Finding truth even if the crowd is wrong

Drazen Prelec[1,2,3], H. Sebastian Seung[3,4], and John McCoy[3]
[1]Sloan School of Management
Departments of [2]Economics, [3]Brain & Cognitive Sciences, and [4]Physics
Massachusetts Institute of Technology
Cambridge MA 02139
dprelec@mit.edu, seung@mit.edu, jmccoy@mit.edu

February 19, 2013

Over a hundred years ago Galton reported on the uncanny accuracy of the median estimate of the weight of an ox, as judged by spectators at a country fair [1]. Since then, the notion that the 'wisdom of the crowd' is superior to any individual has itself become a piece of crowd wisdom, raising expectations that web-based opinion aggregation might replace expert judgment as a source of policy guidance [2, 3]. However, distilling the best answer from diverse opinions is challenging when most people hold an incorrect view [4]. We propose a method based on a new definition of the best answer: it is the one given by respondents who would be least surprised by the true answer if it were revealed. Since this definition is of interest only when the true answer is unknown, algorithmic implementation is nontrivial. We solve this problem by asking respondents not only to answer the question, but also to predict the distribution of others' answers. Previously, it was shown that this secondary information can be used to create incentives for honest responding [5]. Here we prove that this information can also be used to identify which answer is the best answer by our new definition. Unlike multi-item analysis [6, 7] or boosting [8], our method can be applied to a unique question. This capability is critical in knowledge domains that lack consensus about which historical precedents might establish experts' relative track records. Unlike Bayesian models [9, 10, 11, 12, 13] our method does not require user-specified prior probabilities, nor does it require information sharing that might lead to "groupthink" [14]. An experiment demonstrates that the method outperforms algorithms based on democratic or confidence-weighted voting [15, 16, 17, 18, 19].

Imagine that you have no knowledge of U.S. geography, and are confronted with the question

Philadelphia is the capital of Pennsylvania: True or False ?

To find the answer, you pose the question to many people, trusting that the most common answer will be correct. Unfortunately, most people give the incorrect answer ("True"), as shown by the data in Figure 1a. Why is the majority wrong here? Someone who answers "True" may know only that Philadelphia is an important city in Pennsylvania, and reasonably conclude that Philadelphia is the capital. Someone who answers "False" likely possesses a crucial additional piece of evidence, that the capital is actually Harrisburg.

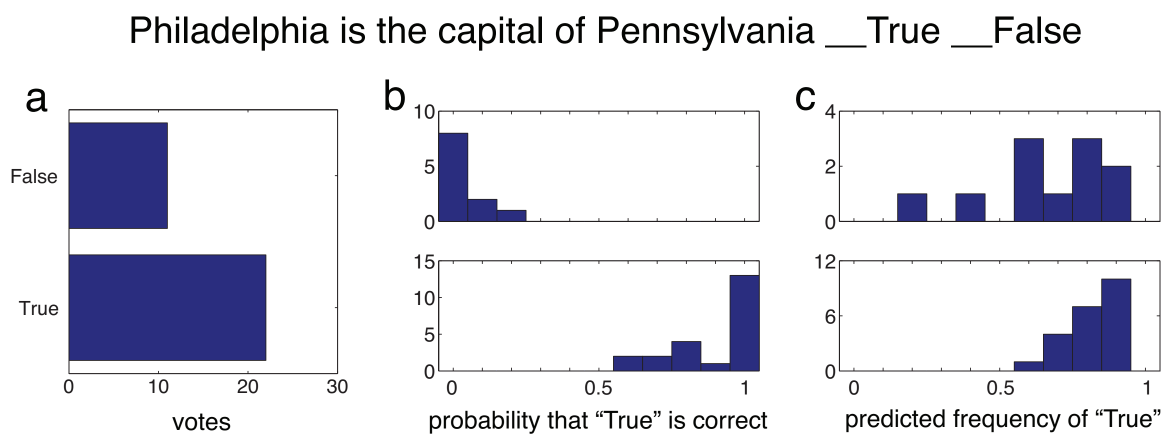

Figure 1: A question that voting (2) fails to answer correctly, while the LST principle (1) succeeds (data are from Study 3 described in text). **a** The wrong answer wins by a large margin in a democratic vote. **b** Respondents are asked to provide estimates of confidence (0.5 to 1), which are combined with their True/False answers to yield estimates of the probability (0 to 1) that Philadelphia is the capital. The histograms show that the correct minority (top) and the incorrect majority (bottom) are roughly equally confident in their answers, so weighting the votes by confidence does not change the outcome. **c** Respondents are asked to predict the frequency of the answer “True.” Those who answer “False” believe that others will answer “True” (top), but not vice versa (bottom). To apply the LST principle, we estimate from (a), $\Pr[X^s = \text{False}|\Omega = i^*] = 0.33$, $\Pr[X^s = \text{True}|\Omega = i^*] = 0.67$, and from (c), $\Pr[X^s = \text{True}|X^r = \text{False}] = 0.76$, $\Pr[X^s = \text{False}|X^r = \text{True}] = 0.22$. Inserting into (4) yields $\Pr[\Omega = i^*|X^r = \text{False}] / \Pr[\Omega = i^*|X^s = \text{True}] = 1.70$, showing that respondents endorsing the correct answer “False” would be less surprised by the truth.

This elementary example reveals a limitation of the one person, one vote approach. If each respondent’s answer is determined by the evidence available to her, the majority verdict will be tilted toward the most widely available evidence, which is an unreliable indicator of truth. The same bias is potentially present in real-world settings, when experts’ opinions are averaged to produce probabilistic assessments of risk, forecasts of key economic variables, or numerical ratings of research proposals in peer review. In all such cases, Galton’s method of counting opinions equally may produce a result that favors shallow information, accessible to all, over specialized or novel information that is understood only by a minority.

To avoid this problem, one might attempt to identify individuals who are most competent to answer the question, or who have the best evidence. A popular approach is to ask respondents to report their confidence [13], typically by a number between 0.5 (no confidence) and 1 (certainty). From their confidence and True/False answers one can infer respondents’ subjective probability estimates that, e.g., Philadelphia is the capital of Pennsylvania. Averaging these probabilities across respondents produces a confidence-weighted vote. This will improve on an unweighted vote, but only if those who answer correctly are also much more confident, which is neither the case in our example, nor more generally [4]. As shown by Figure 1b, the distribution of probabilities for those who answer “True” is approximately the mirror image of the distribution for those who answer “False.” Since confidence is roughly symmetric between the two groups, it cannot override the strong majority in favor of the wrong answer.

Rather than elicit a confidence estimate, our method will ask each respondent to predict how others will respond. For a True/False question, the prediction is a number between 0 and 1 indicating the fraction of respondents who will answer "True." As shown in Figure 1c, those who answer "True" to the Philadelphia question predict that most people will agree and answer "True." On the other hand, those who answer "False" tend to predict that most people will disagree and hence answer "True." This prediction presumably reflects superior knowledge: Respondents who believe that Harrisburg is the capital tend to realize that most people will not know this. The asymmetry between the distributions in Figure 1c is marked, suggesting that predictions of others' answers could provide a signal that is strong enough to override majority opinion. To make use of this information, however, we need a precise definition of the notion of best evidence and best answer.

Consider a probabilistic model in which the state Ω of the world is a random variable taking on values in the set $\{1,\ldots,m\}$ of possible answers to a multiple choice question. The answer X^r given by respondent r is likewise a random variable taking on values in the same set. We assume that the answer is based on a hidden variable, the "signal" or "evidence" available to the respondent. Each respondent reasons correctly from the evidence available to her, and answers honestly. We further assume that respondents with different answers have access to different evidence, and respondents who give the same answer do so based on the same evidence. Therefore a respondent r who answers k assigns probabilities $\Pr[\Omega = i|X^r = k]$ to possible answers i, and this function does not depend on r.

The true value i^* is unknown. If it were revealed, then, under the assumptions of the above model, the probability $\Pr[\Omega = i^*|X^r = k]$ would measure (inversely) the surprise of any respondent r who selected answer k. We define the "best answer" as the one given by respondents who would be "least surprised by the truth" (hereafter LST) or

$$\underset{k}{\operatorname{argmax}} \Pr[\Omega = i^*|X^r = k] \tag{1}$$

By our assumptions, these are also the respondents who possess the best evidence. The best answer to the Philadelphia question should be correct, as those who believe that Philadelphia is the capital will be more surprised by the truth than those who believe that Philadelphia is not the capital. The best answer is not infallible, because evidence unavailable to any respondent might tip the balance the other way. Nevertheless, on average it should be more accurate than the democratic principle, which selects the answer most likely to be offered by respondents:

$$\underset{k}{\operatorname{argmax}} \Pr[X^r = k|\Omega = i^*] \tag{2}$$

The obvious advantage of (2) is that the probabilities can be readily estimated from the relative frequencies of answers even without knowledge of i^*, as answers are by definition sampled from the true state-of-the-world (e.g., in which Harrisburg is the capital of Pennsylvania). As for (1), procedures do exist for eliciting subjective probability distributions over possible states i [13], but (1) requires these probabilities specifically for the unknown value $i = i^*$. To circumvent this difficulty, we note first that finding the maximum in principle (1) can be done by comparing $\Pr[\Omega = i^*|X^r = k]$ and $\Pr[\Omega = i^*|X^s = j]$ for all answers j and k. Using Bayes' Rule, the ratio between these probabilities can be rewritten as

$$\begin{aligned}
\frac{\Pr[\Omega = i^*|X^r = k]}{\Pr[\Omega = i^*|X^s = j]} &= \frac{\Pr[X^r = k|\Omega = i^*]}{\Pr[X^s = j|\Omega = i^*]} \frac{\Pr[X^s = j]}{\Pr[X^r = k]} \qquad (3)\\
&= \frac{\Pr[X^r = k|\Omega = i^*]}{\Pr[X^s = j|\Omega = i^*]} \frac{\Pr[X^s = j|X^r = k]}{\Pr[X^r = k|X^s = j]} \qquad (4)
\end{aligned}$$

In the last expression, $\Pr[\Omega = i^*|X^r = k]$ has been eliminated in favor of $\Pr[X^r = k|\Omega = i^*]$. The latter probability is the same as the one that appears in the majority answer (2), and can be estimated by the frequency of answer k. This simplification comes at the cost of introducing $\Pr[X^r = k | X^s = j]$. We will assume that this probability can be estimated by asking someone who answers j to predict the frequency of answer k (see Figure 1).

Consider how this idea works for the Philadelphia question. The answer $k =$"False" is less common than $j =$"True", so the first ratio in (4) is less than one, and the majority answer (2) is incorrect. On the other hand, the second ratio is greater than one, because of the asymmetry in predictions in Figure 1c, and is actually strong enough to override majority opinion.

For a question with more than two possible answers, it is useful to convert the pairwise comparisons of (4) back into a maximum principle. Taking the logarithm of (4) and performing a weighted sum over j yields

$$\begin{aligned} \log \Pr[\Omega = i^* | X^r = k] \quad &= \quad \log \Pr[X^r = k | \Omega = i^*] + \sum_j w_j \log \frac{\Pr[X^s = j | X^r = k]}{\Pr[X^r = k | X^s = j]} \\ &\quad + \sum_j w_j \log \frac{\Pr[\Omega = i^* | X^s = j]}{\Pr[X^s = j | \Omega = i^*]} \end{aligned} \tag{5}$$

This follows from (4) for any set of weights w_j satisfying $\sum_j w_j = 1$. Since the last term does not depend on k, the best answer is

$$\operatorname*{argmax}_k \left\{ \log \Pr[X^r = k | \Omega = i^*] + \sum_j w_j \log \frac{\Pr[X^s = j | X^r = k]}{\Pr[X^r = k | X^s = j]} \right\} \tag{6}$$

For a practical algorithm that identifies the best answer, we estimate the probabilities in (6) from a population of respondents. Let $x^r_k \in \{0, 1\}$ indicate whether respondent r endorsed answer k ($x^r_k = 1$) , and y^r_k her prediction of the fraction of respondents endorsing answer k. If we estimate $\Pr[X^r = k | \Omega = i^*]$ using the arithmetic mean, $\bar{x}_k = n^{-1} \sum_r x^r_k$, then Eq. (6) takes the form

$$\operatorname*{argmax}_k \left\{ \log \bar{x}_k + \sum_j w_j \log \frac{\bar{y}_{jk}}{\bar{y}_{kj}} \right\} \tag{7}$$

where the estimate $\bar{y}_{kj}$ of $\Pr[X^r = k | X^s = j]$ is based on the predictions y^s_k of respondents s who endorsed answer j. This could be the arithmetic mean $\bar{y}_{kj} = (n\bar{x}_j)^{-1} \sum_s x^s_j y^s_k$ or the geometric mean $\log \bar{y}_{kj} = (n\bar{x}_j)^{-1} \sum_s x^s_j \log y^s_k$. The choice of weights w_j only matters in the case of inconsistencies between the pairwise comparisons. To resolve inconsistencies, one could weight answers equally, $w_j = 1/m$, or, alternatively, weight respondents equally, $w_j = \bar{x}_j$. In the empirical results below, we compute geometric means of predictions and weight respondents equally, and refer to this as the LST algorithm.

To validate the algorithm, we conducted surveys of knowledge of all fifty US state capitals. Each question was like the Philadelphia one above, where the named city was always the most populous in the state. Respondents endorsed "True" or "False" and predicted the distribution of votes by other respondents. The test has some richness because problems range in difficulty, and individual states are challenging for a variety of different reasons. The prominence of a city is sometimes misleading (Philadelphia-Pennsylvania), and sometimes a valid cue (Boston-Massachusetts), and many less populous states have no prominent city. Surveys were administered to three groups of respondents at MIT and Princeton. The True-False votes of the respondents were tallied for each question, and the majority decision was correct on only 31 states in Study 1 ($n = 51$), 38 states in Study 2 ($n = 32$), and 31 states in Study 3 ($n = 33$) (ties counted as 0.5 correct). The LST answers were consistently more accurate, reducing the number of errors from 19 to 9 in Study 1 (matched pair $t_{49} = 2.45$, $p < .01$), from 12 to 6 in Study 2 ($t_{49} = 1.69$, $p < .05$), and from 19 to 4 in Study 3 ($t_{49} = 4.40$, $p < .001$). Our basic empirical finding, that LST outperforms democratic voting, is thus replicated by three separate studies.

In order to compare LST with confidence-weighted voting, Study 3 went beyond the first two studies and asked respondents to report their confidence with a number from 0.5 to 1, as described earlier and in Figure 1. Weighting answers by confidence is indeed more accurate than majority opinion, reducing the number of errors from 19 to 13 ($t_{49} = 2.86$, $p < .01$), but is still less accurate than LST ($t_{49} = 2.64$, $p < .02$). More extreme forms of confidence weighting, such as a policy of only counting the answers of individuals that claim to know the answer for sure (100% confident), or selecting the answer whose adherents are most confident, are likewise not as accurate as LST (Table 1).

For complex, substantive questions, we may prefer a probabilistic answer as a quantitative summary of all available evidence. An estimate of the probability that a city is the capital can be imputed to each respondent based on the True/False answers and confidence estimates collected in Study 3 (Figure 1b). The LST algorithm requires that these probability estimates be discretized and then treated as answers to a multiple choice question. The discretization was done by dividing the [0,1] interval into uniform bins or into nonuniform bins using a scalar quantization algorithm. In Study 3, each respondent was asked to predict the average of others' confidence estimates. This prediction, along with the prediction of the distribution of True/False votes, was used to impute a prediction of the entire distribution of probability estimates. The algorithm selected a bin, and its midpoint served as the best probability according to the LST definition. We compared this with averaging respondents' probabilities.

We found (Table 1) that LST probabilities were more accurate than average probabilities. This was not surprising for questions like Philadelphia-Pennsylvania, for which majority opinion was incorrect. More interestingly, LST outperformed probability averaging even on majority-solvable problems, defined as those for which majority opinion was correct in both Studies 1 and 2. For example, the Jackson-Mississippi question was majority-solvable, but most respondents found it difficult, as judged by their low average confidence. LST not only answered this question correctly, but also more confidently. Other algorithms that put more weight on confidence, such as the logarithmic pool [13] or retaining only the most confident answers, also outperformed probability averaging on majority-solvable problems, but not on majority-unsolvable problems like Philadelphia-Pennsylvania.

In (7), we proposed estimating $\Pr[X^s = k | X^r = j]$ by averaging the predictions of those respondents who answered j. In doing so, we regarded the fluctuations in the predictions between respondents giving the same answer as noise. Alternatively, fluctuations can be regarded as a source of additional information about individual expertise. The LST algorithm (the version of Eq. 7 with $w_j = \bar{x}_j$ and geometric means of predictions) can be rewritten as

$$\operatorname*{argmax}_k \left\{ \frac{1}{n\bar{x}_k} \sum_r x_k^r u^r \right\} \tag{8}$$

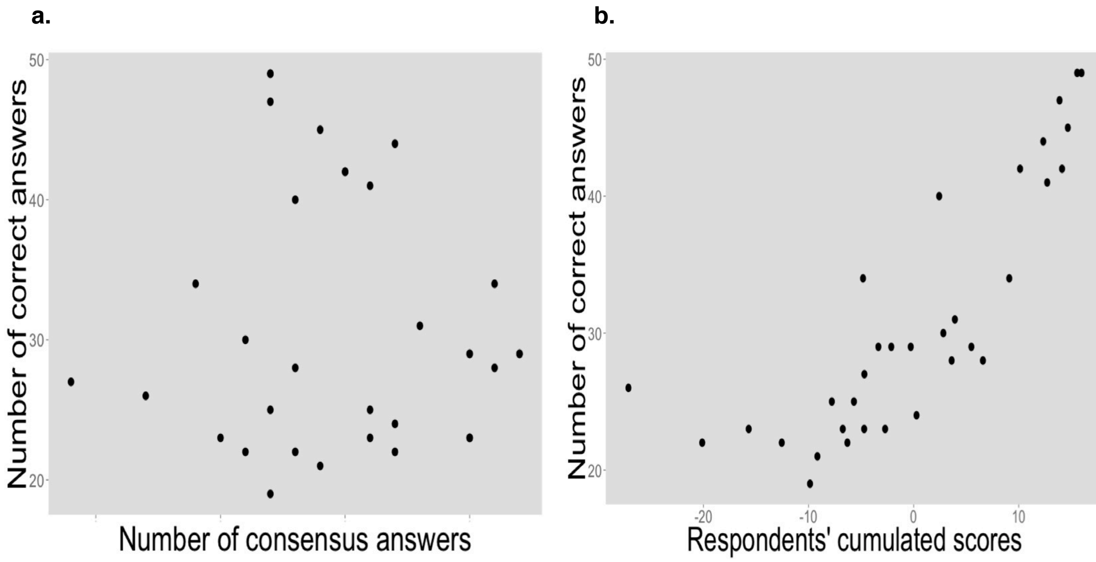

Figure 2: Scoring the expertise of individual respondents. **a** The accuracy of a respondent across all 50 states is uncorrelated with her conformity to conventional wisdom, defined as the number of times he votes with the majority. **b** Accuracy is highly correlated with the individual score u^r of (9) cumulated across fifty states.

where we define a score for each respondent r as

$$u^r = \sum_s \sum_{k,j} x_k^r x_j^s \log \frac{\bar{x}_k y_j^r}{\bar{x}_j y_k^s} = \sum_k x_k^r \log \frac{\bar{x}_k}{\bar{y}_k} - \sum_j \bar{x}_j \log \frac{\bar{x}_j}{y_j^r} \tag{9}$$

and $\bar{y}_j$ is the geometric mean of predicted frequencies of answer j, $\log \bar{y}_j = n^{-1} \sum_r \log y_j^r$. For respondents who give the same answer, the first term of the score is the same, but the second term (a relative entropy) is higher for those who are better able to predict the actual distribution of answers. If the score is an accurate measure of individual expertise, the best answer might be that of the single respondent with the highest score, rather than the LST algorithm, which selects the answer endorsed by the respondents with the highest average score as in (8). We found that the accuracy of the top scoring person was comparable to the LST algorithm (Table 1).

Respondents' individual scores across multiple questions provide an alternative measure of expertise. Figure 2, right panel, shows that the score of an individual respondent, averaged across all 50 states, is highly correlated with his or her objective accuracy (Study 3: $r = 0.82$, $p < .001$). For comparison, we also computed a conventional wisdom (CW) index, defined as the number of states for which a respondent votes with the majority for that state. Because the majority is correct more than half the time, one might expect that respondents with high CW scores will also get more answers correct. However, accuracy and CW are uncorrelated, as shown by the left panel of Figure 2. The score also outperformed several other approaches, such as principal components analysis (SupplementaryInformation).

While these results provide a critical initial test, we are ultimately interested in applying the algorithm to substantive problems, such as assessments of risk, political and economic forecasts, or expert evaluations of competing proposals. Because a verdict in these settings has implications for policy, and truth is difficult or impossible to verify, it is important to guard against manipulation by respondents who may have their own interests at stake. The possibility of manipulation has not been considered in this paper, as we have assumed that respondents gave honest and careful answers. We note, however, that the expertise score of Eq. (9) is identical to the payoff of a game that incentivizes respondents to give truthful answers to questions, even if those answers are nonverifiable. In the context of its application to truthfulness, the score (9) was called the Bayesian Truth Serum or BTS score [5, 20].

This scoring system has features in common with prediction markets, which are gaining in popularity as instruments of crowd-sourced forecasting [21]. Like market returns, the scores in Eq. (9) sum to zero, thus promoting a meritocratic outcome by an open democratic contest. Furthermore, in both cases, success requires distinguishing one's own information from information that is widely shared. With markets, this challenge is implicit — by purchasing a security, a person is betting that some relevant information is not adequately captured by the current price. Our approach makes the distinction explicit, by requesting a personal opinion and a prediction about the crowd. At the same time, we remove a limitation of prediction markets, which is the required existence of a verifiable event. This, together with the relatively simple input requirements, greatly expands the nature and number of questions that can be answered in a short session. Therefore, in combination with the result on incentives [5], the present work points to an integrated, practical solution to the problems of encouraging honesty and identifying truth.

Acknowledgments

Supported by NSF SES-0519141, Institute for Advanced Study (Prelec), and Intelligence Advanced Research Projects Activity (IARPA) via the Department of Interior National Business Center contract number D11PC20058. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions expressed herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

References

[1] Galton, F. Vox populi. *Nature* **75**, 450–451 (1907).

[2] Sunstein, C. *Infotopia: How many minds produce knowledge* (Oxford University Press, USA, 2006).

[3] Surowiecki, J. *The wisdom of crowds* (Anchor, 2005).

[4] Koriat, A. When are two heads better than one and why? *Science* **336**, 360–362 (2012).

[5] Prelec, D. A bayesian truth serum for subjective data. *Science* **306**, 462–6 (2004).

[6] Batchelder, W. & Romney, A. Test theory without an answer key. *Psychometrika* **53**, 71–92 (1988).

[7] Uebersax, J. Statistical modeling of expert ratings on medical treatment appropriateness. *Journal of the American Statistical Association* **88**, 421–427 (1993).

[8] Freund, Y. & Schapire, R. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **55**, 119–139 (1997).

[9] Chen, K., Fine, L. & Huberman, B. Eliminating public knowledge biases in information-aggregation mechanisms. *Management Science* **50**, 983–994 (2004).

[10] Morris, P. Combining expert judgments: A bayesian approach. *Management Science* **23**, 679–693 (1977).

[11] Winkler, R. The consensus of subjective probability distributions. *Management Science* **15**, B–61 (1968).

[12] Yi, S., Steyvers, M., Lee, M. & Dry, M. The wisdom of the crowd in combinatorial problems. *Cognitive science* (2012).

[13] Cooke, R. *Experts in uncertainty: opinion and subjective probability in science* (Oxford University Press, USA, 1991).

[14] Lorenz, J., Rauhut, H., Schweitzer, F. & Helbing, D. How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences* **108**, 9020–9025 (2011).

[15] Austen-Smith, D. & Banks, J. Information aggregation, rationality, and the condorcet jury theorem. *American Political Science Review* 34–45 (1996).

[16] DeGroot, M. Reaching a consensus. *Journal of the American Statistical Association* **69**, 118–121 (1974).

[17] Grofman, B., Owen, G. & Feld, S. Thirteen theorems in search of the truth. *Theory and Decision* **15**, 261–278 (1983).

[18] Hastie, R. & Kameda, T. The robust beauty of majority rules in group decisions. *Psychological review* **112**, 494 (2005).

[19] Ladha, K. The condorcet jury theorem, free speech, and correlated votes. *American Journal of Political Science* 617–634 (1992).

[20] John, L., Loewenstein, G. & Prelec, D. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science* **23**, 524–532 (2012).

[21] Wolfers, j. & Zitzewitz, E. Prediction markets. *Journal of Economic Perspectives* **18**, 107–126 (2004).

	Average probability assigned to the correct answer			Error measure		
Aggregation method	All 50 States	30 majority-solvable states	20 majority-unsolvable states	Number of incorrect answers out of 50	Quadratic error (Brier score)	Log score
Linear pool	61.4	70.1	48.3	13	0.17	-0.52
0/1 Majority vote	N/A	N/A	N/A	19*	N/A	N/A
Logarithmic pool	67.1***	79.8***	48.1	14	0.15*	-0.46**
Counting only 100% confident	70.7***	83.8***	50.9	11.5	0.15	-0.43*
LST algorithm, T/F answers only	N/A	N/A	N/A	4*	N/A	N/A
Top scorer by u^r in each state	81.5***	85.4*	75.6***	4*	0.08**	-0.36*
Average of top 3 scorers by u^r in each state	81.5***	86.7***	73.7***	2.5**	0.06***	-0.24***
Top scorer by u^r across all 50 states	98.8***	100.0***	97.0***	0.5***	0.01***	-0.02***
Probabilistic LST with 2 equal bins	70.8*	79.0	58.6	12	0.15	-0.49
Probabilistic LST with 3 equal bins	85.3***	85.6*	84.9***	3*	0.07**	-0.29*
Probabilistic LST with 5 equal bins	81.8***	78.7	86.6***	7	0.14	-0.59
Probabilistic LST with 2 scalar-quantized bins	81.7***	85.2*	76.6***	4*	0.09*	-0.34*
Probabilistic LST with 3 scalar-quantized bins	84.9***	88.5***	79.5***	5*	0.10	-0.38
Probabilistic LST with 5 scalar-quantized bins	92.7***	95.6***	88.4***	3*	0.06**	-0.31**

Table 1: **Performance of LST compared to baseline aggregation methods**. The table shows the performance of different aggregation methods on data collected in Study 3. Results are shown for baseline aggregation methods (linear and log pools), different implementations of the LST algorithm, and individual respondent BTS score (9). This includes LST applied to the binary True/False answer, and then averaging the probability of the identified experts either per question or across questions. They also include probabilistic LST with equal-sized or scalar-quantized bins. LST algorithms outperform the baseline methods with respect to Brier scores [13], log scores, and the probability they assign to the correct answer. The performance of the algorithms is shown separately on majority-solvable and unsolvable states where solvable states are defined as those for which the majority decision was correct in both Studies 1 and 2. By this definition there were 30 easy and 20 hard states. Significance assessed against the Linear pool, two-tailed matched-pair (t49), *=<.05, **=<.01, ***=<.001.