conditional beliefs in many cases that we have investigated (e.g., 3 events, 3-5 forecasters, a logarithmic scoring rule, logistic normal conditional beliefs).

We feel that these results have important implications for decision makers using probabilities reported by forecasters. If a decision maker takes a forecaster's reported probabilities at face value, assuming that the forecaster is reporting truthfully when in fact she is exaggerating her probabilities toward 0 or 1, the decision maker will be miscalibrated. Even if the forecaster is well-calibrated, exaggeration will cause her reported probabilities to exhibit overconfidence and can, in turn, cause the decision maker to be overconfident also. We suggest that the exaggeration of probabilities, which is a rational strategy if one wants to do better than other forecasters, could partially explain observed overconfidence on the part of forecasters and decision makers who use their forecasts. By appropriately adjusting for such behavior by one forecaster or multiple forecasters, as discussed in §5, the decision maker can still be well-calibrated.

This type of hedging in terms of reporting probabilities is part of a bigger picture. It is another example of how other stakes can create difficulties when we try to interpret information from forecasters or other experts (Kadane and Winkler 1988). Of course, the nature of the misrepresentation of probabilities depends on the nature of the other stakes. For example, a forecaster whose primary worry is about being at or near the bottom of a set of competitors (i.e., who is more concerned about being outperformed by other forecasters than about outperforming them) is more likely to hedge in the opposite

direction, toward 0.5, and to appear underconfident. Such a forecaster is willing to trade off some expected score to reduce the chance of being caught with a low probability for the event that eventually occurs. Some analyses assuming forecasters want to avoid finishing last (e.g., five forecasters with a quadratic scoring rule and a beta common prior) yield this type of behavior, which is consistent with strategies in contests with many winners and few losers (Gaba, Tsetlin, and Winkler 2004). More generally, with multiple forecasters the relative performance term in the forecasters' utility functions could take on many different values corresponding to the ranks a forecaster could achieve among the set of forecasters. Utilities for those ranks could reflect risk attitude toward rank, and probabilities of achieving various ranks would be related to distributions of order statistics, making a game-theoretic analysis difficult. The larger message here is that understanding these other stakes and their implications is important, and such stakes call for appropriate adjustments by users of the forecasts.

The process of updating appropriately in response to reports that are thought to be nontruthful because of competitive strategies or different stakes can improve a decision maker's revised probabilities but can be complicated. Another approach is to minimize the impact of competitive instincts by fostering cooperation among forecasters. We show how this could be implemented through a business-sharing process that highlights shared interests and leads to strictly proper joint scoring rules for the forecasters. These rules are not imposed by the decision maker; they are driven by the nature of the underlying decision problem. The results may not always look like commonly used scoring rules, although they serve the same purpose. An example in §6, however, yields quadratic scores. The managerial implications are that it is important for a decision maker to be aware of competitive stakes and that joint scoring rules can introduce or enhance incentives for forecasters to be cooperative and report truthfully. More generally, anything that can sharpen cooperative stakes and reduce competitive stakes should encourage truthful reporting and thereby improve decision making.

Acknowledgments

The authors are grateful to Giuseppe Lopomo, James E. Smith, Ilia Tsetlin, and the reviewers for helpful comments.

Appendix

Proof of Proposition 3.1. Forecaster 1's expected utility is

$w_1\{p_1F_{p_2|E,p_1}(\beta_1) + (1-p_1)[1-F_{p_2|E',p_1}(\beta_1)]\} + (1-w_1)E_{p_1}[S(\beta_1)]$, so the first-order condition is

$w_1[p_1f_{p_2|E,p_1}(\beta_1) - (1-p_1)f_{p_2|E',p_1}(\beta_1)] + (1-w_1)\partial E_{p_1}[S(\beta_1)]/\partial\beta_1 = 0$. Since S is strictly proper,

$\partial E_{p_1}[S(\beta_1)]/\partial\beta_1 = 0$ iff $\beta_1 = p_1$. Thus, the first-order condition is satisfied at $\beta_1 = p_1$ iff

$p_1f_{p_2|E,p_1}(p_1) - (1-p_1)f_{p_2|E',p_1}(p_1) = 0$, which is equivalent to (4). \square

Proof of Proposition 4.1. Suppose a symmetric, nondecreasing Bayesian Nash equilibrium reporting function b exists and that forecaster j uses it [i.e., she reports $b(p_j)$]. Since $s_E(s_{E'})$ is strictly increasing (decreasing) for any strictly proper scoring rule S (Schervish 1989) and ties have probability zero for all reports β_i in the subset B^+ of the range of b where b is strictly increasing and differentiable, we have

$$P[s_E(b(p_j)) < s_E(\beta_i) | E, p_i] = P[b(p_j) < \beta_i | E, p_i] = F_{p_j|E,p_i}(b^{-1}(\beta_i))$$

$$\text{and } P[s_{E'}(b(p_j)) < s_{E'}(\beta_i) | E', p_i] = P[b(p_j) > \beta_i | E', p_i] = 1 - F_{p_j|E',p_i}(b^{-1}(\beta_i)).$$

Now, forecaster i solves the following problem for reports $\beta_i \in B^+$:

$$\max_{\beta_i \in B^+} \{w[p_i F_{p_j|E,p_i}(b^{-1}(\beta_i)) + (1-p_i)(1-F_{p_j|E',p_i}(b^{-1}(\beta_i)))] + (1-w)[p_i s_E(\beta_i) + (1-p_i)s_{E'}(\beta_i)]\},$$

and the corresponding first-order condition with respect to β_i is

$$w[p_i f_{p_j|E,p_i}(b^{-1}(\beta_i)) - (1-p_i)(f_{p_j|E',p_i}(b^{-1}(\beta_i)))](db^{-1}(\beta_i)/d\beta_i) + (1-w)[p_i s_E'(\beta_i) + (1-p_i)s_{E'}'(\beta_i)] = 0.$$

Since b is a symmetric equilibrium, the optimal β_i must satisfy $\beta_i = b(p_i)$, or $p_i = b^{-1}(\beta_i)$, for all

$\beta_i \in B^+$. Therefore, $db^{-1}(\beta_i)/d\beta_i = 1/b'(p_i)$. Upon substitution, we get the result. \square

Proof of Proposition 6.1. The decision maker, assuming the reports are truthful, will revise his

probability as follows after seeing the reports: $P(E | \beta_1, \beta_2) = 1/\{1 + [p_0(1-\beta_1)(1-\beta_2)/((1-p_0)\beta_1\beta_2)]\}$.

Comparing expected utilities, the decision maker should choose a_1 if $(1-\beta_1)(1-\beta_2)/(\beta_1\beta_2) < k^*$. This

means that from forecaster i 's perspective, $i = 1, 2$, a_1 will be chosen if $\beta_j > (1-\beta_i)/[1 + (k^*-1)\beta_i]$,

where $j \neq i$. If forecaster j reports truthfully (i.e., $\beta_j = p_j$), then forecaster i 's expected utility from reporting β_i can be expressed in the form $p_i s_{i,E}(\beta_i) + (1 - p_i) s_{i,E'}(\beta_i)$, which is the expected score from S_i . From McCarthy (1956) and Savage (1971), S_i is strictly proper if $p_i s_{i,E}(p_i) + (1 - p_i) s_{i,E'}(p_i)$, the expected score under truthful reporting, is strictly convex. Convexity can be shown to hold here by examining the second derivative of the expected score. Therefore, truthful reporting by each forecaster and probability revision by the decision maker assuming such truthful reporting is a perfect Bayesian equilibrium that can be implemented as a strictly proper joint scoring rule $S = \{S_1, S_2\}$. \square

References

- Brier, G.W. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Rev.* **78** 1-3.
- Dawid, A.P., M.H. DeGroot, J. Mortera. 1995. Coherent combinations of experts' opinions. *Test* **4**(2) 263-313.
- de Finetti, B. 1962. Does it make sense to speak of "good probability appraisers"? I.J. Good, ed., *The Scientist Speculates: An Anthology of Partly-Baked Ideas*. Wiley, New York, 357-363.
- Frank, R.H., P.J. Cook. 1995. *The Winner-Take-All Society*. The Free Press, New York.
- Fudenberg, D., J. Tirole. 1991. *Game Theory*. MIT Press, Cambridge, MA.
- Gaba, A., I. Tsetlin, R.L. Winkler. 2004. Modifying variability and correlations in winner-take-all contests. *Oper. Res.* **52**(3) 384-395.
- Kadane, J.B., R.L. Winkler. 1988. Separating probability elicitation from utilities. *J. Amer. Statist. Assoc.* **83**(2) 357-363.
- Kilgour, D.M., Y. Gerchak. 2004. Elicitation of probabilities using competitive scoring rules. *Decision Analysis* **1**(2) 108-113.
- Lebrun, B. 1999. First price auctions in the asymmetric N bidder case. *Int. Econom. Rev.* **40**(1) 125-142.
- Marshall, R.C., M.J. Meurer, J.-F. Richard, W. Stromquist. 2004. Numerical analysis of asymmetric first price auctions. *Games Econom. Behav.* **7**(2) 193-220.
- McCarthy, J. 1956. Measures of the value of information. *Proc. Nat. Acad. Sci. USA* **42** 654-655.
- Morris, P.A. 1974. Decision analysis expert use. *Management Sci.* **20**(9) 1233-1241.
- O'Hagan, A., C.E. Buck, A. Daneshkhah, J.R. Eiser, P.H. Garthwaite, D.J. Jenkinson, J.E. Oakley, T. Rakow. 2006. *Uncertain Judgements: Eliciting Experts' Probabilities*. Wiley, Chichester,

England.

- Ottaviani, M., P.N. Sorensen. 2005. Forecasting and rank-order contests. Working paper, London Business School, London.
- Pfeifer, P.E. 1994. Are we overconfident in the belief that probability forecasters are overconfident? *Organ. Behavior Human Decision Processes* **58** 203-213.
- Prelec, D. 2004. A Bayesian truth serum for subjective data. *Science* **306** 462-466.
- Savage, L.J. 1971. Elicitation of personal probabilities and expectations. *J. Amer. Statist. Assoc.* **66** 783-801.
- Schervish, M.J. 1989. A general method for comparing probability assessors. *Ann. Statist.* **17** 1856-1879.
- Tetlock, P.E. 2005. *Expert Political Judgment*. Princeton University Press, Princeton, NJ.
- Winkler, R.L. 1969. Scoring rules and the evaluation of probability assessors. *J. Amer. Statist. Assoc.* **64**(3) 1073-1078.
- Winkler, R.L. 1996. Scoring rules and the evaluation of probabilities. *Test* **5**(1) 1-60.
- Winkler, R.L., A.H. Murphy. 1970. Nonlinear utility and the probability score. *J. Appl. Meteorol.* **9**(1) 143-148.

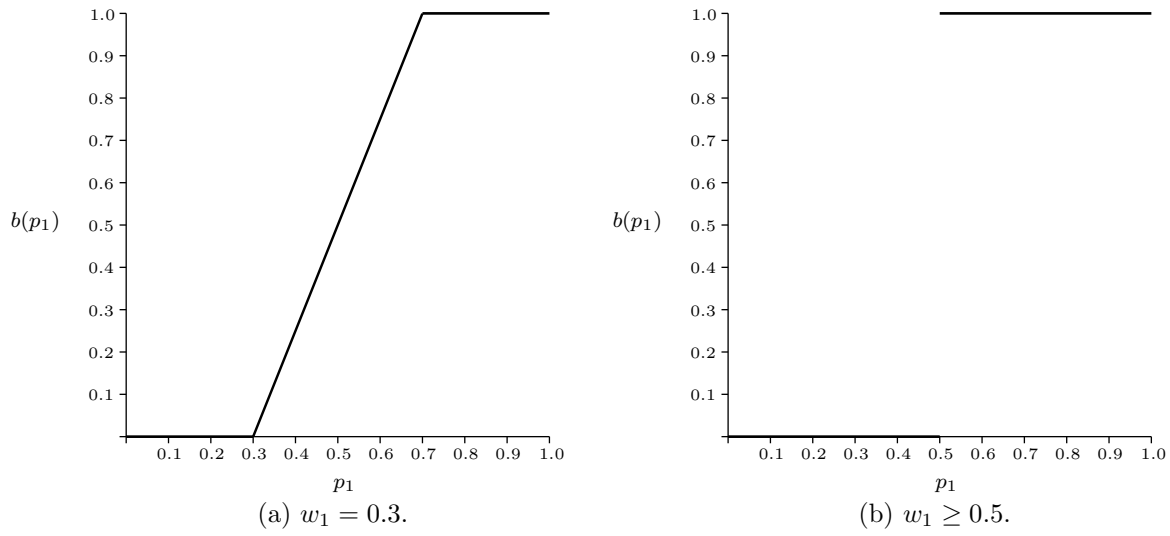


Figure 1: Best response to truthful reporting in Example 3.1.

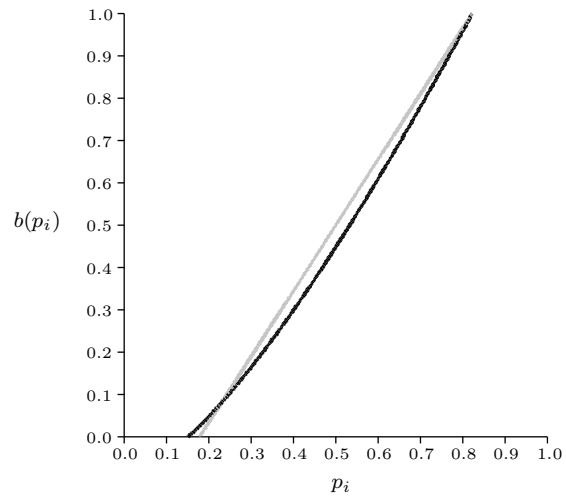


Figure 2: Differential equation solutions in Example 4.1 for $w = 0.3$ and (c, d, c', d') equal to $(2,1,1,2)$ in gray and $(3,1,2,2)$ in black.

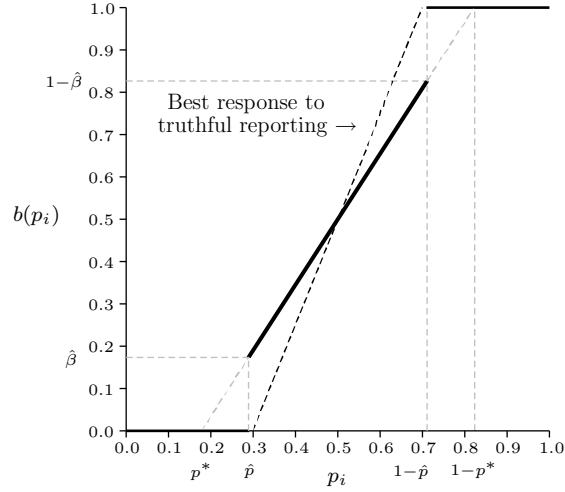
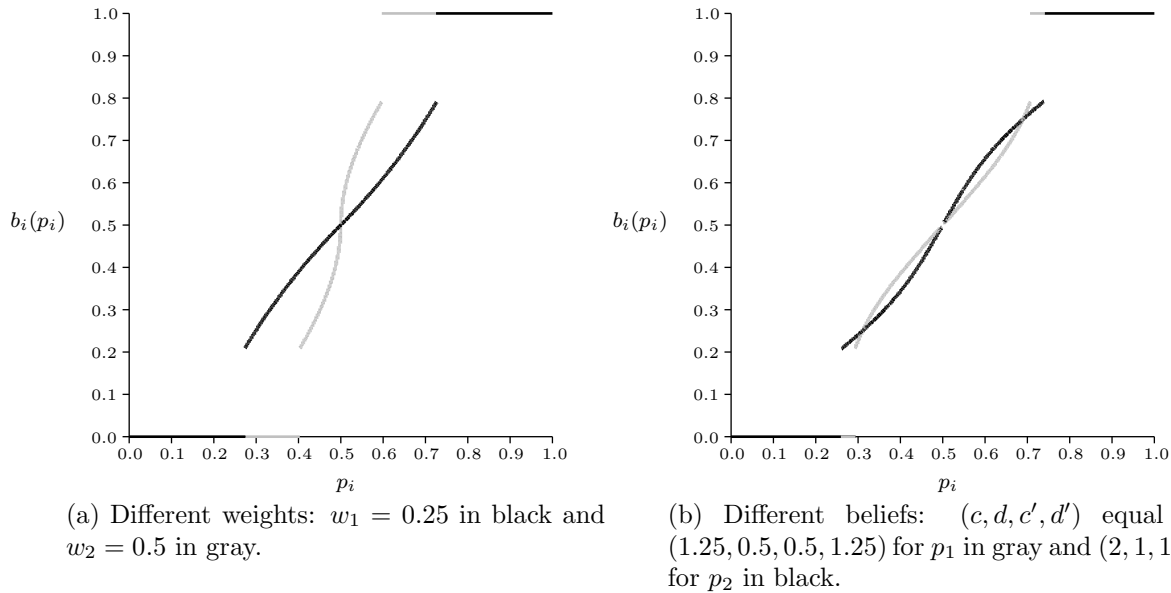


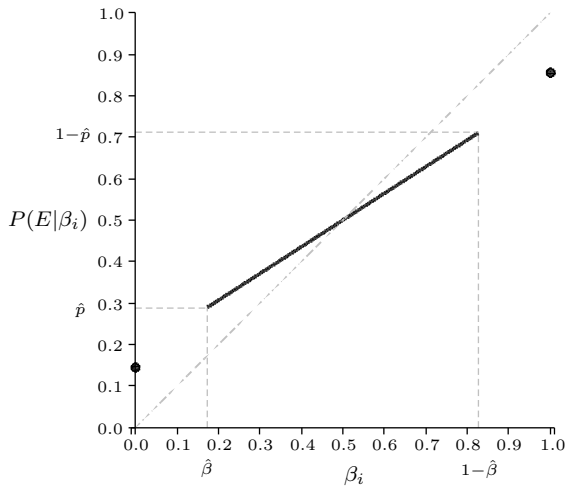
Figure 3: Symmetric equilibrium and best response to truthful reporting in Example 4.1 with $w = 0.3$ and $(c, d, c', d') = (2, 1, 1, 2)$.



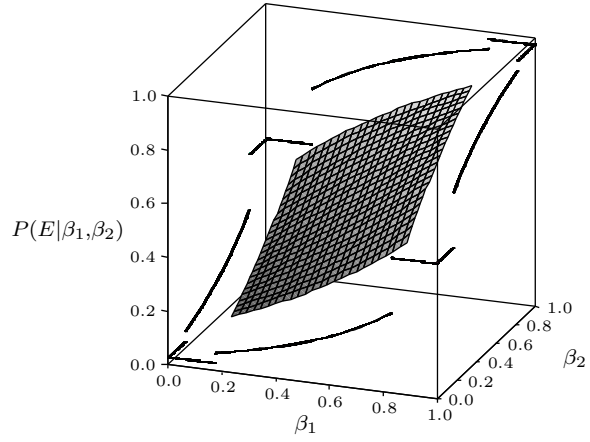
(a) Different weights: $w_1 = 0.25$ in black and $w_2 = 0.5$ in gray.

(b) Different beliefs: (c, d, c', d') equal to $(1.25, 0.5, 0.5, 1.25)$ for p_1 in gray and $(2, 1, 1, 2)$ for p_2 in black.

Figure 4: Equilibria for asymmetric games in Examples 4.2 and 4.3.

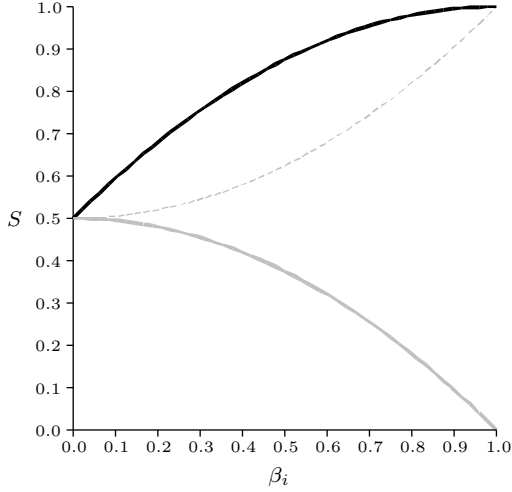


(a) $P(E|\beta_i)$.

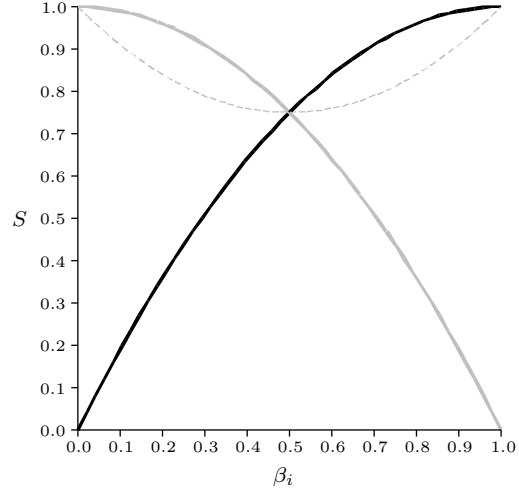


(b) $P(E|\beta_1, \beta_2)$.

Figure 5: Decision maker's revised probability for E in Example 5.1.



(a) $u(a_1, E) = 1, u(a_1, E') = 0, u(a_2, E) = u(a_2, E') = 0.5$.



(b) $u(a_1, E) = u(a_2, E') = 1, u(a_1, E') = u(a_2, E) = 0$.

Figure 6: Joint scoring rules in Example 6.1 with $s_{i,E}$ in black, $s_{i,E'}$ in solid gray, and expected score in dashed gray.