http://pubsonline.informs.org/journal/mnsc/

MANAGEMENT SCIENCE
Articles in Advance, pp. 1–19
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Extracting the Wisdom of Crowds When Information Is Shared

Asa B. Palley,[a] Jack B. Soll[b]

[a] Kelley School of Business, Indiana University, Bloomington, Indiana 47405; [b] Fuqua School of Business, Duke University, Durham, North Carolina 27708

Contact: apalley@indiana.edu, http://orcid.org/0000-0002-6724-4654 (ABP); jsoll@duke.edu, http://orcid.org/0000-0002-4496-1976 (JBS)

Received: September 18, 2015
Revised: December 14, 2016; November 14, 2017; December 19, 2017
Accepted: December 29, 2017
Published Online in Articles in Advance:

https://doi.org/10.1287/mnsc.2018.3047

Abstract. Using the wisdom of crowds—combining many individual judgments to obtain an aggregate estimate—can be an effective technique for improving judgment accuracy. In practice, however, accuracy is limited by the presence of correlated judgment errors, which often emerge because information is shared. To address this problem, we propose an elicitation procedure in which respondents are asked to provide both their own best judgment and an estimate of the average judgment that will be given by all other respondents. We develop an aggregation method, called pivoting, which separates individual judgments into shared and private information and then recombines these results in the optimal manner. In several studies, we investigate the method and examine the accuracy of the aggregate estimates. Overall, the empirical data suggest that the pivoting method provides an effective judgment aggregation procedure that can significantly outperform the simple crowd average.

History: Accepted by Manel Baucells, decision analysis.
Funding: This work was supported by Duke University's Fuqua School of Business.
Supplemental Material: Data and the online appendix are available at https://doi.org/10.1287/mnsc.2018.3047.

Keywords: judgment aggregation • wisdom of crowds • forecasting • shared information

1. Introduction

Obtaining accurate estimates of uncertain variables is an important problem across a broad range of applications, ranging from managerial decision problems to macroeconomics to geopolitics. These estimates often comprise a crucial input to decision making in practice, offering insight into questions such as the following: "How many units of this product will be sold if we develop it and charge this price?" "How much will GDP grow over the next three years if we implement this policy?" or "What are the chances that the United Kingdom will exit the European Union by the end of 2019?"

Often, there are many pieces of useful information that can help generate a better estimate, but they may be scattered across different institutions and people. Combining many individual judgments to obtain an aggregate estimate can, therefore, provide an effective technique for improving judgment accuracy. This idea is known as the wisdom of crowds (Surowiecki 2005, Page 2008), and evidence for its effectiveness dates back more than a century (Galton 1907). Since that time, many studies have shown that the simple average performs remarkably well in a wide variety of settings (Makridakis and Winkler 1983, Clemen and Winkler 1986, Clemen 1989, Larrick et al. 2012). One reason for the success of averaging is that as long as judgments bracket the truth (i.e., some are too high and others too low), averaging is guaranteed to be more accurate than the average individual (Larrick and Soll 2006).

Theoretically, the simple average provides an ideal crowd estimate when individual judgment errors are independent and identically distributed. These conditions are rarely met in practice, and a variety of techniques have been proposed to improve accuracy. Some, such as trimming and Winsorizing (Armstrong 2001, Jose and Winkler 2008), have proven modestly beneficial and robust across different error distributions. Other approaches strive to exploit the covariance structure in past data to derive optimal weights (Winkler 1981, Clemen and Winkler 1986, Budescu and Chen 2015). To be effective, these methods require a large amount of data and a sufficiently stable estimation environment. In practice, perhaps because these conditions are usually not met, simple averaging often performs as well as and sometimes better than other approaches (Clemen 1989).

Regardless of the aggregation method, the potential benefit from combining estimates is greatly limited by correlation in judgment errors. For example, if judgment errors have a common variance and pairwise correlations of $\rho = 0.25$, the information that could be extracted from even a very large number of judgments will be less than only four equivalent independent judges (Clemen and Winkler 1985). In practice, dependence often emerges because shared information leads to similar judgments. Even though the shared information may be highly informative, the benefits of averaging across many judges are limited because a

significant proportion of the average judgment represents the same information being repeated. We refer to this as the *shared-information problem*.

In this paper, we introduce a judgment aggregation method that can mitigate the shared-information problem and yield a more accurate crowd estimate. To do this, we propose augmenting the standard elicitation procedure, so that in addition to providing their own judgments, individuals also guess how others will respond (Prelec 2004). By asking this additional question, the decision analyst can estimate which part of each judgment is shared and which part is private information. The aggregation method, which we call *pivoting*, then builds a crowd estimate by recombining each of these shared and private pieces of information in a Bayesian manner. This reduces the judgment error of the aggregate estimate to its minimum as the crowd size grows large. We introduce several variations on pivoting, each of which makes different assumptions about how information may be distributed across judges. Finally, we present four studies that examine the accuracy of pivoting in both controlled and real-world contexts.

2. Literature Review

Past efforts at addressing the shared-information problem can be viewed as falling into two categories—information-focused and judge-focused. Information-focused methods seek to reduce the weight on shared information or a common prior and to increase the weight on private information. An example is a winner-take-all forecasting contest, in which judges may have an incentive to shade forecasts toward their own private information to differentiate themselves and increase their chances of winning (Ottaviani and Sørensen 2006). Because of the greater weight on private information, the average forecast can be more accurate in a competitive crowd (Lichtendahl et al. 2013). An alternative approach asks forecasters to reveal their private and shared information, leaving it to a decision analyst to combine them. Chen et al. (2004) show how such decomposed judgments can be obtained in a coordination game.

Another information-focused approach is to have the decision analyst make an educated guess about the shared information or common prior. For example, the decision analyst might assume a common prior of 0.5 when assessing the probability of a binary event. If judges individually possess little private information, their individual judgments are likely to coalesce near the prior. Across judges, however, private information may be substantial. The decision analyst can account for this by "extremizing" the combined judgment relative to the assumed prior, pushing probabilities closer to 0 and 1 (Baron et al. 2014). Similar logic applies to point estimates. Kim et al. (2001) propose a model in which formerly private information becomes commonly known over time. In a multi-period setting, they suggest using average past forecasts as a proxy for current-period shared information. The overweighting of shared information can be remedied "by adding a positive multiple of the change in the mean forecast to the current mean forecast" (p. 335).

Judge-focused methods seek to assign more weight to more knowledgeable, accurate, or contributory judges (Budescu and Chen 2015). It is plausible that better judges possess extra information in addition to what is commonly held. Thus, by weighting the answers of more accurate judges more heavily, these methods may have the added benefit of reducing the weight given to shared information. To identify these judges, the decision analyst can rely on cues to expertise, such as past accuracy (e.g., Armstrong 2001, Mannes et al. 2014).

Prelec et al. (2017) develop an innovative, nonparametric, categorical approach they call the "surprisingly popular" (SP) algorithm (first introduced by Prelec and Seung 2006). Judges report both their own choice and estimate the frequencies of others' choices. The SP algorithm selects the surprisingly popular answer—the one that more people vote for than expected. The algorithm, which requires no past data, is judge-focused in the sense that it seeks to identify the answer favored by the most knowledgeable judges. The methodology has been extended to multi-question contexts and noisy responses (McCoy and Prelec 2017) and can handle continuous quantities by discretizing the response scale. In theory, the SP algorithm will identify the correct answer as the crowd size grows large. In practice, the algorithm has been shown to outperform other methods, such as majority rule and confidence-weighted voting.

The pivoting method that we propose in this paper is information-focused. The goal is to estimate the judgment of a hypothetical rational agent with access to all of the information distributed across all of the judges. Pivoting shares several features in common with the SP algorithm. Both methods ask judges to predict the responses of others, which provides information about the signal-generating process. This process is assumed to be known to judges but not to the analyst. In addition, both methods improve upon simple aggregation rules by making a contrarian modification to the crowd opinion. The SP algorithm accomplishes this by selecting the surprisingly popular answer. In contrast, pivoting modifies the crowd opinion by adjusting away from the prediction of others. This adjustment is analogous to extremizing and similar to the Kim et al. (2001) adjustment away from the previous period's mean as described earlier. In fact, extremizing can be viewed as a special case of pivoting in which judgments are adjusted away from a central value. Finally, both pivoting and SP provide aggregation rules, which

distinguish them from peer prediction mechanisms that improve crowd judgments by incentivizing truth telling (Miller et al. 2005, Jurca and Faltings 2009) and discouraging low-knowledge individuals from participating (Witkowski et al. 2013). In contrast to SP, pivoting is explicitly designed for continuous judgments and can, in theory, completely alleviate the shared-information problem when the number of judges is small.

One might hope that judges can address the shared-information problem through information exchange and discussion. However, research on group decision making suggests the contrary. Shared information has greater impact on group decisions than uniquely held information—a phenomenon known in social psychology as the common knowledge effect (Gigone and Hastie 1993). Also, people often lack the correct intuitions about correlation and shared information (Soll 1999, Budescu and Yu 2007, Yaniv et al. 2009) and, thus, may not recognize it as a problem to overcome in group settings. Just as in the algorithmic combination of judgments, shared information in a group setting enters the aggregate multiple times and, thus, has an outsized impact on the group decision.

3. The Shared-Information Problem

In the most general setting, information may be distributed haphazardly among judges as illustrated by the exemplar general structure in Figure 1. Some information set s is held by all judges, and additional information may be held uniquely or by various subsets of judges, represented by the overlap in the ovals in the figure. Although averaging can be an effective way to combine judgments, the shared signal s may bias the result in this setting—each judge will use s in the judge's own judgment, meaning that s will comprise an outsized portion of the average judgment. The general structure would be difficult if not impossible to analyze because the possible sets of overlapping information are combinatorially complex. Therefore, we consider several idealized structures that approximate a range of real-world situations. Whereas the models differ substantially in their assumptions about how information is distributed, they all imply a form of pivoting as the optimal aggregation policy.

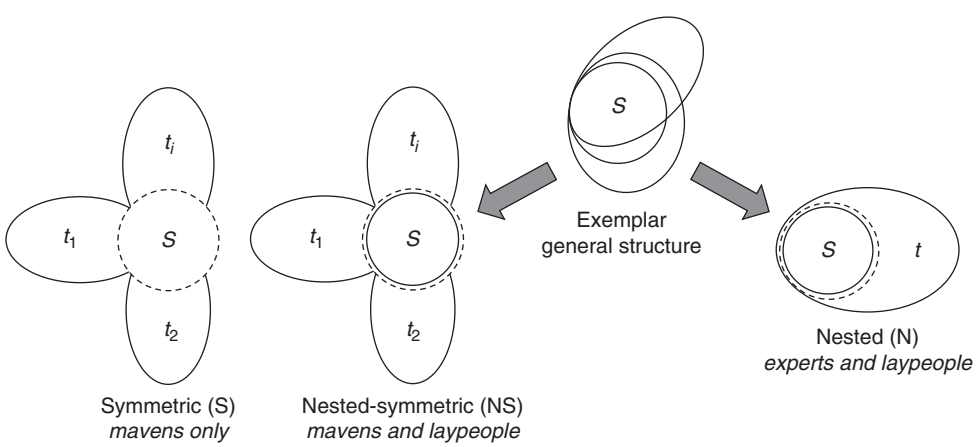
3.1. Laypeople, Experts, and Mavens

Our idealized models incorporate three types of people: laypeople, experts, and mavens. Laypeople have access to the shared signal and nothing else. Experts have access to additional information, which they all share. Mavens also have additional information beyond the shared signal; except in contrast to experts, they hold it uniquely. The key difference between experts and mavens is that whereas additional mavens are always valuable because they bring new information, the value of additional experts is limited to reducing noise. In general structures, there is a continuum between experts and mavens. Judges are more maven-like to the extent that they possess additional information that is held by few others. For the purpose of this paper, we consider structures that have either experts or mavens but not both.

3.2. Three Idealized Information Structures

The symmetric (S) information structure is the predominant setup used in previous literature (e.g., Kim et al. 2001, Ottaviani and Sørensen 2006, Lichtendahl et al. 2013). This model assumes that all judges have access to the same shared information s, and additionally, each judge i receives an equal amount of private information, represented by t_i in Figure 1. All judges are mavens, and any given piece of information is held by either everyone or by only one person. The S structure is a special case of the nested–symmetric (NS)

Figure 1. Comparison of Information Structures



Notes. A solid line around s indicates the presence of laypeople who observe shared information only. A dashed line indicates that, in addition to s, mavens and experts observe another signal t_i or t.

structure. In the NS structure, some judges are laypeople who only observe s, and others are mavens, each of whom observes a private signal t_i in addition to s. Finally, the nested (N) structure is comprised of laypeople and experts. As in the other two structures, all judges have access to s, but experts also observe a common signal t. In all structures, the decision analyst does not know which judges possess additional information yet wishes to extract their extra knowledge from everyone's responses. While the amount of total information increases with the number of judges in the S and NS structures, only two total signals are available to judges in the N structure.

Although the assumptions of each model will rarely hold exactly, the range of structures captures essential features of the shared-information problem. For example, the NS structure could be a plausible approximation to market situations in which judges supplement public information with their own research. The N structure can approximate situations in which some people are very familiar with a judgment context and others know less, such as predicting the outcome of a college basketball game.

3.3. Linear Aggregation Problems

Let X denote the random variable being estimated, which is distributed according to a known cumulative distribution function $F(X \mid \theta)$ with unknown mean θ. There are n judges, who share a common prior belief $\pi_0(\theta)$ over the variable of interest θ with mean μ_0 and finite variance σ_0^2. We assume a finite prior predictive variance $V_0 \equiv \mathbb{E}[(X-\theta)^2] < \infty$, where the expectation is taken over $F(X \mid \theta)$ and $\pi_0(\theta)$. All judges observe the same shared signal s_1, which equals the average of m_1 independent observations of X. In addition, in the N structure, some judges receive an additional signal t, which equals the average of l independent observations of X. In the NS structure, some judges receive an additional signal t_i, which equals the average of l_i independent and judge-specific observations with equal numbers of observations, $l_i = l$, across judges. We restrict our attention to families of distributions where the posterior expectation can be written as a linear combination of the prior expectation and the information observed.

Definition 1. We say that the information aggregation problem is *linear* if the posterior expectation of θ given prior $\pi_0(\theta)$, shared signal s_1, and any such collection of private signals $\{t_1, \ldots, t_K\}$ is given by $\mathbb{E}[\theta \mid \pi_0, s_1, t_1, \ldots, t_K] = (m_0\mu_0 + m_1 s_1 + l\sum_{k=1}^{K} t_k)/(m_0 + m_1 + lK)$ for some m_0.

For linear information aggregation problems, the prior mean μ_0 can be thought of as representing m_0 observations of independent realizations of X. It will be convenient to define $m \equiv m_0 + m_1$ and $s \equiv (m_0\mu_0 + m_1 s_1)/m$. Note that s, which we refer to as the shared signal, is an amalgamation of the information contained in μ_0 and s_1. After receiving the pair of signals (s, t_i), judge i updates the judge's prior about θ according to Bayes' rule. The judge combines information according to the weight $w = l/(m+l)$, which represents the relative informativeness of private versus shared information. We assume that this information structure and the parameters m_0, m_1, and l are commonly known by all judges. Three examples of random variables X, which yield linear information aggregation problems, include (1) a normal variable with unknown mean θ, (2) a binomial variable with unknown probability θ of success on each of the trials, and (3) a gamma variable with known shape and unknown scale. A complete characterization of the parameters, signals, and natural conjugate families for each of these examples is provided in an electronic companion to this article.

The model treats information held by judges as equivalent to independent samples that vary in size. This modeling strategy makes it possible to capture the relevant statistical features of judgment problems, such as varying levels of expertise and information overlap. Ideally, the decision analyst would like to determine the *global posterior expectation* (GPE) of X, the estimate $\mathbb{E}[X \mid \pi_0, s_1, t_1, \ldots, t_K]$ that efficiently combines all signals $\{s_1, t_1, \ldots, t_K\}$ observed by all judges (Frongillo et al. 2015).[1]

3.4. Addressing the Shared-Information Problem With Pivoting

We illustrate the intuition behind pivoting by considering the case of information distributed according to the S structure. The decision analyst would like to use responses from each judge to build a crowd estimate of X. Each judge produces a point estimate $f_i = (1-w)\cdot s + wt_i$ equal to the mean of the judge's predictive distribution for this quantity. This implies that a simple average of the individual judgments f_i is a weighted average with weight $(1-w)$ on the shared signal and weight w on the average of the private signals. Each individual is partially repeating s, and therefore, the simple average over-weights the shared signal relative to what is optimal. As the number of judges, n, grows large, the average of the private signals converges to the true mean θ by the law of large numbers while the specific realization s of the shared signal remains the same. As a result, when taking a simple average of the f_i, any error resulting from s cannot be reduced by simply including more judges. We call this residual error *shared-information bias*, defined as $\mathbb{E}[\bar{f} - X \mid s, \theta]$, the expected difference between the average judgment and X given the shared information and the true mean of X. In the S structure, the shared-information bias equals $(1-w)(s-\theta)$.

The decision analyst can alleviate this bias by asking judges an additional question to help identify what

part of their information is shared: "What do you think will be the average judgment of the other $n-1$ judges?" We show that, by asking for both a personal judgment f_i of X and also an estimate g_i of others' judgments, the decision analyst can distinguish between information that is common to all judges and information that is specific to judge i. To incentivize judges, the decision analyst may choose any two strictly proper scoring rules that elicit mean beliefs for both f_i and g_i.[2] Judge i knows that each other judge j observes both the shared signal s and a different private signal t_j, which has mean θ, so the judge should provide a response of $g_i = (1-w)s + w\mathbb{E}[\theta \mid s, t_i] = (1-w)s + wf_i = (1-w^2)s + w^2 t_i$.

Although both responses will lie between the shared and private signals, f_i lands closer to the private information while g_i lands closer to the shared information. This holds true on average as well; the average judgment $\bar{f}$ is shaded closer toward the average private signal $\bar{t}$ while the average guess of others $\bar{g}$ is shaded closer to the shared signal s. The decision analyst can then exploit this difference to identify the shared and private information, which can be recombined to eliminate shared-information bias.

Our proposed method starts at $\bar{f}$ and *pivots* in different directions to produce estimates of $\bar{t}$ and s as shown in Figure 2. If the relative information weight w is known, the decision analyst can estimate $\bar{t}$ by beginning with $\bar{f}$ and pivoting in the direction opposite of the average guess of others $\bar{g}$ by a factor of $1/w$. Note that when the decision analyst estimates $\bar{t}$, the decision analyst completely removes the shared information from the average judgment, but the decision analyst's ultimate estimate of θ should use s to some extent as well. The decision analyst can estimate s by starting with $\bar{f}$ and pivoting in the direction toward $\bar{g}$ by a factor of $1/(1-w)$. The decision analyst then combines the estimates $\hat{\bar{t}}$ of the average private signal and $\hat{s}$ of the shared signal by weighting them in the optimal Bayesian fashion. Details of these calculations can be found in Equations (2)–(4) in Section 3.5, setting $p = 1$.

Of course, this example makes a number of simplifying assumptions. In the following sections, we explore different information structures and relax the assumptions that responses are noiseless and parameters are known to the decision analyst.

3.5. Developing a General Pivoting Procedure

While existing literature typically considers a symmetric information structure with normally distributed variables, in general, both the information structure and distribution shape could vary. In this section, we present a general pivoting procedure that accommodates any linear information aggregation problem in each of the S, N, and NS structures.

In many situations, judges may possess varying levels of expertise or access to information with some knowing more than others. To study these settings, suppose that a proportion $1-p$ of judges are *laypeople*, who only observe the shared signal s_1, while the remaining proportion p of judges are *experts* (in the N structure) or *mavens* (in the NS structure), who observe both s_1 and an additional signal that allows them to make better judgments. In the N structure, experts receive an additional common signal t, which is independent of s_1 conditional on θ. In the NS structure, each maven observes the maven's own conditionally independent private signal t_i. Experts (mavens) are assumed to share a common expectation p of the proportion of other experts (mavens) in the crowd. The S structure is a special case of NS with $p = 1$.

Proposition 1. *If the information aggregation problem is linear and judges are incentivized to provide their mean beliefs, then it is optimal for each layperson i to provide responses of $f_i^* = g_i^* = s$ and for each expert or maven i to provide responses of*

$$\begin{aligned} f_i^* &= (1-w)s + w\tau_i, \\ g_i^* &= (1-pw\eta)s + pw\eta\tau_i, \end{aligned} \tag{1}$$

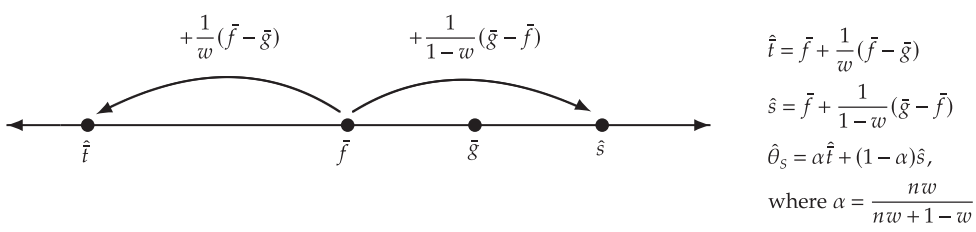
where

$$(\tau_i, \eta) = \begin{cases} (t, 1) & \text{if N information structure,} \\ (t_i, w) & \text{if NS information structure.} \end{cases}$$

Proofs of all results can be found in the appendix.

To develop a general pivoting procedure, we begin with the simple case where the decision analyst knows w and p, and we then expand the procedure in Section 4 to allow for uncertainty about these parameters. Recall that the S structure is a special case of NS where $p = 1$.

Figure 2. Estimating the Private and Shared Signals in the S Information Structure When $w \in (0,1)$ Is Known



1. Estimate the average private signal according to

$$\hat{\bar{t}} = \begin{cases} \bar{f} + \dfrac{1}{pw}(\bar{f} - \bar{g}) \\ \quad \text{if } pw > 0 \text{ and S or NS structure,} \\ \bar{f} + \dfrac{1-pw}{pw(1-p)}(\bar{f} - \bar{g}) \\ \quad \text{if } pw > 0, p < 1, \text{ and N structure,} \\ \bar{f} \quad \text{otherwise.} \end{cases} \tag{2}$$

2. Estimate the shared signal according to

$$\hat{s} = \begin{cases} \bar{f} + \dfrac{1}{1-pw}(\bar{g} - \bar{f}) \\ \quad \text{if } pw < 1 \text{ and S or NS structure,} \\ \bar{f} + \dfrac{1}{1-p}(\bar{g} - \bar{f}) \\ \quad \text{if } p < 1 \text{ and N structure,} \\ \bar{f} \quad \text{otherwise.} \end{cases} \tag{3}$$

3. Compute the aggregate estimate

$$\hat{\theta} = \alpha\hat{\bar{t}} + (1-\alpha)\hat{s},$$
$$\text{where } \alpha = \begin{cases} \dfrac{npw}{npw + 1 - w} & \text{if S or NS structure,} \\ w & \text{if N structure.} \end{cases} \tag{4}$$

The weight α in Equation (4) captures the relative amount of information contained in the average private and shared signals, respectively. In two cases, $\bar{f}$ and $\bar{g}$ are equal, and the procedure simply returns the simple average $\bar{f}$ as the crowd estimate: If $w = 0$, then all information is shared, and $\bar{f} = \bar{g} = s$. If $w = 1$, then all information is privately held, and $\bar{f} = \bar{g} = \sum_{i=1}^{n} t_i/n$.

Proposition 2. *If the information aggregation problem is linear and judges report their true beliefs, then the crowd estimate $\hat{\theta}$ given in Equation (4) is the GPE of X.*

Proposition 2 shows that, in principle, the pivoting procedure can recover the optimal point estimate of the crowd. It is *as if* the decision analyst had direct access to all the signals. In other words, pivoting completely overcomes the problem of shared-information bias. Of course, this will typically be unattainable because judges report imperfectly and w and p are not known.

3.6. Noisy Responses

In practice, responses may deviate from optimality for a variety of reasons. Cognitive limitations and mental costs often prevent people from fully processing all of the available information and working out the optimal responses, leading to the use of simplifying heuristics (Payne et al. 1993). In addition, judges may be inconsistent in attending to and weighting information, may have heterogeneous prior beliefs, and may make mistakes when thinking about the behavior of others. Relaxing the model of response behavior to account for these deviations from optimality is a necessary step to work with data from multiple judges because not doing so would lead to an intractable, contradictory set of equations.

Extending the results in Proposition 1 to accommodate response variation, assume that laypeople, on average, estimate the shared signal s and, lacking additional information, guess the same for the average response of others, so $f_i = g_i = s + \delta_i$, where δ_i is an independent mean-zero error. Experts will use both signals s and t, combining them according to the appropriate Bayesian weight w. In addition, they know that they possess more knowledge than laypeople and expect that p of the other judges are experts and $1-p$ of the other judges are laypeople. In the N structure, experts will, therefore, provide a guess of others' judgments of $g_i = pf_i + (1-p)s$. In the NS structure, mavens expect that the judgment of other mavens will be $(1-w)s + w\bar{t}_{-i}$, where $\bar{t}_{-i}$ is their average private signal. A maven's best estimate of $\bar{t}_{-i}$ is the maven's own judgment f_i of the mean θ. Making the appropriate substitutions and allowing for noise in responses, a behavioral model of responses for an expert or maven i (for all three structures) can be written as $f_i = (1-w)s + w\tau_i + \varepsilon_i$ and $g_i = (1-pw\eta)s + pw\eta\tau_i + p\eta\varepsilon_i + \gamma_i$, where ε_i and γ_i are independent mean-zero error terms. If this behavioral model holds, pivoting will still outperform the average judgment[3] for a sufficiently large crowd size n for each of the idealized information structures as stated formally by the next proposition.

Proposition 3. *Even with noisy responses, if the information aggregation problem is linear, the tailored pivoted estimate $\hat{\theta}$ specified in Equation (4) for a particular information structure provides a lower expected squared error than the average judgment $\bar{f}$ for a sufficiently large crowd. In the N structure, $\mathbb{E}[(\hat{\theta}_N - \theta)^2] \to (1-w)^2((m_0^2\sigma_0^2 + m_1^2(V_0/m_1))/m^2) + w^2(V_0/l)$ as $n \to \infty$ while $\mathbb{E}[(\bar{f} - \theta)^2] \geq (1-pw)^2((m_0^2\sigma_0^2 + m_1^2(V_0/m_1))/m^2) + p^2w^2(V_0/l)$. In the S and NS structures, $\mathbb{E}[(\hat{\theta}_{NS} - \theta)^2] \to 0$ as $n \to \infty$ while $\mathbb{E}[(\bar{f} - \theta)^2] \geq (1-pw)^2((m_0^2\sigma_0^2 + m_1^2(V_0/m_1))/m^2)$.*

4. Practical Implementation of Pivoting

In practice, the decision analyst typically does not know the parameters w and p of these information models and must estimate them from response data. In this section, we present one possible approach to modifying the pivoting procedure presented in Section 3.5 to account for uncertainty in the parameter values in each of the N, S, and NS structures. We also recognize that the decision analyst may not know which structure best corresponds to the judges' information setting. To address

this, in Section 4.2, we propose a simple universal procedure called minimal (M) pivoting.

4.1. A Procedure When Parameters Are Uncertain

To implement the N and NS procedures, the decision analyst must use the response data to estimate the proportion p of experts and mavens. We propose a simple approach in which the decision analyst assumes that $f_i = g_i$ if and only if judge i is a layperson. The decision analyst can then simply calculate the proportion of judges who provide a guess about others that differs from their own judgment and use this in place of p in Equation (4). Although some classification errors are inevitable, preliminary analysis using the data from Study 1 suggests that the accuracy of the method is robust to moderate relaxations of the classification criteria on the closeness between f and g. More formal statistical classification procedures (e.g., cluster analysis) could offer an alternative method for identifying which judges are laypeople.

Next, in the S and NS structures, the decision analyst needs an estimate of pw to implement the pivoting procedure. The behavioral model assumes that experts and mavens provide responses of $g_i = (1-pw)s + pwf_i + \gamma_i$, where γ_i is a noise term with mean zero that is independent of s. If we assume that γ_i is normal and define $q \equiv pw$, we can estimate q using the slope coefficient $\hat{q}$ from a least-squares linear regression of g on f *among mavens*. Let d be the number of mavens who provide judgments. If $d > 2$, $(\hat{q} - q)/s_q$ follows a t-distribution with $d-2$ degrees of freedom. Taking a uniform prior over the possible values of q ($\pi(q) = \mathbf{1}_{[0,1]}$) and using the standard error $s_{\hat{q}}$ as an estimate of s_q, Bayes' theorem allows us to use the responses to derive the posterior distribution $f(q \mid \hat{q}, s_{\hat{q}}) \propto t_{d-2}^{\hat{q}, s_{\hat{q}}}(q)\mathbf{1}_{[0,1]}$. In the general estimation procedure in Equation (4), q appears in the denominator when estimating the average private and shared signals in steps 1 and 2. This implies that any uncertainty in the estimate of q will cause "over-pivoted" estimates of both $\bar{t}$ and s, leading to a biased estimate of θ in step 3. To solve this problem, the decision analyst should instead pivot according to a hedged weight h that minimizes the expected squared error in estimating these signals. The following algorithm provides an analogue of the pivoting procedure in Equation (4) that accounts for this uncertainty in q:

1. Compute $\bar{f}$ and $\bar{g}$ using the full set of responses $\{(f_i, g_i)\}_{i=1}^n$.
2. Count the number of judges d who gave a different guess about others than their own judgment (count all i such that $f_i \neq g_i$) and compute $\hat{p} = d/n$. In the S structure, the decision analyst considers all judges to be mavens and sets $\hat{p} = 1$ and $d = n$.
3. In the N structure or if $d \leq 2$, the aggregate judgment is

$$\hat{\theta}_N^* = \begin{cases} \bar{f} + (\bar{f} - \bar{g})/\hat{p} & \text{if } \hat{p} > 0, \\ \bar{f} & \text{if } \hat{p} = 0. \end{cases}$$

4. In the S and NS structures when $d > 2$, estimate the slope coefficient $\hat{q}$ and its standard error $s_{\hat{q}}$ from a simple least-squares linear regression of g on f within the subset of data $\{(f_i, g_i) \mid f_i \neq g_i\}$.
5. Estimate the average private signal according to

$$\hat{\bar{t}}^* = \bar{f} + \frac{1}{h_{private}^*(\hat{q}, s_{\hat{q}}, d)}(\bar{f} - \bar{g}),$$

$$\text{where } h_{private}^*(\hat{q}, s_{\hat{q}}, d) = \frac{\int_0^1 q^2(1-q)^2 t_{d-2}^{\hat{q}, s_{\hat{q}}}(q)\,dq}{\int_0^1 q(1-q)^2 t_{d-2}^{\hat{q}, s_{\hat{q}}}(q)\,dq}.$$

6. Estimate the shared signal according to

$$\hat{s}^* = \bar{f} + \frac{1}{1 - h_{shared}^*(\hat{q}, s_{\hat{q}}, d)}(\bar{g} - \bar{f}),$$

$$\text{where } h_{shared}^*(\hat{q}, s_{\hat{q}}, d) = \frac{\int_0^1 q^3(1-q) t_{d-2}^{\hat{q}, s_{\hat{q}}}(q)\,dq}{\int_0^1 q^2(1-q) t_{d-2}^{\hat{q}, s_{\hat{q}}}(q)\,dq}.$$

7. Compute the aggregate estimate

$$\hat{\theta}_{NS}^* = \alpha^* \hat{\bar{t}}^* + (1 - \alpha^*)\hat{s}^*, \tag{5}$$

where[4]

$$\alpha^* = \frac{nq^*}{nq^* + 1 - w^*}, \quad q^* = \frac{\int_0^1 q\, t_{d-2}^{\hat{q}, s_{\hat{q}}}(q)\,dq}{\int_0^1 t_{d-2}^{\hat{q}, s_{\hat{q}}}(q)\,dq},$$

$$\text{and} \quad w^* = \frac{\int_0^{\hat{p}} q\, t_{d-2}^{\hat{q}, s_{\hat{q}}}(q)\,dq}{\hat{p}\int_0^{\hat{p}} t_{d-2}^{\hat{q}, s_{\hat{q}}}(q)\,dq}.$$

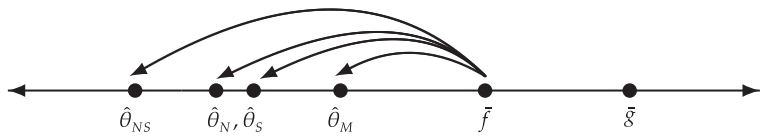
The symmetric estimate $\hat{\theta}_S^*$ is defined as a special case of Equation (5) when the decision analyst considers all judges to be mavens and sets $\hat{p} = 1$ and $d = n$ in step 2 and uses the entire set of responses to estimate the regression in step 4.

Proposition 4. *The hedged weights $h_{private}^*(\hat{q}, s_{\hat{q}}, d)$ and $h_{shared}^*(\hat{q}, s_{\hat{q}}, d)$ minimize the expected squared error in the estimated average private signal $\hat{\bar{t}}^*$ and shared signal $\hat{s}^*$, respectively.*

Aside from hedging, the decision analyst may also want to correct the pivoted crowd estimate to avoid logically impossible values. For example, if individuals are estimating the probability of a binary event, it would be reasonable to Winsorize the pivoted estimate at 0 and 1.

To conclude, the general procedure specified in Equation (5) allows the decision analyst to implement pivoting in each of the S, N, and NS structures. We next propose a simpler alternative that can remove shared-information bias in all three structures.

Figure 3. Pivoting Procedure to Calculate the Crowd Estimate in Different Information Structures for Large n When the Parameters Are Known and $pw > 0$



If w, p known:

$$\hat{\theta}_M = \bar{f} + (\bar{f} - \bar{g})$$

$$\hat{\theta}_S \approx \bar{f} + \frac{1}{w}(\bar{f} - \bar{g})$$

$$\hat{\theta}_N = \bar{f} + \frac{1}{p}(\bar{f} - \bar{g})$$

$$\hat{\theta}_{NS} \approx \bar{f} + \frac{1}{pw}(\bar{f} - \bar{g})$$

Notes. In the N and M procedures, the equations given are exact for any n. In the S and NS procedures, the figure approximates the exact formulas for $\hat{\theta}_S$ and $\hat{\theta}_{NS}$ given by Equation (4).

4.2. A Minimal Pivoting Procedure

If we set both parameters $w = p = 1$, we obtain a minimal (M) pivoting procedure. The minimal pivoting estimate, defined as $\hat{\theta}_M \equiv \bar{f} + (\bar{f} - \bar{g}) = 2\bar{f} - \bar{g}$, pivots by a factor of one and provides a lower bound pivot (assuming a large crowd) for the other three procedures as shown in Figure 3. For a small crowd, this may not hold because of the weight $(1 - \alpha)$ given to the estimate $\hat{s}$ of the shared signal in Equation (3). However, as crowd size n increases, α converges to one, and the pivoted estimate is approximated by the estimate $\hat{\bar{t}}$ of the average private signal in Equation (2), reproduced in the figure for each structure. Since both w and p are, by definition, less than or equal to one, the factors $1/w$, $1/p$, and $1/pw$ in the S, N, and NS estimates, respectively, are all greater than or equal to one, leading to a pivot at least as large as that in the minimal procedure.

Because minimal pivoting adjusts away from the average judgment in the correct direction, putting less weight on the shared information, we can expect it to outperform the average judgment. This advantage of minimal pivoting is summarized in the following proposition.

Proposition 5. *If $wp > 0$, then the estimate $\hat{\theta}_M$ obtained from the minimal pivoting procedure provides a lower expected squared error than the simple average $\bar{f}$ with a sufficiently large crowd in each of the three information structures.*

Minimal pivoting thus offers a simple variant of the general pivoting procedure that does not require parameter estimates for w or p. We expect that it will at least partially remove shared-information bias in a large crowd in each of the S, N, and NS structures, thus making it robust to differences in the information environment. Although the minimal procedure pivots less than the optimal amount in these three settings, there are reasons to use it instead of a tailored procedure. First, the three information structures we have studied are only idealized forms; the real setting is likely to be more general, in which case it is virtually impossible to develop a pivoting procedure that produces the GPE. The analyst can make a guess about the closest idealization, but if this guess is incorrect, the resulting pivot can be unboundedly inaccurate. Meanwhile, there is a limit to how far minimal pivoting might deviate from the average judgment, which protects it from very large errors. Second, judges may deviate from the behavioral model we have specified, may misperceive w and p, and may misidentify shared and private information. All of these deviations could cause the regression to estimate the wrong w and lead to over-pivoting or pivoting in the wrong direction. Again, these errors can be unboundedly large, imposing a significant risk on applying the model. For these reasons, minimal pivoting provides a conservative alternative that will often outperform the simple average but is not prone to egregious mistakes.

In the next sections, we present the results of four studies that investigate the accuracy of the individual and crowd estimates in several settings and demonstrate the benefits of the pivoting procedures in removing shared-information bias and improving the crowd estimate.

5. Study 1: Testing Pivoting in Different Information Settings

Study 1 focuses on two critical questions: (1) Does the behavioral model of judgments provide a reasonable characterization of participants' responses? and (2) Does the pivoting method offer an effective technique for producing an accurate estimate of the true mean of the variable being estimated, especially in comparison with simple averaging? To answer these questions, we developed carefully controlled information settings that correspond closely to our theoretical framework. This allows us to study how participants respond to specific private and shared signals and how pivoting aggregates information empirically.

Participants predicted how many heads would appear in 100 flips of a *biased* two-sided coin. The bias (i.e., the probability of heads) was unknown to participants, who were told that it could be any number between 1% and 99%. In the notation of Section 3, the

binomial random variable X was the realized number of heads in the $N = 100$ flips of the coin, and the true mean θ was the bias of the coin. At the start of the experiment, θ was drawn from a uniform distribution on $[0.01, 0.99]$, and this realized bias was then used to generate sample flips that made up the information that served as a basis for judgments. Shared information was comprised of sample flips that everyone saw, and private information was comprised of sample flips seen by only one or some people. We examined nine information settings, each with a different group of participants as shown in Table 1. Participants in each setting repeated the task for eight separate coins with different biases and sample flips. Altogether, this gave us 72 coins with which to test the pivoting method.

In the S structure, the parameter w is varied through the number of shared and private flips. The weight of $w = 0.25$ corresponds to a condition with nine commonly observed flips and three private flips (each judge receives the judge's own set of private flips). The weight of $w = 0.5$ corresponds to a condition with nine commonly observed flips and nine private flips, and the weight of $w = 0.75$ corresponds to a condition with three commonly observed flips and nine private flips. In the NS structure, we set the weight at $w = 0.5$ and varied the proportion p of mavens who each received their own nine private flips in addition to the nine commonly observed flips. The N structure was similar except that the nine additional flips were shared by all experts. Therefore, the NS structure contains much more information than the N structure because in the former structure there is a large quantity of private flips distributed across mavens. Because of this informational disparity, we expect that pivoting will have the most benefit in the S and NS structures.

5.1. Methods

5.1.1. Participants. Responses were gathered by running a number of online forecasting challenges on Amazon Mechanical Turk, each with a targeted size of 100 participants. Participants were recruited in three batches, one for each of the information structures. Within a batch, participants were randomly assigned to a low, medium, or high w setting (S) or to a low, medium, or high p setting (NS and N). In the latter two structures, participants were also randomly assigned to be *laypeople*, *experts*, or *mavens* as appropriate in proportion to p. Participants were quizzed to make sure they understood the task and response scales—330 potential participants "screened out" and never performed our task. An additional 222 participants started the study but did not finish it, and 1,030 participants ($M_{age} = 34$, 60% male) provided complete data and were included in the analysis. They spent an average of 10 minutes on the task and were compensated \$0.50 plus an average bonus of \$1.19.

5.1.2. Procedure. Participants were told that they would be participating in eight separate coin prediction tasks and that a new coin (with chances of heads anywhere between 1% and 99%) would be randomly selected for each task. To help develop an intuition for the task, they were told the following: *We will be showing you flips from biased coins. A biased coin is one that is tilted in favor of coming up either heads or tails. In case you're curious, these are virtual coins. The flipping is done by computer using a random number generator*.

They were also informed that after the experiment we would randomly select one of the coins and flip it 100 times (virtually) to calculate their bonus payment. For their forecast of the number of heads, they started with \$1, and five cents were subtracted for every unit by which the forecast differed from the actual number of heads observed in 100 new flips of the coin. In addition, the same payment scheme was used to reward their accuracy in guessing the average prediction of others. Bonuses for each response were not allowed to fall below zero.

Participants provided answers for one practice coin to become familiar with the task and eight additional coins presented in a random order. For a given participant, the coins all shared the same information structure as well as the same w and p values. The coins differed in their true biases θ and the samples generated based on that θ. In the N and NS structures, participants kept the same role as an expert, maven, or layperson for all coins (role labels were not communicated to them). Participants in the layperson role were not aware that other participants would be getting additional flips, but participants in the maven (expert) roles were informed of the proportion of participants who were mavens (experts). Participants were asked to provide their own best forecast of the number of heads and a guess of the average forecast of others on sliding scales with an initial position of 50 (see Figure 4). In the N condition, the label for private flips was changed to "Additional Flips" to help clarify that these flips were shared by the fraction of participants who observed them. After completing this task for all eight coins, participants answered several demographic questions.

5.2. Results

5.2.1. Is There Evidence for the Behavioral Model? The experiment allows us to control the signals participants receive and to compare their responses with normative benchmarks. Given the information that each participant saw, we imputed the relative weight that the participant placed on private versus shared information. For all experts and mavens i, we estimated the implied weight on private information for each response r_i according to $(r_i - s)/(\tau_i - s)$. Implied weights were not calculated for laypeople, who receive only a shared signal s, or for cases in which s and τ_i

Table 1. Study 1 Experimental Design and Number of Participants in Each Condition

	Nested–symmetric			Symmetric	Nested		
	$p = 0.25$	$p = 0.5$	$p = 0.75$	$p = 1$	$p = 0.25$	$p = 0.5$	$p = 0.75$
$w = 0.25$				119			
$w = 0.5$	111	112	117	101	112	119	114
$w = 0.75$				125			

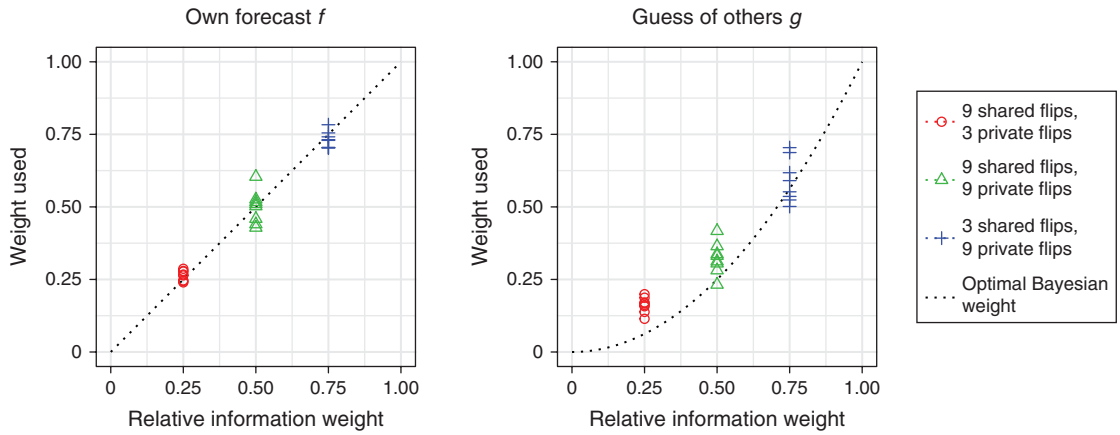
were equal. Finally, weights outside the range $[0, 1]$ were Winsorized (i.e., set equal to the nearer of zero and one) since responses outside the interval suggest that the two signals were not being combined.

We first consider the average of these implied weights across all participants for each coin in the S information structure, where we varied the true relative information weight w across conditions. The left panel of Figure 5 shows the average weights implied by the individual forecasts f_i and the optimal weight w, represented by the dashed identity line. Not surprisingly, participants combined their private and shared information in a way that was close to optimal when providing their own forecast. The correlation between the average implied weight and the Bayesian optimal weight for forecasts across the 24 coins was 0.98.

However, to be effective, the pivoting method also requires judges to guess others' forecasts accurately, which depends crucially on their ability to *think vicariously*. As a result, an equally important question is whether participants were able to correctly use available information when guessing the average forecast of the other participants. In other words, in the S structure, for example, did they, on average, place a weight of w^2 on their private signal and a weight of $(1 - w^2)$ on the shared signal when reporting g_i? The right panel

Figure 4. Example of the Experimental Interface for a Maven in the NS Structure

Figure 5. Average Winsorized Implied Weights Placed on Private Signal t_i Relative to the Shared Signal s Across the 24 Coins in the S Information Structure



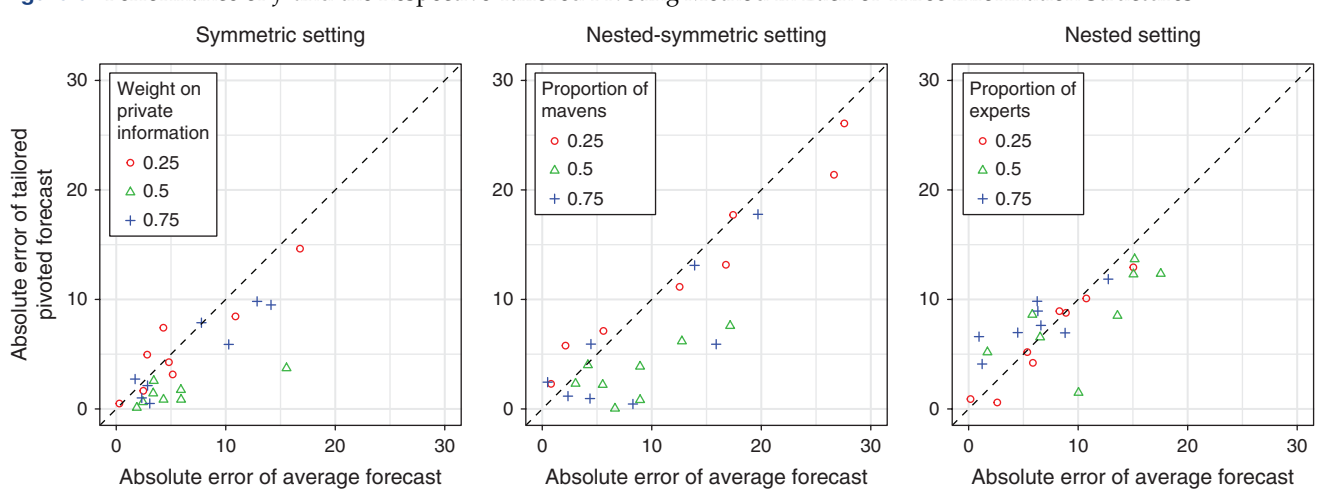
of Figure 5 shows the average weights implied by g_i across judges for each coin and the dashed optimal weight w^2 from Proposition 1. Although the individual weights varied substantially, participants, on average, used weights that were close to optimal here as well. The correlation between the average implied weight and the Bayesian optimal weight for guessing others across the 24 coins was 0.96.

Next, we considered the implied weights in the NS and N structures, where we fixed $w = 0.5$ and varied the proportion p. On average, experts and mavens put a weight of 0.29 on their private information in predicting others, compared with an average optimal weight of 0.19 across the coins in these settings. Contrary to the theory, participants were insensitive to p ($r = -0.01$ across the two structures). One potential reason for this is that a constrained optimal range (between 0.06 and 0.38) in these settings may have made it difficult to detect a relationship. Also, participants may have had difficulty coming up with a single estimate that aggregates over laypeople and experts (mavens), leading to noisy estimates.

5.2.2. Does the Pivoting Method Outperform Simple Averaging of Forecasts? Finally, and most importantly, we test how well the pivoting method performs at recovering the true mean θ for each coin. We applied the four pivoting algorithms $\hat{\theta}^*_S$, $\hat{\theta}^*_{NS}$, $\hat{\theta}^*_N$, and $\hat{\theta}_M$ to the response data, comparing the accuracy of the pivoted estimates against the accuracy of the simple average forecast $\bar{f}$. Figure 6 displays the absolute differences between each of these crowd estimates and the true mean θ for each coin, using the appropriate pivoting method in each information structure. Points below the dotted line indicate coins for which the absolute error of $\bar{f}$ was higher than the absolute error of the pivoted estimate $\hat{\theta}^*$ for that respective structure.

In the S structure, the pivoted estimate $\hat{\theta}^*_S$ had a significantly lower absolute error than $\bar{f}$ in estimating θ ($|\hat{\theta}^*_S - \theta| = 4.02$ ($s.d. = 3.81$) $< |\bar{f} - \theta| = 6.06$ ($s.d. = 4.76$), paired $t(23) = 3.43$, $p = 0.002$), and outperformed the

Figure 6. Performance of $\bar{f}$ and the Respective Tailored Pivoting Method in Each of Three Information Structures



average forecast for 19 of the 24 coins. In the NS structure, the pivoted estimate $\hat{\theta}^*_{NS}$ also had a significantly lower absolute error than $\bar{f}$ in estimating θ ($|\hat{\theta}^*_{NS} - \theta| = 7.48$ ($s.d. = 7.22$) $< |\bar{f} - \theta| = 10.25$ ($s.d. = 7.78$), paired $t(23) = 3.52$, $p = 0.002$), outperforming the average forecast for 18 of the 24 coins. In the N structure, however, the pivoted estimates provided only a small improvement, reducing the average absolute error from $|\bar{f} - \theta| = 7.91$ ($s.d. = 4.99$) to $|\hat{\theta}^*_N - \theta| = 7.63$ ($s.d. = 3.71$, paired $t(23) = 0.42$, $p = 0.68$) and outperforming the average forecast for only 13 of the 24 coins. This lower performance is perhaps unsurprising because of the inherently restricted information in this setting—there are only two total signals available to the judges in the N structure, so even perfectly removing shared-information bias offers limited benefit in forecast accuracy, particularly when the proportion of experts is high.

In practice, the decision analyst may not know the appropriate information structure, so to test the robustness of the pivoting method, we studied the performance of each estimation procedure across all three structures. We also considered the average performance of the individual forecasts f_i for each coin. The average absolute errors for each estimation procedure are displayed in Table 2. First, we note the unsurprising result that the simple wisdom of crowds offers a sizable improvement in accuracy. Over all structures, the average forecast reduced error by 32% relative to the individual forecasts. If the decision analyst chooses the correct tailored procedure for each of the three structures in Table 2 (e.g., $\hat{\theta}^*_S$ in S), pivoting achieves an average error of 6.38, representing a 46% reduction in error relative to the individual forecasts. Minimal pivoting performs nearly as well with an average error of 6.59—a 44% reduction in error relative to the individual forecasts.

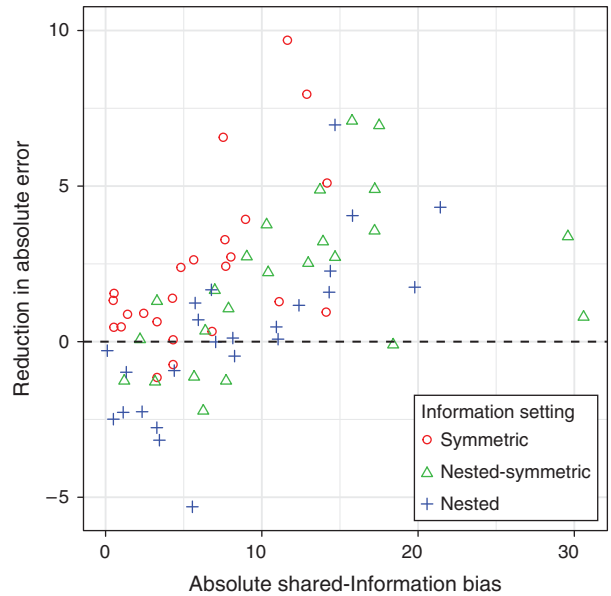
Additionally, we studied the performance of the global posterior expectation $\hat{\theta}_{GPE}$ for each coin and summarized the results in Table 2. These errors quantify the theoretically achievable accuracy of the estimate if all signals were directly available to the decision analyst, offering a useful benchmark against which to compare pivoting. Over all 72 coins, minimal pivoting achieved 29% of this potential improvement (the gap between the accuracy of averaging and the accuracy of the GPE), ranging from 10% of the theoretically achievable error reduction in the N structure to 57% of the theoretically achievable error reduction in the S structure.

Table 2. Performance of Different Estimation Procedures in the Different Information Structures

	Information structure			Over all
	Symmetric	Nested–symmetric	Nested	structures
Estimation procedure				
f_i	10.84	13.83	10.78	11.82
$\bar{f}$	6.06	10.25	7.91	8.07
θ^*_S	4.02	7.65	7.72	6.46
θ^*_{NS}	4.86	7.48	8.04	6.79
θ^*_N	3.33	8.05	7.63	6.34
$\hat{\theta}_M$	3.76	8.34	7.68	6.59
$\hat{\theta}_{GPE}$	2.02	1.41	5.51	2.98

Notes. Each entry in the table provides the average of the absolute distance between the estimate $\hat{\theta}$ and the true θ across the set of 24 coins in that structure. The final row provides the average over all 72 coins considered.

Figure 7. Reduction in Absolute Error Achieved by Minimal Pivoting (Relative to the Simple Average) Versus Absolute Shared-Information Bias for All 72 Coins



Finally, we studied the improvement of minimal pivoting as a function of the shared-information bias that was present for each coin, displayed in Figure 7. When the shared signal is highly accurate with near-zero bias, pivoting cannot offer any improvement over averaging and may even reduce accuracy in the crowd estimate because of noise introduced by its added complexity. However, as the shared-information bias grows larger, we observe steadily increasing average benefits from the pivoting procedure. In practice, of course, the shared-information bias is not known a priori. The results show that the pivoting method, on average, leads to improvement in all three information structures with strongest performance in the S structure, where the ratio of private to shared information is highest.

6. Study 2: Comparing Incentive Schemes

We designed a follow-up study to answer two questions: (1) How does the minimal pivoting procedure compare with existing alternatives, such as a

Table 3. Study 2 Experimental Design and Number of Participants in Each Condition

	Elicit guess of others g_i?					
	No			Yes		
	$w = 0.25$	$w = 0.5$	$w = 0.75$	$w = 0.25$	$w = 0.5$	$w = 0.75$
Incentives						
Accuracy	55	59	62	57	66	70
Winner-take-all	71	66	66	68	46	60

winner-take-all forecasting contest (Lichtendahl et al. 2013) for removing shared-information bias and increasing forecast accuracy? and (2) Does the addition of the guessing others question affect the individual forecasts in a significant way? To address these questions, we designed a 2 (payment incentive: accuracy versus winner take all) × 2 (guess others: yes versus no) × 3 (information weight: 0.25 versus 0.5 versus 0.75) experiment with participants randomly assigned to one of the 12 conditions. The forecasting task corresponded to the symmetric condition in Study 1. Participants repeated the task for eight separate coins.

6.1. Participants and Procedure

Participants were recruited in a single batch on Amazon Mechanical Turk (1,356 individuals started the survey, and 1,208 completed it). As in Study 1, several screens were used to ensure that participants understood the task and the payment scheme for their condition. As a result, 412 participants were screened out and never completed the main task. Because of the importance of the response format and incentive scheme, an additional 50 responses with duplicate IP addresses were removed, leaving 746 participants that were subjected to analysis ($M_{age} = 36$, 57% male) as displayed in Table 3. In each accuracy–incentives condition, participants were paid a bonus of \$1 minus five cents for each unit of absolute error for both their forecast of the number of realized heads in the 100 new flips and their guess of the average forecast of others when applicable. In each winner-take-all condition, the participant whose forecast was closest to the realized number of heads (with ties broken randomly) was paid a \$50 bonus. Participants spent an average of 11 minutes completing the survey and received a fixed \$0.50 payment plus an average bonus of \$1.23.

6.2. Results

Altogether, the design provides 96 coins with which to compare elicitation procedures and incentives. To avoid low cell sizes, we collapsed across the three levels of information weight. As shown in Table 4, the simple average $\bar{f}$ improved accuracy substantially over the average individual f_i. Across all coins, averaging reduced error by 47% compared with the average individual judge. There were no significant differences across conditions.

Table 4. Mean Absolute Errors $|\hat{\theta} - \theta|$ Across the 24 Coins for Different Crowd Estimation Procedures

	Accuracy incentives		Winner-take-all	
	Elicit f_i only	Elicit f_i and g_i	Elicit f_i only	Elicit f_i and g_i
Estimation procedure				
f_i	10.19	10.61	10.17	10.74
$\bar{f}$	6.22	6.11	5.99	5.82
$\hat{\theta}_M$	—	4.20	—	4.00
$\hat{\theta}_{GPE}$	2.11	2.44	2.74	2.12

Next, to compare aggregation methods, we included only the 48 coins for which participants guessed the estimates of others. We employed a mixed ANOVA model with aggregation method as a within-participant factor and payment incentives as a between-participants factor. As expected, minimal pivoting (mean improvement = 62.1%, $s.d.$ = 27.0%) outperformed the simple average (mean improvement = 48.4%, $s.d.$ = 25.8%), $F = 19.1$, $p < 0.001$. There was no effect of payment incentive ($F = 0.14$, $p = 0.712$) and no interaction ($F = 0.01$, $p = 0.904$).

Replicating Study 1, pivoting provided a large and clear advantage over simple averaging. Winner-take-all incentives provided an additional 3% error reduction compared with the average judge although this result was not significant. Of course, judges in the present study were untrained and inexperienced. A competitive crowd may be more effective with experienced judges who, over time, learn that they can obtain higher expected payoffs by shading responses toward their private information.

7. Study 3: Estimating Grocery Prices

Studies 3 and 4 tested the pivoting method using real-world stimuli. For Study 3, we purchased 10 different bundles of nonperishable grocery items at a Target store near Duke University. Bundles were composed of three items, which were displayed on a table so that participants had the opportunity to pick up and physically inspect them while providing their answers. Examples of items include a bottle of 190 L'il Critters Gummy Vites Sour Complete multivitamins (\$10.93), a 5-oz. can of Wild Planet wild albacore tuna in extra

virgin olive oil ($4.19), and an 11-oz. bag of Stauffer's Animal Crackers ($1.00).

7.1. Participants and Procedure

We recruited 49 volunteers passing through the student union to estimate the total price of the items in each bundle. Participants were compensated with a fixed payment of $5. For each bundle, participants estimated the total cost to purchase the three items at regular price and then guessed the average estimate that would be given by all of the other participants. Actual prices of the bundles ranged from $2.01 to $33.21, with a mean price of $15.00 and a median price of $13.62.

7.2. Results

For each bundle, we compared individual judgments f_i, the average judgment $\bar{f}$, and the estimates θ^* from the S, N, NS, and M pivoting procedures to the true price θ at the store. The mean absolute judgment error of a single individual ($\sum_{i=1}^{49} |f_i - \theta|$) across the 10 questions was $8.10. The average percentage improvement over the mean individual judgment error was 5.92% for $\bar{f}$. Each of the pivoting procedures yielded significantly better average percentage improvements than $\bar{f}$: 11.77% for the M procedure $\hat{\theta}_M$ (paired $t(9) = 2.17, p = 0.058$), 13.34% for the S procedure $\hat{\theta}^*_S$ (paired $t(9) = 2.27, p = 0.049$), 14.25% for the N procedure $\hat{\theta}^*_N$ (paired $t(9) = 2.22, p = 0.054$), and 13.91% for the NS procedure $\hat{\theta}^*_{NS}$ (paired $t(9) = 2.19, p = 0.056$). Because of the small number of grocery bundles, we also ran Wilcoxon signed rank tests for each comparison and obtained similar results ($p = 0.064$ for all four comparisons). The pivoted estimates $\hat{\theta}_M$, $\hat{\theta}^*_S$, $\hat{\theta}^*_N$, and $\hat{\theta}^*_{NS}$ each outperformed the average judgment $\bar{f}$ on 7 out of the 10 questions.

When guessing what other judges would say, on average, 70% of g_i were equal to the participant's own judgment f_i (yielding $\hat{p} = 0.30$). In addition, the average judgment and average guess about others were generally not far apart (mean $|\bar{f} - \bar{g}| = \$0.65, s.d. = \$0.54$). As a result, the pivoted estimates were fairly close to the simple average of participants' individual judgments. Despite this, the small adjustments made to the average judgment still allowed pivoting to outperform the average judgment, roughly doubling the improvement from simple averaging.

8. Study 4: Predicting NCAA Men's Basketball Tournament Games

In our fourth study, we tested the pivoting procedures in a binary-outcome setting by asking participants to estimate the probability that each team would win across 120 different games in the early rounds of the 2014, 2015, and 2016 NCAA Division I Men's Basketball Tournaments.

8.1. Participants and Procedure

Participants were recruited through ClearVoice Research in 2014 and through Amazon Mechanical Turk in 2015 and 2016 and were invited to participate in a web survey in exchange for a payment of $0.50 and a bonus of up to $2 based on the accuracy of their responses. On the consent screen, participants were informed that they would be predicting the outcomes of upcoming NCAA tournament ("March Madness") games and directed to participate if they were fans of college basketball or excited about the tournament.

For those predicting the "Round of 64," the games were divided into two sets so that each participant predicted 16 of the games. Participants predicting the "Round of 16" predicted all eight games in that round. For each of the 16 or eight games, participants selected a winner and assigned a probability on a scale from 50% to 100% that their chosen team would win. After providing this estimate, they then estimated the average probability of winning (on a scale from 0% to 100%) that other participants would assign to their chosen team. Each game was presented on a separate screen. Bonus payments were based on a rescaled quadratic scoring rule, which was displayed graphically above each game they were asked to predict. At the end of the survey, one game was randomly selected, and participants were paid based on the outcome of the game. Accuracy in guessing others was not compensated. Altogether, a total of 712 individuals participated in the study. Because the data were collected over time, the number of participants varied for different sets of games: $n = 73$ for 32 games (2014); $n = 164$ for eight games (2014); $n = 52$ for 16 games (2015); $n = 48$ for 16 games (2015); $n = 101$ for eight games (2015); $n = 65$ for 16 games (2016); $n = 56$ for 16 games (2016); $n = 80$ for eight games (2016). Participants spent an average of 10 minutes completing the survey and received an average bonus of $1.54.

8.2. Results

We used two different standards to examine the performance of estimation procedures: outcomes and market. The outcomes approach entails scoring estimates by comparing them with the binary outcomes across each of the 120 games using a proper rule. The Brier score for each game r is calculated by comparing a particular estimate y_r of team 1 winning with the indicator variable X_r for whether team 1 won the game or not according to the formula $(y_r - X_r)^2$. Lower Brier scores indicate more accurate estimates with a score of 0.25 corresponding to a "know-nothing" judge who reports 50% for every game. The market approach involves comparing the closeness of the different procedures with the probabilities implied by online betting websites. There is good reason to believe that bookmakers and participants in betting markets are more interested

in the games and possess more information than participants in the study. As a result, the market probabilities provide a natural benchmark to measure the accuracy of other estimates in the absence of objective probabilities for each game.

To obtain the market probability κ_r of team 1 winning in game r, we collected the decimal odds for bets on each team winning from the websites 5Dimes, Bovada, topbet, BetOnline, MyBookie.ag, BetDSI, BookMaker, GT Bets, SportBet, SportsBetting.com, and RealBet when available. For each game, there are two decimal odds quoted—one for betting on the event that team 1 will win and the other for betting on the event that team 2 will win. There is typically a small spread between the probabilities implied by each of these odds, which is part of the bookmaker's profit. Averaging these two decimals to get the probability of team 1 winning based on website k's odds accounts for the implied probabilities for the two teams summing to a number greater than one. Letting o_{1rk} be the decimal odds on team 1 and o_{2rk} be the decimal odds on team 2, we computed the implied probability κ_{rk} for each website k by taking the average of $1/o_{1rk}$ and $1 - 1/o_{2rk}$ and then computed the market probabilities κ_r of team 1 winning by averaging these implied probabilities κ_{rk} across all of the websites with available odds.

As expected, the Brier score of individual estimates f_i (mean score $= 0.232, s.d. = 0.121$) were improved significantly by combining responses across individuals (mean score for $\bar{f} = 0.189, s.d. = 0.122$), paired $t(119) = 41.9, p < 0.001$. Also as expected, the market probabilities were even better (mean score for $\kappa = 0.162, s.d. = 0.190$) than the average estimate, paired $t(119) = 3.01, p = 0.003$. Although each of the pivoted estimates showed improvement over the simple average (mean score $= 0.184$ for $\hat{\theta}^*_S$, 0.185 for $\hat{\theta}^*_N$, 0.183 for $\hat{\theta}^*_{NS}$, 0.186 for $\hat{\theta}_M$), none of these differences were statistically significant (e.g., paired $t(119) = 1.36, p = 0.177$ for the comparison between $\bar{f}$ and $\hat{\theta}_M$).

Comparing with market probabilities, individual estimates are again significantly improved by averaging (mean $|f_i - \kappa| = 0.198 >$ mean $|\bar{f} - \kappa| = 0.125$, paired $t(119) = 17.1, p < 0.001$). The pivoting procedures bring the crowd estimates further toward the market probabilities than the simple average. The mean difference $|\hat{\theta}_M - \kappa| = 0.116$ between the minimal estimates and the market probabilities was significantly lower than $|\bar{f} - \kappa|$ (paired $t(119) = 3.57, p < 0.001$). This difference was also significantly lower for the N estimates, mean $|\hat{\theta}^*_N - \kappa| = 0.112$ (paired $t(119) = 3.41, p < 0.001$), but was not significantly lower for the S estimates, mean $|\hat{\theta}^*_S - \kappa| = 0.116$ (paired $t(119) = 1.80, p = 0.075$), or the NS estimates, mean $|\hat{\theta}^*_{NS} - \kappa| = 0.119$ (paired $t(119) = 1.05, p = 0.298$).

The superior performance of the market probabilities is not surprising, because they represent the aggregate views of bookmakers and casinos with significant sums of money on the line. However, market probabilities may not be available in other prediction settings, and a decision analyst might need to choose between other estimation procedures, such as simple averaging and minimal pivoting. In our setting, these alternatives can be compared based on how close they come to the benchmark κ (i.e., the market probabilities). We found that minimal pivoting led to a directionally better Brier score and significantly improved estimates using this benchmark. Overall, the results replicate the finding that minimal pivoting can improve upon the average estimate and, moreover, can be beneficial in the aggregation of probabilities.

9. General Discussion

The fact that people rely on a combination of shared and private information poses a major challenge to the aggregation of judgments. It is well established that one of the most effective ways to combine judgments is to average responses. However, shared information induces correlation in judgment errors, which severely limits the ability to reduce error through averaging (Clemen and Winkler 1985). For any given question, shared information will typically bias all judgments in the same direction. This shared-information bias cannot be removed by averaging. We discussed this problem in the context of several information settings. Whereas past research has focused on an S structure in which each judge has both shared and private information, we have additionally studied two more structures—N and NS. In both of these structures, some judges are *laypeople* who have access to only shared information and others are *experts* or *mavens* who have access to more information (which is shared among experts in N and unique for each maven in NS).

For each structure, we developed a specialized pivoting procedure that, under idealized conditions, removes the entire shared-information bias. The procedures involve asking judges for both a judgment of the criterion and an estimate of what other judges will say on average. In principle, judges should weight shared information more heavily in predicting others than in predicting the criterion itself. We showed that this difference can be exploited by starting at the mean judgment $\bar{f}$ and pivoting away, by an appropriate amount, from the mean estimate of others $\bar{g}$. The logic behind pivoting is that the shared information contained in $\bar{g}$ is over-weighted in $\bar{f}$ because it is part of everyone's judgment. Pivoting away from $\bar{g}$ thus reduces the influence of the shared information to an appropriate level.

In practice, the implementation of pivoting is complicated by error-prone judges and general information structures that do not perfectly match the idealized forms that we have studied. We addressed the first problem by proposing a behavioral model that

accounts for judgmental error. By regressing the individual responses g_i on f_i, the decision analyst can recover the weight that judges put on private versus shared information, which is one of the inputs in determining how much to pivot. We address the problem of generalizability by offering a minimal pivoting procedure that makes no assumptions. This procedure entails less correction than the others and pivots by an amount equal to the distance between $\bar{f}$ and $\bar{g}$, such that $\hat{\theta}_M = 2\bar{f} - \bar{g}$.

We tested the three specialized procedures and minimal pivoting in four studies and found that pivoting consistently outperformed the simple average. In Study 1, we constructed the settings, so we knew which procedure should perform the best. Although overall the tailored procedures performed best in Study 1, all of the procedures achieved similar levels of accuracy across studies. Moreover, it is also the case that some procedures pivot more than others, which suggests that there is some risk of overcorrecting for shared information if the incorrect procedure is used. The risk may be more serious in practice because real-world information structures are unlikely to perfectly match the idealized forms we have studied. We offer the minimal pivoting procedure as an alternative that carries the least risk of overcorrecting, is simple to implement, and achieves similar performance to the other procedures.

9.1. Pivoting as a Solution to the Shared-Information Problem

Pivoting may seem like an unlikely approach to aggregation because it appears to contradict the common wisdom that an aggregate answer should be inside the range of the estimates (in this case, $\bar{f}$ and $\bar{g}$). Ideally, judgments will bracket the truth by falling on both sides of it, and thus, error can be canceled out by averaging (Larrick and Soll 2006). Bracketing is more likely, to the extent that judges have different perspectives, rely on different information, and form judgments independently to avoid anchoring (Soll and Larrick 2009). One approach to extracting more crowd wisdom, therefore, has been to focus on selection methods that recruit sufficient diversity of thought and group processes that facilitate independent judgment (Larrick et al. 2012). However, it may not always be possible to assemble a crowd with sufficient diversity to overcome the shared-information problem. In such low-bracketing situations, averaging will make little headway in reducing error compared with the average individual. Pivoting can help here, adjusting the average using the additional question about others' opinions, which points to the location of the shared information.

The aforementioned discussion provides some insight into the strong relationship between the effectiveness of pivoting and the size of the shared-information bias (see Figure 7). To illustrate this, we calculated the bracketing rate for each of the 72 coins from Study 1. As expected, shared-information bias and bracketing were strongly negatively correlated ($r = -0.74$). When the shared-information bias was low (less than five units), two randomly selected judges bracketed the truth 44% of the time (two independent judges can be expected to bracket 50% of the time). This implies that there were similar numbers of judgments above and below the truth and that averaging eliminated a large proportion of the error. In contrast, when the shared-information bias was high (greater than 15 units), the bracketing rate dropped to 12%, leaving plenty of room for improving upon averaging, provided that pivoting adjusts in the correct direction.

9.2. Limitations and Future Directions

We do not expect pivoting to lead to improvement in every instance. It may be worse than simple averaging when the shared information happens, by chance, to be very accurate (this explains the left side of Figure 7). It also cannot be expected to outperform simple averaging when all information is either private or shared. In these cases, we expect pivoting to perform about as well as averaging. When all information is shared, judges are likely to predict that others will give the same judgment as their own. Even if they do deviate, it is unlikely to be systematically in a given direction. Thus, $\bar{f}$ and $\bar{g}$ will be close together, and pivoting will produce an estimate very similar to $\bar{f}$. The same is likely to happen when all information is private. Thus, in situations where pivoting is unlikely to help, the method is unlikely to result in much pivoting anyway. As long as judges use their information to provide reasonably sensible estimates of both the criterion and what others will say, pivoting can correct for shared-information bias to improve aggregate estimates. It is also possible that people may not be able to distinguish between private and shared information or may only be able to make limited distinctions between what is shared and what is unique to themselves. To the extent that this is true, the average guess about others will be similar to the average judgment, and pivoting will give a similar result to simple averaging.

In principle, the potential to improve accuracy by combining opinions is greatest when there are many signals dispersed throughout the crowd, such as when most judges are mavens. Indeed, the more mavens in the conditions of Study 1, the more pivoting excelled. This happened not only because there were more signals, but also because pivoting came closer to the GPE when there were more mavens—it did a better job of solving the shared information problem. This can be seen by calculating relative improvement statistics for the NS structure. Minimal pivoting captured

3%, 29%, 46%, and 57% of the potential improvement available above and beyond the accuracy of averaging when the proportion of mavens was 0.25, 0.50, 0.75, and 1 (the S structure), respectively. Also, pivoting did not help as much in the N structure where there were experts instead of mavens. Together, these observations suggest that there is a benefit to composing crowds that are more S-like by including more mavens. However, doing so would require identifying judges' types—a nontrivial task when data are scarce. Because mavens' unique information may cause them to express discordant views, it may be difficult to distinguish true mavens from noisy laypeople or, worse yet, crackpots. Incorporating the opinions of legitimate mavens while simultaneously discounting the opinions of less-informed judges poses a continuing challenge for future research.

9.3. Conclusion

We have developed a novel method for dealing with shared-information bias in the aggregation of judgments. By asking judges to estimate the average judgment of others, it is possible to separate out shared and private signals. The method proceeds by pivoting off the mean judgment in a direction away from the shared signal, which corrects for the fact that the simple average over-weights shared information. Tailored variations of pivoting optimize the procedure for different types of information settings. However, when in doubt about the information setting or response behavior, we have shown that a simple minimal pivoting procedure achieves most of the benefits without a high risk of pivoting too far.

Overcoming shared-information bias is difficult, and much of the headway in addressing it to date has been theoretical. The bias greatly limits the extent to which aggregating judgments can reduce error, so the potential rewards from mitigating it are substantial. Future research should develop new approaches, refine existing ones, and seek to identify what is most effective. The results we have reported for pivoting suggest that it is a worthy contender.

Acknowledgments

The authors are grateful to Yael Grushka-Cockayne and Casey Lichtendahl for their input, discussion, and feedback during the formative stages of this project. In addition, we would like to thank Manel Baucells, Bob Clemen, Bob Nau, Drazen Prelec, Bob Winkler, and three anonymous reviewers for providing thoughtful comments and suggestions.

Appendix A. Proofs

Proof of Proposition 1. The optimal judgment for a layperson is $f^*(s_1) = \mathbb{E}[X \mid \pi_0, s_1] = \mathbb{E}[\mathbb{E}[X \mid \theta] \mid \pi_0, s_1] = \mathbb{E}[\theta \mid \pi_0, s_1] = s$. Likewise, the optimal judgment for expert or maven i is $f^*(s_1, \tau_i) = \mathbb{E}[X \mid \pi_0, s_1, \tau_i] = \mathbb{E}[\mathbb{E}[X \mid \theta] | \pi_0, s_1, \tau_i] = \mathbb{E}[\theta \mid \pi_0, s_1, \tau_i] = (1-w)s + w\tau_i$. The optimal guess about others for layperson i is $g^* = \mathbb{E}[\sum_{j\neq i} f_j/(n-1) \mid \pi_0, s_1]$. Conditioning over the types of the other judges and assuming that all judges follow the strategies f^* above, $g^* = \mathbb{E}[(1-p)f^*(s_1) + pf^*(s_1, \tau_j) \mid \pi_0, s_1] = (1-w)s + w\mathbb{E}[\tau_j \mid \pi_0, s_1]$. Iterating expectations yields $\mathbb{E}[\tau_j | \pi_0, s_1] = \mathbb{E}[\mathbb{E}[\tau_j \mid \theta] | \pi_0, s_1] = \mathbb{E}[\theta \mid \pi_0, s_1] = s$ and $g^* = s$. Likewise, the optimal guess of others for expert or maven i is $g^* = \mathbb{E}[\sum_{j\neq i} f_j/(n-1) \mid \pi_0, s_1, \tau_i]$. Again conditioning over the types of the other judges and substituting the optimal reporting strategies f^* above, $g^* = \mathbb{E}[(1-p)f^*(s_1) + pf^*(s_1, \tau_j) \mid \pi_0, s_1, \tau_i]$. In the N structure, $\tau_j = \tau_i$, so this simplifies to $g^* = (1-pw)s + pw\tau_i$. In the NS structure, $\mathbb{E}[\tau_j \mid \pi_0, s_1, \tau_i] = \mathbb{E}[\mathbb{E}[\tau_j \mid \theta] \mid \pi_0, s_1, \tau_i] = \mathbb{E}[\theta \mid \pi_0, s_1, \tau_i] = (1-w)s + w\tau_i$, so $g^* = \mathbb{E}[(1-p)s + p((1-w)s + w\tau_j) \mid \pi_0, s_1, \tau_i] = (1-pw^2)s + pw^2\tau_i$.

Proof of Proposition 2. In the NS structure (recall that S is a special case of NS with $p = 1$), if $pw > 0$, then $\bar{f} = (1-pw)s + pw\sum_{i=1}^{pn}(t_i/(pn))$, $\bar{g} = (1-p^2w^2)s + p^2w^2\sum_{i=1}^{pn}(t_i/(pn))$, and $\hat{\theta}_{NS} = ((npw)/(npw+1-w))(\bar{f} + (1/(pw))(\bar{f} - \bar{g})) + ((1-w)/(pnw+1-w))(\bar{f} + (1/(1-pw))(\bar{g} - \bar{f})) = ((1-w)/(pnw+1-w))s + ((npw)/(npw+1-w))\sum_{i=1}^{pn}(t_i/(pn))$. We can then write $\hat{\theta}_{NS} = (ms + l\sum_{i=1}^{pn} t_i)/(m+lpn) = \mathbb{E}[\theta \mid \pi_0, s_1, t_1, \ldots, t_{pn}]$, the GPE of X. If $pw = 0$, then either $p = 0$ and s_1 is the only signal observed by any judge or $w = 0$ and $\mathbb{E}[\theta \mid \pi_0, s_1, t_1, \ldots, t_{pn}] = \mathbb{E}[\theta \mid \pi_0, s_1]$. In each case, the GPE of X is $s = \bar{g} = \hat{\theta}_{NS}$. In the N structure, if $pw > 0$ and $p < 1$, then $\bar{f} = (1-pw)s + pwt$, $\bar{g} = (1-p^2w)s + p^2wt$, and $\hat{\theta}_N = w(\bar{f} + ((1-pw)/(pw(1-p)))(\bar{f} - \bar{g})) + (1-w)(\bar{f} + (1/(1-p))(\bar{g} - \bar{f})) = (1-w)s + wt$. Then $\hat{\theta}_N = (ms + lt)/(m+l) = \mathbb{E}[\theta \mid \pi_0, s_1, t]$, the GPE of X. If $pw = 0$, then either $p = 0$ and s_1 is the only signal or $w = 0$ and $\mathbb{E}[\theta \mid \pi_0, s_1, t] = \mathbb{E}[\theta \mid \pi_0, s_1]$. In both cases, the GPE of X is $s = \bar{g} = \hat{\theta}_N$. If $p = 1$, then all judges are experts and $\hat{\theta}_N = \bar{f} = (ms + lt)/(m+l)$, the GPE of X.

Proof of Propositions 3 and 5. For each structure (recall S is a special case of NS with $p = 1$), we show that as $n \to \infty$ the squared error in estimating θ when $wp > 0$ is smallest when using the tailored pivoting procedure given in Equation (4), larger when using $\hat{\theta}_M$, and largest when using $\bar{f}$.

In the NS structure, $\bar{f} = (1-pw)s + pw\sum_{i=1}^{pn}(t_i/(pn)) + (1-p)\sum_{i=1}^{n-pn}(\delta_i/(n-pn)) + p\sum_{i=1}^{pn}(\varepsilon_i/(pn))$ and $\bar{g} = (1-p^2w^2)s + p^2w^2\sum_{i=1}^{pn}(t_i/(pn)) + (1-p)\sum_{i=1}^{n-pn}(\delta_i/(n-pn)) + p^2w\sum_{i=1}^{pn}(\varepsilon_i/(pn)) + p\sum_{i=1}^{pn}(\gamma_i/(pn))$. Then $\mathbb{E}[(\bar{f} - \theta)^2] = \mathbb{E}[\mathbb{E}[(((1-pw)m_0/m)(\mu_0 - \theta) + ((1-pw)m_1/m)(s_1 - \theta) + (pw/(pn))\sum_{i=1}^{pn}(t_i - \theta) + ((1-p)/(n-pn))\sum_{i=1}^{n-pn}(\delta_i - 0) + (p/(pn))\sum_{i=1}^{pn}(\varepsilon_i - 0))^2 \mid \theta]]$, where the outer expectation is taken over $\pi_0(\theta)$ and the inner expectation is taken over the signals and error terms. Conditional on θ, the first term is a constant, and all other terms are independent, so the inner expectation is $((1-pw)^2m_0^2/m^2)(\mu_0 - \theta)^2 + ((1-pw)^2m_0^2/m^2)\mathrm{Var}(s_1 \mid \theta) + (w^2/n^2)\sum_{i=1}^{pn}\mathrm{Var}(t_i \mid \theta) + (1/n^2)\sum_{i=1}^{n-pn}\mathrm{Var}(\delta_i) + (1/n^2)\sum_{i=1}^{pn}\mathrm{Var}(\varepsilon_i)$, and the full expression simplifies to $\mathbb{E}[(\bar{f} - \theta)^2] = ((1-pw)^2m_0^2/m^2)\sigma_0^2 + ((1-pw)^2m_1^2/m^2)(V_0/m_1) + (pw^2/(nl))V_0 + ((1-p)/n)\mathrm{Var}(\delta) + (p/n)\mathrm{Var}(\varepsilon)$. As $n \to \infty$, $\mathbb{E}[(\bar{f} - \theta)^2] \to (1-pw)^2((m_0^2\sigma_0^2 + m_1^2(V_0/m_1))/m^2)$. Next, $\hat{\theta}_M = 2\bar{f} - \bar{g} = (1 - 2pw + p^2w^2)s + (2pw - p^2w^2)\sum_{i=1}^{pn}(t_i/(pn)) + (1-p)\sum_{i=1}^{n-pn}(\delta_i/(n-pn)) + (2p - p^2w)\sum_{i=1}^{pn}(\varepsilon_i/(pn)) - p\sum_{i=1}^{pn}(\gamma_i/(pn))$. Then $\mathbb{E}[(\hat{\theta}_M - \theta)^2] = \mathbb{E}[\mathbb{E}[(((1-pw)^2m_0/m)(\mu_0 - \theta) + ((1-pw)^2m_1/m)(s_1 - \theta) + ((pw(2-pw))/(pn))\sum_{i=1}^{pn}(t_i - \theta) + ((1-p)/(n-pn)) \cdot \sum_{i=1}^{n-pn}(\delta_i - 0) + ((2p - p^2w)/(pn))\sum_{i=1}^{pn}(\varepsilon_i - 0) - (p/(pn))\sum_{i=1}^{pn}(\gamma_i - 0))^2 \mid \theta]] = \mathbb{E}[((1-pw)^4m_0^2/m^2)(\mu_0 - \theta)^2 + ((1-pw)^4m_1^2/m^2)\mathrm{Var}(s_1 \mid \theta) + (w^2(2-pw)^2/n^2)\sum_{i=1}^{pn}\mathrm{Var}(t_i \mid \theta)$

$+ (1/n^2)\sum_{i=1}^{n-pn}\mathrm{Var}(\delta_i) + ((2-pw)^2/n^2)\sum_{i=1}^{pn}\mathrm{Var}(\varepsilon_i) + (1/n^2)\cdot\sum_{i=1}^{pn}\mathrm{Var}(\gamma_i)] = ((1-pw)^4 m_0^2/m^2)\sigma_0^2 + ((1-pw)^4 m_1^2/m^2)\cdot(V_0/m_1) + ((pw^2(2-pw)^2)/n)(V_0/l) + ((1-p)/n)\mathrm{Var}(\delta) + (p(2-pw)^2/n)\mathrm{Var}(\varepsilon) + (p/n)\mathrm{Var}(\gamma)$. As $n\to\infty$, $\mathbb{E}[(\hat{\theta}_M-\theta)^2]\to(1-pw)^4((m_0^2\sigma_0^2 + m_1^2(V_0/m_1))/m^2)$. If $wp>0$, this outperforms the limiting squared error obtained by using $\bar{f}$.

Finally, $\hat{\theta}_{NS} = ((npw)/(npw+1-w))(\bar{f} + (1/(pw))\cdot(\bar{f}-\bar{g})) + ((1-w)/(pnw+1-w))(\bar{f} + (1/(1-pw))(\bar{g}-\bar{f})) = ((1-w)/(pnw+1-w))s + ((npw)/(npw+1-w))\cdot\sum_{i=1}^{pn}(t_i/(pn)) + (1-p)\sum_{i=1}^{n-pn}(\delta_i/(n-pn)) + (np/(npw+1-w))\cdot\sum_{i=1}^{pn}(\varepsilon_i/(pn)) + ((np^2w - np + p - pw)/((npw+1-w)\cdot(1-pw)))\sum_{i=1}^{pn}(\gamma_i/(pn))$ and $\mathbb{E}[(\hat{\theta}_{NS}-\theta)^2] = \mathbb{E}[\mathbb{E}[(((1-w)/(pnw+1-w))(m_0/m)(\mu_0-\theta) + ((1-w)/(pnw+1-w))\cdot(m_1/m)(s_1-\theta) + (w/(pnw+1-w))\sum_{i=1}^{pn}(t_i-\theta) + ((1-p)/(n-pn))\sum_{i=1}^{n-pn}(\delta_i-0) + (1/(pnw+1-w))\sum_{i=1}^{pn}(\varepsilon_i-0) + ((np^2w - np + p - pw)/(pn(npw+1-w)(1-pw)))\cdot\sum_{i=1}^{pn}(\gamma_i-0))^2 \mid \theta]] = \mathbb{E}[((1-w)^2m_0^2/((pnw+1-w)^2m^2))\cdot(\mu_0-\theta)^2 + ((1-w)^2m_1^2/((pnw+1-w)^2m^2))\mathrm{Var}(s_1\mid\theta) + (w^2/((pnw+1-w)^2))\sum_{i=1}^{pn}\mathrm{Var}(t_i\mid\theta) + (1/n^2)\sum_{i=1}^{n-pn}\mathrm{Var}(\delta_i) + (1/(pnw+1-w)^2)\sum_{i=1}^{pn}\mathrm{Var}(\varepsilon_i) + ((np^2w - np + p - pw)^2/(p^2n^2(npw+1-w)^2(1-pw)^2))\sum_{i=1}^{pn}\mathrm{Var}(\gamma_i)] = ((1-w)^2m_0^2/((pnw+1-w)^2m^2))\sigma_0^2 + ((1-w)^2m_1^2/((pnw+1-w)^2m^2))\cdot(V_0/m_1) + ((pnw^2)/(pnw+1-w)^2)(V_0/l) + ((1-p)/n)\cdot\mathrm{Var}(\delta) + (pn/(pnw+1-w)^2)\mathrm{Var}(\varepsilon) + ((np^2w - np + p - pw)^2/(pn(npw+1-w)^2(1-pw)^2))\mathrm{Var}(\gamma)$. If $wp>0$, $\mathbb{E}[(\hat{\theta}_{NS}-\theta)^2]\to 0$ as $n\to\infty$, which outperforms the limiting squared error of $\hat{\theta}_M$.

In the N structure, $\bar{f} = (1-pw)s + pwt + (1-p)\cdot\sum_{i=1}^{n-pn}(\delta_i/(n-pn)) + p\sum_{i=1}^{pn}(\varepsilon_i/(pn))$ and $\bar{g} = (1-p^2w)s + p^2wt + (1-p)\sum_{i=1}^{n-pn}(\delta_i/(n-pn)) + p^2\sum_{i=1}^{pn}(\varepsilon_i/(pn)) + p\sum_{i=1}^{pn}(\gamma_i/(pn))$. $\mathbb{E}[(\bar{f}-\theta)^2] = \mathbb{E}[\mathbb{E}[(((1-pw)m_0/m)(\mu_0-\theta) + ((1-pw)m_1/m)(s_1-\theta) + pw(t-\theta) + ((1-p)/(n-pn))\sum_{i=1}^{n-pn}(\delta_i-0) + (p/(pn))\sum_{i=1}^{pn}(\varepsilon_i-0))^2\mid\theta]]$. The inner expectation is $((1-pw)^2m_0^2/m^2)(\mu_0-\theta)^2 + ((1-pw)^2m_0^2/m^2)\mathrm{Var}(s_1\mid\theta) + p^2w^2\mathrm{Var}(t\mid\theta) + ((n-pn)/n^2)\mathrm{Var}(\delta) + (pn/n^2)\mathrm{Var}(\varepsilon)$, so $\mathbb{E}[(\bar{f}-\theta)^2] = ((1-pw)^2m_0^2/m^2)\sigma_0^2 + ((1-pw)^2m_1^2/m^2)(V_0/m_1) + p^2w^2(V_0/l) + ((1-p)/n)\mathrm{Var}(\delta) + (p/n)\mathrm{Var}(\varepsilon)$. As $n\to\infty$, $\mathbb{E}[(\bar{f}-\theta)^2]\to(1-pw)^2((m_0^2\sigma_0^2 + m_1^2(V_0/m_1))/m^2) + p^2w^2(V_0/l)$. Next, $\hat{\theta}_M = 2\bar{f} - \bar{g} = (1-2pw+p^2w)s + (2pw-p^2w)t + (1-p)\sum_{i=1}^{n-pn}(\delta_i/(n-pn)) + (2p-p^2)\sum_{i=1}^{pn}(\varepsilon_i/(pn)) - p\sum_{i=1}^{pn}(\gamma_i/(pn))$ and $\mathbb{E}[(\hat{\theta}_M-\theta)^2] = \mathbb{E}[\mathbb{E}[(((1-2pw+p^2w)m_0/m)(\mu_0-\theta) + ((1-2pw+p^2w)m_1/m)(s_1-\theta) + (2pw-p^2w)(t-\theta) + ((1-p)/(n-pn))\sum_{i=1}^{n-pn}(\delta_i-0) + ((2p-p^2)/(pn))\sum_{i=1}^{pn}(\varepsilon_i-0) - (p/(pn))\cdot\sum_{i=1}^{pn}(\gamma_i-0))^2\mid\theta]] = \mathbb{E}[((1-2pw+p^2w)^2m_0^2/m^2)(\mu_0-\theta)^2 + ((1-2pw+p^2w)^2m_1^2/m^2)\mathrm{Var}(s_1\mid\theta) + (2pw-p^2w)^2\mathrm{Var}(t\mid\theta) + (1/n^2)\sum_{i=1}^{n-pn}\mathrm{Var}(\delta_i) + ((2-p)^2/n^2)\sum_{i=1}^{pn}\mathrm{Var}(\varepsilon_i) + (1/n^2)\sum_{i=1}^{pn}\mathrm{Var}(\gamma_i)] = ((1-2pw+p^2w)^2m_0^2/m^2)\sigma_0^2 + ((1-2pw+p^2w)^2m_1^2/m^2)(V_0/m_1) + (2pw-p^2w)^2(V_0/l) + ((1-p)/n)\cdot\mathrm{Var}(\delta) + (p(2-p)^2/n)\mathrm{Var}(\varepsilon) + (p/n)\mathrm{Var}(\gamma)$. As $n\to\infty$, $\mathbb{E}[(\hat{\theta}_M-\theta)^2]\to(1-2pw+p^2w)^2((m_0^2\sigma_0^2 + m_1^2(V_0/m_1))/m^2) + (2pw-p^2w)^2(V_0/l)$.

Finally, $\hat{\theta}_N = \bar{f} + (1/p)(\bar{f}-\bar{g}) = (1-w)s + wt + (1-p)\cdot\sum_{i=1}^{n-pn}(\delta_i/(n-pn)) + \sum_{i=1}^{pn}(\varepsilon_i/(pn)) - \sum_{i=1}^{pn}(\gamma_i/(pn))$. If $p>0$ then $\mathbb{E}[(\hat{\theta}_N-\theta)^2] = \mathbb{E}[\mathbb{E}[((1-w)(m_0/m)(\mu_0-\theta) + (1-w)\cdot(m_1/m)(s_1-\theta) + w(t-\theta) + ((1-p)/(n-pn))\sum_{i=1}^{n-pn}(\delta_i-0) + (1/(pn))\sum_{i=1}^{pn}(\varepsilon_i-0) - (1/(pn))\sum_{i=1}^{pn}(\gamma_i-0))^2\mid\theta]] = \mathbb{E}[((1-w)^2m_0^2/m^2)(\mu_0-\theta)^2 + ((1-w)^2m_1^2/m^2)\mathrm{Var}(s_1\mid\theta) + w^2\mathrm{Var}(t\mid\theta) + (1/n^2)\sum_{i=1}^{n-pn}\mathrm{Var}(\delta_i) + (1/(p^2n^2))\sum_{i=1}^{pn}\mathrm{Var}(\varepsilon_i) + (1/(p^2n^2))\sum_{i=1}^{pn}\mathrm{Var}(\gamma_i)] = ((1-w)^2m_0^2/m^2)\sigma_0^2 + ((1-w)^2\cdot m_1^2/m^2)(V_0/m_1) + w^2(V_0/l) + ((1-p)/n)\mathrm{Var}(\delta) + (1/(pn))\cdot\mathrm{Var}(\varepsilon) + (1/(pn))\mathrm{Var}(\gamma)$. As $n\to\infty$, $\mathbb{E}[(\hat{\theta}_N-\theta)^2]\to(1-w)^2\cdot((m_0^2\sigma_0^2 + m_1^2(V_0/m_1))/m^2) + w^2(V_0/l)$.

Define $V(z)\equiv(1-z)^2((m_0^2\sigma_0^2 + m_1^2(V_0/m_1))/m^2) + z^2(V_0/l)$, which is a convex function of z. The posterior mean of θ conditional on the prior $\pi_0(\theta)$ and both observed signals s_1 and t is $(1-w)(m_0/m)\mu_0 + (1-w)(m_1/m)s_1 + wt$. Since this mean is the Bayes estimate of θ with respect to quadratic loss, it provides the minimal expected squared error, which is $\mathbb{E}[((1-w)s+wt-\theta)^2] = \lim_{n\to\infty}\mathbb{E}[(\hat{\theta}_N-\theta)^2] = V(w)$, which therefore must be less than or equal to both $\lim_{n\to\infty}\mathbb{E}[(\hat{\theta}_M-\theta)^2] = V(2pw-p^2w)$ and $\lim_{n\to\infty}\mathbb{E}[(\bar{f}-\theta)^2] = V(pw)$. Observe that $p(2-p)$ is a concave quadratic function of p that takes a maximum value of one at $p=1$, so $pw(2-p)\le w$ for $p\in[0,1]$ and $w\ge 0$. In addition, $pw\le pw(2-p)$ since $p\in[0,1]$ and $w\ge 0$. Then $pw\le pw(2-p)\le w$, and since $pw(2-p) = (1-p)pw + pw$, the convexity of V implies that $V(pw(2-p))\le(1-p)V(pw) + pV(w)$. $V(w)\le V(pw)$, so $V(pw(2-p))\le(1-p)V(pw) + pV(pw) = V(pw)$. This establishes our result that $V(pw) = \lim_{n\to\infty}\mathbb{E}[(\bar{f}-\theta)^2]\ge V(2pw-p^2w) = \lim_{n\to\infty}\mathbb{E}[(\hat{\theta}_M-\theta)^2]\ge V(w) = \lim_{n\to\infty}\mathbb{E}[(\hat{\theta}_N-\theta)^2]$.

Proof of Proposition 4. Define $\bar{t}^h\equiv\bar{f} + (1/h)(\bar{f}-\bar{g})$ and $\bar{t} = \sum_{i=1}^{pn}(t_i/(pn))$. The decision analyst wants to choose h to minimize the expected squared error $\mathscr{E}_{private}(h,q) = (\mathbb{E}[\bar{t}^h\mid\bar{t},s,q]-\bar{t})^2$, given the decision analyst's uncertainty about q. The optimally hedged weight is $h^*_{private} = \mathrm{argmin}_h\,\mathscr{D}_{private}(h)$, where $\mathscr{D}_{private}(h) = \int_0^1\mathscr{E}_{private}(h,q)f(q\mid\{(f_i,g_i)\}_{i=1}^n)dq$. Using the expressions for $\bar{f}$ and $\bar{g}$ from the proof of Proposition 2, we can write $\mathbb{E}[\bar{t}^h\mid\bar{t},s,q] = \bar{f} + (1/h)(\bar{f}-\bar{g}) = (1-q)s + q\bar{t} + (q/h)\cdot(1-q)(\bar{t}-s)$, $\mathscr{E}_{private}(h,q) = (\bar{t}-s)^2(1-q)^2(1-q/h)^2$, and $\mathscr{D}_{private}(h) = (s-\bar{t})^2\int_0^1(1-q/h)^2(1-q)^2\cdot f(q\mid\{(f_i,g_i)\}_{i=1}^n)dq$. The first-order condition (FOC) with respect to h yields $\int_0^1(q/h)q(1-q)^2f(q\mid\{(f_i,g_i)\}_{i=1}^n)dq = \int_0^1 q(1-q)^2f(q\mid\{(f_i,g_i)\}_{i=1}^n)dq$, or $h^*_{private} = (\int_0^1 q^2(1-q)^2f(q\mid\{(f_i,g_i)\}_{i=1}^n)dq)/(\int_0^1 q(1-q)^2f(q\mid\{(f_i,g_i)\}_{i=1}^n)dq)$. It is straightforward to show that $h^*_{private}$ is minimal.

Likewise, define $s^h\equiv\bar{f} + (1/(1-h))(\bar{g}-\bar{f})$. The decision analyst wants to minimize the expected squared error $\mathscr{E}_{shared}(h,q) = (\mathscr{E}[s^h\mid\bar{t},s,q]-s)^2$, given the decision analyst's uncertainty about q. In other words, the decision analyst wants to find $h^*_{shared} = \mathrm{argmin}_h\,\mathscr{D}_{shared}(h)$, where $\mathscr{D}_{shared}(h) = \int_0^1\mathscr{E}_{shared}(h,q)f(q\mid\{(f_i,g_i)\}_{i=1}^n)dq$. $\mathscr{E}[s^h\mid\bar{t},s,q] = (\bar{g}-h\bar{f})/(1-h) = s + q(\bar{t}-s)((q-h)/(1-h))$, $\mathscr{E}_{shared}(h,q) = q^2(\bar{t}-s)^2\cdot((q-h)/(1-h))^2$, and $\mathscr{D}_{shared}(h) = (s-\bar{t})^2\int_0^1((q-h)/(1-h))^2q^2f(q\mid\{(f_i,g_i)\}_{i=1}^n)dq$. The FOC is $\int_0^1 h(1-q)q^2f\cdot(q\mid\{(f_i,g_i)\}_{i=1}^n)dq = \int_0^1(1-q)q^3f(q\mid\{(f_i,g_i)\}_{i=1}^n)dq$, or $h^*_{shared} = (\int_0^1(1-q)q^3f(q\mid\{(f_i,g_i)\}_{i=1}^n)dq)/(\int_0^1(1-q)q^2f(q\mid\{(f_i,g_i)\}_{i=1}^n)dq)$, which is minimal.

Endnotes

[1] Alternatively, one could view μ_0 as part of the shared signal and assume a diffuse prior over θ by letting $m_0\to 0^+$. In that case, the linearity condition assumes that the GPE equals the minimum-variance unbiased linear combination of all signals that were observed. The judgment aggregation problem can then be viewed as a parameter estimation problem given the set of data $\{s_1,t_1,\ldots,t_K\}$.

[2] Specifically, the judge can be rewarded with the sum of $R_1(X,f_i)$ for the judge's own judgment and $R_2(\sum_{j\neq i}f_j/(n-1),g_i)$ for the judge's guess of others' judgments. Each scoring rule should be strictly proper in the sense that reporting the mean of the judge's subjective

distribution for X provides the judge with the highest expected score among all possible reports.

[3] The pivoting method works well when judgments are unbiased in the variable of interest. Distributions of judgment errors that satisfy this property are possible for variables with domains that are unbounded (e.g., normal), bounded on one side (e.g., gamma), and double-bounded (e.g., binomial). Alternative error models that have been considered in the existing literature may not satisfy this property. For example, logit models have been used to aggregate probability judgments (in which case judgments would be unbiased in log-odds but biased when measured as probabilities).

[4] The estimation of w^* here is analogous to the estimation of q^* with one modification—we know that w must be in $[0,1]$ by definition, and $w = q/p$, so q must be in the interval $[0,p]$. Since we are using $\hat{p}$ as an estimate of p, restricting the distribution of q to $[0,\hat{p}]$ ensures that the estimate w^* will stay in $[0,1]$.

References

Armstrong SJ (2001) Combining forecasts. Armstrong SJ, ed. *Principles of Forecasting: A Handbook for Researchers and Practitioners* (Kluwer Academic Publishers, Norwell, MA), 417–439.

Baron J, Mellers BA, Tetlock PE, Stone E, Ungar LH (2014) Two reasons to make aggregated probability forecasts more extreme. *Decision Anal.* 11(2):133–145.

Budescu DV, Chen E (2015) Identifying expertise to extract the wisdom of crowds. *Management Sci.* 61(2):267–280.

Budescu DV, Yu HT (2007) Aggregation of opinions based on correlated cues and advisors. *J. Behavioral Decision Making* 20(2): 153–177.

Chen K, Fine LR, Huberman BA (2004) Eliminating public knowledge biases in information-aggregation mechanisms. *Management Sci.* 50(7):983–994.

Clemen RT (1989) Combining forecasts: A review and annotated bibliography. *Internat. J. Forecasting* 5(4):559–583.

Clemen RT, Winkler RL (1985) Limits for the precision and value of information from dependent sources. *Oper. Res.* 33(2): 427–442.

Clemen RT, Winkler RL (1986) Combining economic forecasts. *J. Bus. Econom. Statist.* 4(1):39–46.

Frongillo RM, Chen Y, Kash IA (2015) Elicitation for aggregation. Bonet B, Koenig S, eds. *Proc. Twenty-Ninth AAAI Conf. Artificial Intelligence* (AAAI Press, Palo Alto, CA), 900–906.

Galton F (1907) Vox populi. *Nature* 1949(75):450–451.

Gigone D, Hastie R (1993) The common knowledge effect: Information sharing and group judgment. *J. Personality Soc. Psych.* 65(5):959–974.

Jose VRR, Winkler RL (2008) Simple robust averages of forecasts: Some empirical results. *Internat. J. Forecasting* 24:163–169.

Jurca R, Faltings B (2009) Mechanisms for making crowds truthful. *J. Artificial Intelligence Res.* 34(1):209–253.

Kim O, Lim SC, Shaw KW (2001) The inefficiency of the mean analyst forecast as summary of forecast of earnings. *J. Accounting Res.* 39:329–335.

Larrick RP, Soll JB (2006) Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Sci.* 52(1):111–127.

Larrick RP, Mannes AE, Soll JB (2012) The social psychology of the wisdom of crowds. Krueger JI, ed. *Social Psychology and Decision Making* (Psychology Press, New York), 227–242.

Lichtendahl KC, Grushka-Cockayne Y, Pfeifer PE (2013) The wisdom of competitive crowds. *Oper. Res.* 61(6):1383–1398.

Makridakis S, Winkler RL (1983) Averages of forecasts: Some empirical results. *Management Sci.* 29(9):987–996.

Mannes AE, Soll JB, Larrick RP (2014) The wisdom of select crowds. *J. Personality Soc. Psych.* 107(2):276–299.

McCoy J, Prelec D (2017) A statistical model for aggregating judgments by incorporating peer predictions. Working paper.

Miller N, Resnick P, Zeckhauser R (2005) Eliciting informative feedback: The peer-prediction method. *Management Sci.* 51(9): 1359–1373.

Ottaviani M, Sørensen PN (2006) The strategy of professional forecasting. *J. Financial Econom.* 81:441–466.

Page SE (2008) *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies* (Princeton University Press, Princeton, NJ).

Payne JW, Bettman JR, Johnson EJ (1993) *The Adaptive Decision Maker* (Cambridge University Press, New York).

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

Prelec D, Seung HS (2006) An algorithm that finds truth even if most people are wrong.

Prelec D, Seung HS, McCoy J (2017) A solution to the single-question crowd wisdom problem. *Nature* 541:532–535.

Soll JB (1999) Intuitive theories of information: Beliefs about the value of redundancy. *Cognitive Psych.* 38(2):317–346.

Soll JB, Larrick RP (2009) Strategies for revising judgment: How (and how well) people use others' opinions. *J. Experiment. Psych.: Learn., Memory, Cognition* 35(3):780–805.

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

Winkler RL (1981) Combining probability distributions from dependent information sources. *Management Sci.* 27(4):479–488.

Witkowski J, Bachrach Y, Key P, Parkes DC (2013) Dwelling on the negative: Incentivizing effort in peer prediction. *First AAAI Conf. Human Comput. Crowdsourcing*.

Yaniv I, Choshen-Hillel S, Milyavsky M (2009) Spurious consensus and opinion revision: Why might people be more confident in their less accurate judgments? *J. Experiment. Psych.: Learn., Memory, Cognition* 35(2):558–563.