

An algorithm that finds truth even if most people are wrong*

Drazen Prelec

Massachusetts Institute of Technology

Sloan School, Cambridge MA 02139

dprelec@mit.edu

H. Sebastian Seung

Howard Hughes Medical Institute

and

Massachusetts Institute of Technology, Brain and Cognitive Sciences

Cambridge MA 02139

seung@mit.edu

Abstract

Effective use of subjective judgment is essential in all fields of knowledge. We present a method for finding truth when the subjective judgments of multiple respondents are the only evidence available, and majority opinion may be wrong. Respondents are scored for their own judgments and for their metaknowledge of others' judgments. In a probabilistic model of belief formation, their scores converge to the subjective probabilities they implicitly assign to the truth. Hence, the judgments of the highest scoring respondents reveal truth, irrespective of majority opinion. An experiment demonstrates that the method outperforms majority rule in selecting truth, and in identifying experts.

* Acknowledgments: Support by NSF SES-0519141, Institute for Advanced Study and John Simon Guggenheim Foundation (Prelec). Hospitality of Aspen Center for Physics, Zagreb Math Dept, Tom Palfrey and Princeton Lab for Exp Soc Sci. Comments and discussion: Dilip Abreu, Shane Frederick, Michael Kearns, Eric Maskin, Danica Mijovic-Prelec, Whitman Richards, Rebecca Saxe, Ray Weaver.

Subjective judgment is an irreducible component of pure and applied knowledge (1). Experts estimate the likelihood of rare accidents, assess the evolutionary significance of bone fragments, pronounce on aesthetic value and give interpretation on points of law. Even the touchstone of scientific practice — peer review of articles and proposals — rests on referees' subjective assessments of originality and significance. But experts often disagree, and can be wrong (2, 3). The lack of accepted procedures for reconciling opinions feeds skepticism about subjective judgment, and promotes controversy when expertise is claimed but difficult to prove (4).

New Web-based technologies have made it easier to tap the knowledge of dispersed individuals, raising hopes that automated opinion aggregation mechanisms might provide an impartial, efficient alternative to expert judgment (2, 5). Indeed, electronic information (prediction) markets have proved remarkably accurate in some domains (6, 7). However, markets suffer from a fundamental limitation: They can only trade claims whose final value is defined by a public event. For example, one can organize a market to predict the outcome of a jury trial, which is of course a public event, but not to determine the defendant's true guilt or innocence (5).

We present a method for finding truth when subjective judgment remains the only source of evidence and there is a possibility that most people are wrong. The method selects the judgments of respondents who reveal superior metaknowledge — knowledge of others' opinions. The explicit metaknowledge criterion distinguishes our method from other approaches, such as voting algorithms (8-13), multi-item analysis (14, 15), or Bayesian updating (1, 16-18).

The instrument for measuring metaknowledge is a mathematical formula called the Bayesian Truth Serum (BTS) (19). As input to the formula, respondents provide their judgments and also their predictions of the distribution of judgments in a sample of peers. These dual reports then undergo competitive, zero-sum scoring as a function of the sample statistics. Previously, it was shown that BTS scoring provides incentives for respondents to answer truthfully (19), but the problem of identifying judgments that are closer to objective, impersonal truth was left open. Here, the same formula is shown to select judgments that are objectively true, under fairly mild assumptions about how truth and belief are related. Hence, when these assumptions hold, the BTS scoring system provides an integrated solution to both problems — rewarding truthfulness, and aggregating judgments to discover truth (20).

Metaknowledge is an effective truth diagnostic whenever information is unevenly distributed among respondents, so that some pieces are widely shared but other pieces are known only by a minority (16, 21). As an elementary example, consider the (false) proposition that Chicago is the capital of the state of Illinois. Respondents might form different opinions about the truth of the proposition, depending on whether they knew: (a) that Chicago is a large city, (b) that it is located in the state of Illinois, (c) that Springfield is the actual capital of Illinois, and so on. If the typical person is aware of

(a) and (b) but not of (c), then the majority of those queried might vote for the incorrect answer, that the proposition is True.

A democratic poll ignores the asymmetry in metaknowledge between respondents who know the right answer and those who do not. Those who know that Chicago is not the capital of Illinois can imagine that many others will be misled. A comparable insight into the opinions of others is not available to those who falsely believe the answer is Yes (22). Our scoring method in effect reweights the votes so as to reflect different levels of metaknowledge associated with each possible answer. If the method works as claimed, the true answer should emerge as the winner, regardless of how many respondents endorse it. Later on, this claim will be justified within a Bayesian model of belief formation, using the mathematical theorem about the BTS formula given in Box 1. But first we give a description of the method itself.

The scoring works at the level of a single question with m answers, which is posed to n respondents, indexed by $r, s, \dots \in \{1, 2, \dots, n\}$. The question should admit only one right answer, and it should be reasonable to believe that if one could somehow pool information from all respondents, then that collective knowledge would determine the right answer. For example,

- (a) Chicago is the state capital of Illinois ($m=2$: True, False)
- (b) The best current estimate of the global temperature increase between now and 2100 is _____
($m=5$: $< 2^\circ\text{C}$, $2^\circ - 4^\circ\text{C}$, $4^\circ - 6^\circ\text{C}$, $6^\circ - 8^\circ\text{C}$, $> 8^\circ\text{C}$)
- (c) On current evidence, the probability of intelligent extraterrestrial life in the universe is _____
($m=10$: numerical probability scale, divided into 10 categories)

The respondent is asked to endorse the answer mostly likely to be true, and to predict the proportion of the sample that will endorse each possible answer. Let $x_k^r \in \{0, 1\}$ indicate whether respondent r has endorsed answer k , and $y = (y_1^r, \dots, y_m^r)$ her prediction of the sample proportions ($y_k^r \geq 0, \sum_k y_k^r = 1$). The truth-selection algorithm proceeds in four steps.

Step 1 Calculate the average \bar{x}_k of the endorsements and the geometric mean \bar{y}_k of the predictions:

$$\bar{x}_k = \frac{1}{n} \sum_{r=1}^n x_k^r, \quad \log \bar{y}_k = \frac{1}{n} \sum_{r=1}^n \log y_k^r \quad (1)$$

Step 2 Calculate the BTS score (23) of each individual r :

$$u^r = \sum_{k=1}^m x_k^r \log \frac{\bar{x}_k}{\bar{y}_k} + \sum_{k=1}^m \bar{x}_k \log \frac{y_k^r}{\bar{x}_k} \quad (2)$$

Step 3 For each answer k , calculate the average BTS score \bar{u}_k of all individuals endorsing answer k :

$$\bar{u}_k = \frac{1}{n} \sum_{r=1}^n x_k^r u^r \quad (3)$$

Step 4 Select the answer k that maximizes \bar{u}_k .

This algorithm will be validated both theoretically and experimentally, but first a few words about our methodology are in order. The theoretical justification is based on a model of belief formation presented below. While this model is intuitively reasonable, its quality as an approximation to real-world situations cannot be decided by purely theoretical means. Therefore we will also provide a simple empirical test. In this application the true answers are known independently, so that the accuracy of the method can be quantified.

The model formalizes the relation between truth and subjective beliefs of an ideal Bayesian respondent. The truth $\square = i$ is associated with one of the m answers, and is drawn from a probability distribution $\Pr[\square = k]$. Each respondent r then receives a signal T^r , representing information relevant to the question. Respondents are identical apart from the value of their signal, and signals of different respondents are presumed independent when conditioned on \square . Therefore, the signal generation process is fully specified by a *signal matrix* $S_{kj} = \Pr[T^r = k | \square = j]$.

The complete probability distribution over all variables, also called the *common prior*, is assumed to be common knowledge among respondents (24). In that case, Bayesian respondents can compute a *belief matrix* $B_{jk} = \Pr[\square = j | T^r = k]$, which expresses subjective uncertainty about truth, and a *metaknowledge matrix*, $M_{jk} = \Pr[T^s = j | T^r = k]$, which expresses uncertainty about signals received by others. We simplify the belief matrix by assuming that exactly one signal favors each possible answer. Then there is no ambiguity in taking $T^r = k$ to mean that respondent r believes that $\square = k$ is most likely to be true,

$$\Pr[\square = k | T^r = k] > \Pr[\square = j | T^r = k], \text{ for all } j \neq k.$$

For dichotomous questions this completes the model. For $m \geq 3$, one further assumption on the structure of the belief matrix is necessary. This is *truth sensitivity*,

$$\Pr[\square = i | T^r = i] > \Pr[\square = i | T^r = k] \quad (4)$$

for any k different from the truth i . In other words, if we compare the degree of beliefs in the true answer $\square = i$, those who received the signal $T^r = i$ believe more strongly in the true answer than those who received other signals. The reasonableness of this assumption may be more evident when rewritten using Bayes' rule, as

$$\frac{\Pr[T^r = i | \square = i]}{\Pr[T^r = k | \square = i]} > \frac{\Pr[T^r = i]}{\Pr[T^r = k]}$$

which asserts that conditioning upon truth strengthens belief in truth relative to belief in untruth (25).

Above we stated that respondents' signals are independent when conditioned on truth. From this assumption one might jump to the mistaken conclusion that the predictions of all respondents should be the same, regardless of the answers they endorse. In fact, the signals of the respondents have statistical dependencies induced by uncertainty about θ . Calculation of the metaknowledge matrix by Bayes' rule therefore requires averaging over all possible values of θ , giving rise to differences in metaknowledge and causing predictions of respondents to depend on which answer they endorse. This dependence is consistent with Bayesian reasoning and is also well-corroborated by experiments (26).

Observe that the common prior contains information about counterfactual scenarios. For example, what fraction of people would believe that Chicago is the capital of Illinois, if Chicago were actually the capital of Illinois? What is the prior probability that Chicago is the capital of Illinois? Introspectively, these seem to be difficult and speculative questions, and our method does not require respondents to answer them. Instead, we only request that respondents pick the answer they believe most likely to be true, and predict the distribution of endorsements in the sample.

These paired reports are in turn directly related to the elements of the belief model. The endorsement of a respondent is identified with his signal. The predictions of a respondent are interpreted as a noisy report of the single column of the metaknowledge matrix associated with his signal. Therefore, if respondents are truthful, their predictions and their distribution of endorsements will estimate, respectively, the full metaknowledge matrix and a *single* column of the signal matrix, the column corresponding to the true answer.

Nothing in the belief model rules out the possibility that the largest entry in this column is associated with a wrong answer, which is to say, that a wrong answer will receive the most votes. The model is noncommittal about the circumstances that might give rise to such widespread false beliefs, but it is worth outlining an example for concreteness (Box 2). Suppose that most respondents only recognize Chicago as the largest city in Illinois. They are likely to conclude that Chicago is the state capital, whether or not that statement is true. A smaller number of respondents have more specific information, because they remember hearing that another city is capital. This memory is not 100% reliable, but is sufficient to tip judgment against the proposition. These differences in knowledge could give rise to the numbers shown in Box 2.

To sharpen the truth-detection problem, the numbers have been fine-tuned to create a situation where respondents with the wrong signal should be more confident. Those endorsing Yes should infer from their signal (by applying Bayes' rule) that there is an 83% probability that they are correct, while those endorsing No should infer that there is only a 73% probability that they are correct. Hence, if opinions are weighted by confidence the collective decision would still be wrong, as would any decision reached by mathematical aggregation of subjective probabilities.

In the actual world Chicago is not the capital of Illinois. However, this model explains how most people might think that it is, even though they understand the informational mechanism and follow the rules of Bayesian inference. How does BTS discern the true answer in this situation? The user of BTS asks respondents to predict the beliefs of others. If respondents compute the metaknowledge matrix, those with the incorrect opinion will predict that 89% will agree with them, while those with the correct opinion will predict that only 30% will agree, that is, they expect to be in the minority. The user would record these two sets of predictions, and would also record that about 60% of the sample endorses the proposition that Chicago is the capital. Applying (3) to these statistics, she would arrive at a score of -0.57 for Yes and $+0.86$ for No, and conclude that majority opinion is wrong.

To prove that this will work in the general case, we must specify how actual reports relate to the probabilistic model in the large sample limit. An ideal respondent, facing the scoring system, would carry out the Bayesian computations required to find the metaknowledge matrix. In reality, errors in these computations and noise in the reports are likely, but they are assumed to average out. Namely, the average log prediction of all respondents endorsing answer j converges to the j th column of the log metaknowledge matrix, $\log M_{kj}$. Second, the average endorsement frequencies converge to the column of the signal matrix associated with the true answer.

With these inputs, the BTS formula reveals, for each possible signal, the probability that that signal assigns to the true answer (modulo a constant). This is proven in the Theorem of Box 1, which states that the average score of respondents endorsing answer k is

$$\lim_{n \rightarrow \infty} \bar{u}_k = \log \Pr[\square = i \mid T^r = k] + C, \quad (5)$$

where $\square = i$ is the true answer, and C does not depend on k . The theorem proves that the BTS user can rank respondents according to how well they anticipate the truth. Indeed, the theorem goes further, proving that the asymptotic BTS score is identical to the score that respondents would have received with the logarithmic proper scoring rule (27), as administered by a user with independent access to truth.

Equation 5 is the central mathematical result of this paper. When combined with truth sensitivity it immediately implies that maximizing the average BTS score \bar{u}_k with respect to k is a prescription for finding truth, in the limit of a large number of respondents. Therefore the truth selection property of the BTS method is proven, at least for the particular model of belief formation considered here (28).

Our theorem predicts that the BTS decision rule should outperform majority rule in recovering true answers. We test this with a survey of knowledge of US state capitals, presented to students at MIT ($n=51$) and Princeton ($n=32$) (details in SOM). The survey contained fifty questions, of the form: “Is Chicago the capital of Illinois,” where the named city was always the most populous in the state.

Students answered Yes or No, and predicted the distribution of Yes and No votes in the sample. Although elementary, these questions should give rise to differences in knowledge that the method can exploit. Given natural variation in backgrounds, it is likely that for most states some subset — the *local experts* — would know the correct answer. At the same time, their number might be small and their influence diluted by the uninformed majority. The survey tests our main theoretical claim — that BTS is able to reduce such dilution of expert votes.

The survey proved difficult — the average respondent was slightly above chance, endorsing 29.5 correct answers at MIT and 31 at Princeton. The collective judgment was slightly better: The majority decision was correct for 31 states at MIT, and for 36 states (and 4 ties) at Princeton. As implied by the Bayesian model, endorsements had a strong impact on predictions. Respondents who agreed that a city was the capital predicted on average that 70.3% of others would vote Yes, while those who denied the proposition predicted only 49.8% for Yes (29). Predictions made by respondents with correct answers were on average more accurate, closer to actual percentages in 37 states at MIT, 46 states at Princeton.

Can the BTS algorithm determine the true answers, as implied by theory? Following Steps 1-4, we find that in the MIT sample, BTS decision differs from the majority decision on 18 states: for 14 states (including Illinois) it corrects a wrong majority decision, while for 4 states it reverses a correct majority decision. The 14-4 split is significantly different from chance ($p < .02$, binomial test). The total number of mistakes drops from 19 to 9 (matched pair $t_{49} = 2.45$, $p < .01$). In the Princeton sample, the split is 12-4 in favor of BTS ($p < .05$, binomial test). If ties are coded as 0.5, the total number of mistakes drops from 12 to 6 (matched pair $t_{49} = 1.69$, $p < .05$).

BTS does, therefore, improve on majority decision when applied to individual questions. However, BTS scores can also be used identify expert respondents if knowledge correlates across multiple questions. BTS suggests two possible indices for rating the expertise of a single respondent r . The first, *individual index* is the BTS score u^r of the respondent, averaged across all questions. The second, *pooled index* is the average BTS score of the answer endorsed by respondent r , $\bar{u}_r = \frac{1}{m} \sum_{k=1}^m x_r^k u_k$, averaged over all questions. The two indices coincide in theory, because respondents endorsing the same answer should have identical BTS scores. An advantage of the pooled index is that it filters out individual differences in prediction competence (i.e., longrun calibration of predictions).

Figure 2 shows how the pooled BTS index of expertise correlates with actual knowledge of states' capitals. As benchmark, we included an index of *conventional wisdom* (CW), defined as the number of states for which a respondent votes with the majority opinion for that state. Because majority opinion is correct more than half the time, one might expect that respondents with high CW scores will also get more answers right. However, accuracy and CW are uncorrelated, as shown by the two left

panels of Figure 2. CW is worthless as a predictor of genuine knowledge for this domain (30). By contrast, the panels on the right show that the BTS score is able to sort respondents by true accuracy (MIT: $r=+.84$, $p<.0001$; Princeton: $r=+.94$, $p<.0001$). The pooled BTS index is thus an excellent predictor of genuine knowledge.

The effectiveness of the BTS expertise index suggests hybrid decision policies. For example, one could follow the majority opinion of expert subsets, where expertise is determined by BTS scores across all fifty questions. Judgments of other respondents would be ignored in this elitist version of majority rule. Figure 2 shows the performance of such policies for the two expertise indices displayed in Figure 1, as a function of subset size. Each line starts with a subset of one — the accuracy of the single respondent who scored highest on the corresponding criterion. For any size of expert subset (indicated by the numerical value on the x-axis) the y-axis gives the number of correct answers produced by majority rule applied to that subset.

The blue lines in the Figure show that there is no benefit in eliminating respondents with unconventional judgments. However, if the sample is purified according to BTS scores (red lines), then small subsets achieve impressive accuracy levels. Retaining the votes of the top 20% of the sample results in 45 (MIT) or 47 (Princeton) correct majority opinions. The majority vote of the top three respondents by BTS score is a nearly perfect indicator of truth in either sample (47 at MIT, 48 at Princeton). These results highlight the latent superiority of expert subsets and establish the possibility of using purely internal criteria to select them. As the simple survey shows, it may be better to follow a handful of individuals with high metaknowledge than the majority opinion of the full sample.

In summary, we have presented here an algorithmic method for determining true answers, when private opinions are the only source of information. The method aims to extend the advantages of competitive information aggregation mechanisms to subjective domains of knowledge, such as aesthetics or law. Like an information market, the method requires minimal inputs from the user — only the question and possible answers. Although the supporting theory assumes a Bayesian model of belief formation, the user is not challenged to specify prior probabilities and in fact need not be a Bayesian at all.

The algorithm is naturally suited to open access implementation, which is often presented as a special virtue of markets (2, 6, 7). Since truth is not associated with a statistical average but with the opinions of winners, there is no need to screen for expertise. The less informed individuals will not disturb the outcome, but will only subsidize those who are more informed. Over time, self-selection and attrition should lead to a situation where experts predominate in the active player pool. In this sense, the BTS algorithm enforces a meritocratic outcome by an open democratic process.

Notes

1. R. M. Cooke, *Experts in Uncertainty* (Oxford University Press, New York, 1991).
2. J. Surowiecki, *The Wisdom of Crowds* (Random House, New York, 2004).
3. P. E. Tetlock, *Expert Political Judgment: How good is it? How can we know?* (Princeton University Press, Princeton, 2006).
4. For example, from the back cover of the recent best-seller (2): “Large groups of people are smarter than an elite few, no matter how brilliant — better at solving problems, fostering innovation, coming to wise decisions, even predicting the future.”
5. C. R. Sunstein, *Infotopia* (Oxford University Press, New York, 2006).
6. D. M. Pennock, C. L. Giles, *Science* **291**, 987 (2001).
7. J. Wolfers, E. Zitzewitz, *J Econ Perspect* **18**, 107 (2004).
8. D. Austen-Smith, J. S. Banks, *Am Pol Sci Rev* **90**, 34 (1995).
9. M. d. Condorcet, in *Condorcet: Selected Writings* K. M. Baker, Ed. (Bobbs-Merrill Co., Indianapolis, 1785/1976) pp. 33-70.
10. M. De Groot, *J Am Stat Assoc* **69** (1974).
11. B. Grofman, G. Owen, S. L. Feld, *Theory and Decision* **15** (1983).
12. R. Hastie, T. Kameda, *Psychol Rev* **112**, 494 (2005).
13. K. K. Ladha, *Am J Pol Sci* **36**, 617 (1992).
14. W. H. Batchelder, A. K. Romney, *Psychometrika* **53**, 71 (1988).
15. J. S. Uebersax, *J Am Stat Assoc* **88**, 421 (1993).
16. K.-Y. Chen, L. R. Fine, B. A. Huberman, *Manage Sci* **50**, 983 (2004).
17. P. Morris, *Manage Sci* **20**, 1233 (1977).
18. R. Winkler, *Manage Sci* **15**, B61 (1968).
19. D. Prelec, *Science* **306**, 462 (2004).
20. Interestingly, majoritarian decision rules need not be truthfulness-inducing even in the absence of any variation in preferences (8). Condorcet’s jury theorem may be the first statistical argument for majority rule as an algorithm for arriving at the truth of a proposition (9). Condorcet assumes independence of judgments conditional on truth, and individual competence — that each person has a greater than even chance of knowing the truth of a proposition. Majority opinion converges on the true answer as the number of polled individuals increases. This remains an influential paradigm, with many refinements (10-14). Our framework denies the individual competence assumption.
21. R. Winkler, *Manage Sci* **27**, 479 (1981).
22. S. Frederick, *J Econ Perspect* **19**, 24 (2005).
23. The BTS score has two components: the first evaluates the person’s answer, while the second evaluates the prediction. Answers score high to the extent that they are surprisingly common — more common than collectively expected. Predictions are scored for accuracy, measured by the Kullback-Leibler divergence. For smaller samples, the formula can be adjusted to exclude a person’s own answers from the statistics that determine his score (19). A respondent who believes that other respondents are truthful should behave truthfully provided a single answer has negligible impact on sample statistics (and his sole objective is to maximize expected score). Formally, truth-telling is a strict Bayesian Nash equilibrium in a large enough sample (19). The scoring formula is zero-sum. Zero-one predictions are trimmed, e.g., to .01 and .99.
24. J. Harsanyi, *Manage Sci* **14**, 159 (1967).
25. Truth sensitivity holds automatically for the dichotomous case $m=2$, as $Pr[\square = i | T^r = i] > 0.5$, $i=1, 2$, jointly imply: $Pr[\square = 1 | T^r = 1] > 0.5 > 1 - Pr[\square = 2 | T^r = 2] = Pr[\square = 1 | T^r = 2]$, and similarly for $\square=2$.
26. R. M. Dawes, *J Exp Soc Psychol* **25**, 1 (1989).
27. L. J. Savage, *J Am Stat Assoc* **66**, 783 (1971).

28. The superiority of this algorithm over majority decision should be detectable even with small samples. Observe that by Bayes' rule, the BTS score of equation 5 is the log ratio of $\Pr[T^r = k | \square = i]$ to $\Pr[T^r = k]$. The first quantity is the endorsement frequency in a large sample. This is normalized (weighted inversely) by the second quantity, the probability of endorsements unconditioned on the truth. The normalization is what makes it possible for a minority opinion to win. If small sample frequencies deviate substantially from $\Pr[T^r = k | \square = i]$, this will potentially affect not only the BTS decision but also the majority decision, which goes exclusively by these frequencies. The benefit of weighting the frequencies accrues to BTS alone.
29. Pooling across all questions and both samples. For example, with the Chicago question those who endorsed Yes believed that 80% would agree, while those endorsing No predicted a 50-50% split; the actual percentage was 59% for Yes. The true answer has a very small positive impact on predictions (+2.5%), after controlling for the respondent's answer and subject fixed effects. This is consistent with our assumption that respondents' information can be approximated by a binary signal favoring one or the other answer. State-by-state statistics are in the SOM.
30. Average *percent* agreement with other respondents also fails to predict accuracy (MIT: $r = +.09$; Princeton: $r = +.22$). Another index of interest is *skepticism*, defined as the total number of No answers. Skepticism does correlate with accuracy (+.66 & +.60). However, in a simultaneous regression on both variables, the partial correlation of accuracy and BTS remains very high (+.71 & +.91), while the partial correlation of accuracy and skepticism disappears (+.26 & -.02). This rules out the interpretation that the objective accuracy of individuals with high BTS scores derives from a nonspecific, but correct suspicion that the largest city is often not the capital. Additional indices, including principal component scores, are considered in the SOM, with similar results (Tables S5, S6).

Box 1 The BTS formula reveals the subjective probability of truth

Consider the model of belief formation described in the main text. Suppose that the average of n endorsements converges to the i -th column of the signal matrix:

$$\lim_{n \rightarrow \infty} \bar{x}_j = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{s=1}^n x_j^s = S_{ji} \quad (a)$$

and that the average of log predictions converges to the logarithm of the metaknowledge matrix:

$$\lim_{n \rightarrow \infty} \frac{1}{n \bar{x}_j} \sum_{s=1}^n x_j^s \log y_k^s = \log M_{kj} \quad (b)$$

Theorem. *It follows from (a) and (b) above that the averages of BTS scores for experts endorsing the same answer $j=1, \dots, m$,*

$$\bar{u}_j = \lim_{n \rightarrow \infty} \frac{1}{n \bar{x}_j} \sum_{s=1}^n x_j^s u^s, \quad (c)$$

converge to the log of the row of the belief matrix $B_{ij} = \Pr[\square = i \mid T^r = j]$ corresponding to the true answer $\square = i$, plus a constant C that is independent j ,

$$\lim_{n \rightarrow \infty} \bar{u}_j = \log \Pr[\square = i \mid T^r = j] + C, \quad (d)$$

Proof: Our task is to compute (c), the average of the BTS scores (see Eq. 2)

$$u^r = \prod_{k=1}^m x_k^r \log \frac{\bar{x}_k}{\bar{y}_k} \prod_{k=1}^m \bar{x}_k \log \frac{\bar{x}_k}{\bar{y}_k} \quad (e)$$

over all experts who endorsed answer j , and then take the limit $n \rightarrow \infty$. Averaging the first term selects the $k=j$ element from the summation, leaving $\log(\bar{x}_j / \bar{y}_j)$. In the limit, assumption (a) allows the

replacement of \bar{x}_j by S_{ji} . Furthermore, \bar{y}_j is replaced by $\lim_{n \rightarrow \infty} \log \bar{y}_j = \sum_{k=1}^m S_{ki} \log M_{jk}$, which follows from

assumption (b). Averaging the second term of (e) and taking the limit $n \rightarrow \infty$ replaces the predictions by the metaknowledge matrix, according to assumption (b). In this term also, the average endorsements

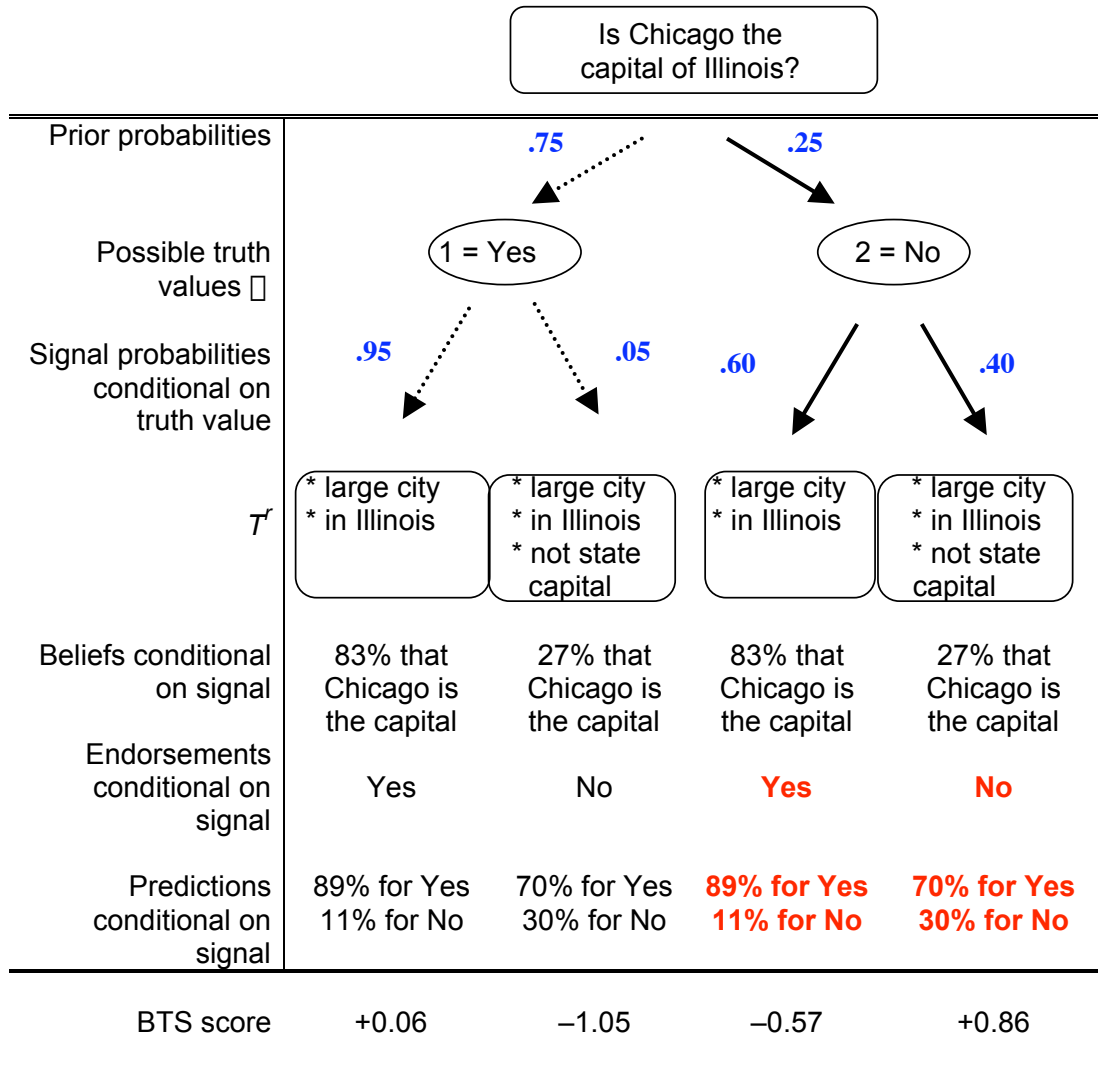
\bar{x}_k are replaced by S_{ki} . The final result is $\lim_{n \rightarrow \infty} \bar{u}_j = \sum_{k=1}^m S_{ki} \log \frac{S_{ji}}{M_{jk}} \prod_{k=1}^m S_{ki} \log \frac{S_{ki}}{M_{kj}}$. By Bayes' rule

$M_{jk} \prod_k = M_{kj} \prod_j$, yielding the desired average $\lim_{n \rightarrow \infty} \bar{u}_j = \log \frac{S_{ji}}{\prod_j} \prod_{k=1}^m S_{ki} \log \frac{S_{ki}}{\prod_k}$. The second term does not

depend on j . Application of Bayes' rule to the first term yields $S_{ji} / \prod_j \propto B_{ij}$. Equation (d) of the

Theorem follows.

Box 2 Illustration of the theorem in context of the Chicago – Illinois problem



The blue numbers are the unobserved common prior. The red numbers are what the BTS user will observe. The prior odds are 3:1 that a prominent city like Chicago is a state capital. Solid arrows represent the actual scenario, that Chicago is not the capital; dotted arrows, the counterfactual scenario. Respondents receive either a null signal (no further information), or a signal suggesting a different capital. On the counterfactual scenario, there is a small probability (5%) of getting the wrong signal. On the actual scenario, the probability of getting the correct signal is 40%. Predicted percentages are derived by Bayes' rule: The null signal implies a 83% probability that Chicago is the capital, leading to a $89\% = (.83)(.95) + (.17)(.60)$ prediction of fraction voting Yes; the other signal implies a 27% probability that Chicago is the capital, leading to a $70\% = (.27)(.95) + (.73)(.60)$ prediction for Yes. From these predictions and the 60-40 split in actual votes, the user calculates BTS scores of: $-.57$ for Yes and $+0.86$ for No, and concludes that majority decision is incorrect. Counterfactually, were Chicago the capital of Illinois, the observed votes would split 95-5 in favor, leading to BTS scores of $+0.06$ for Yes, and -1.05 for No, again in favor of the correct answer.

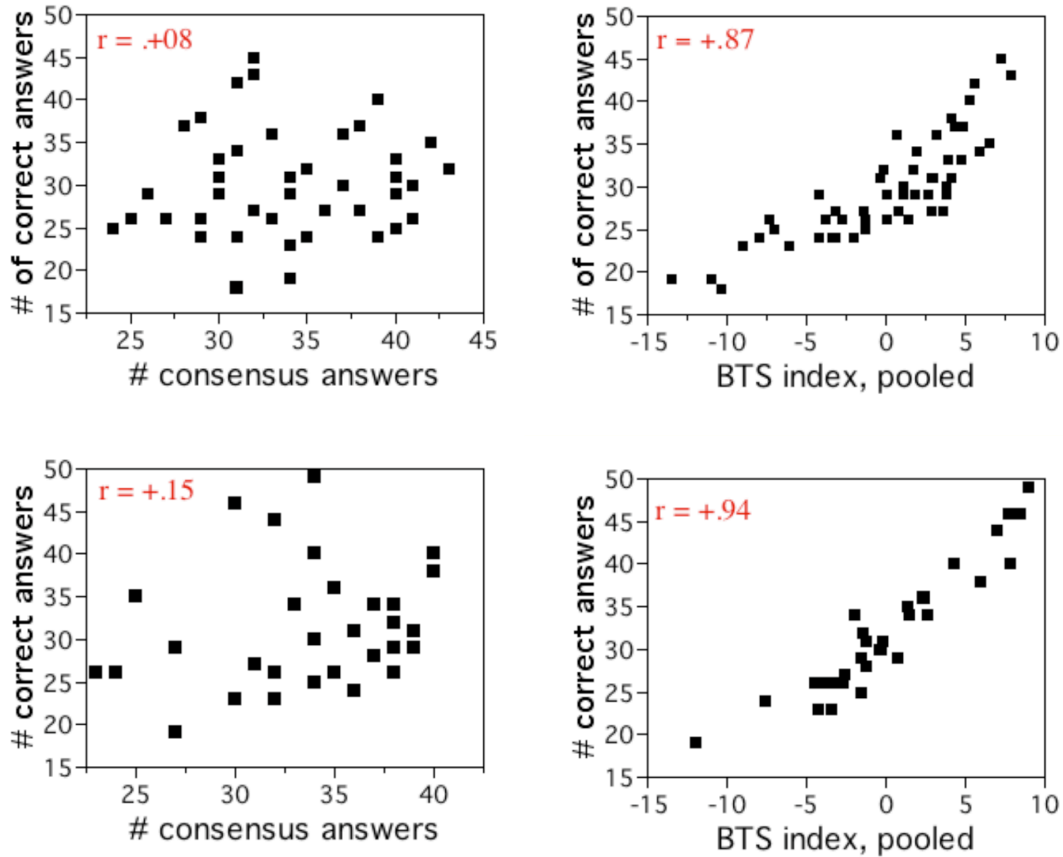


Figure 1 Expertise correlates with BTS scores but not with Conventional Wisdom.

Conventional wisdom is defined as the number of consensus answers (answers consistent with majority). The two top panels are taken from the MIT study; the bottom panels from the Princeton study. The y-axis is the number of correct answers out of fifty. The x-axis is: (left panels) the number of states where a respondent's answer matches majority opinion; (right) for each subject r , the sum over 50 states of the BTS score for the answer that they endorsed

(averaged across individuals who endorsed that answer), or: $\sum_{k=1}^m x_k^r \bar{u}_k$.

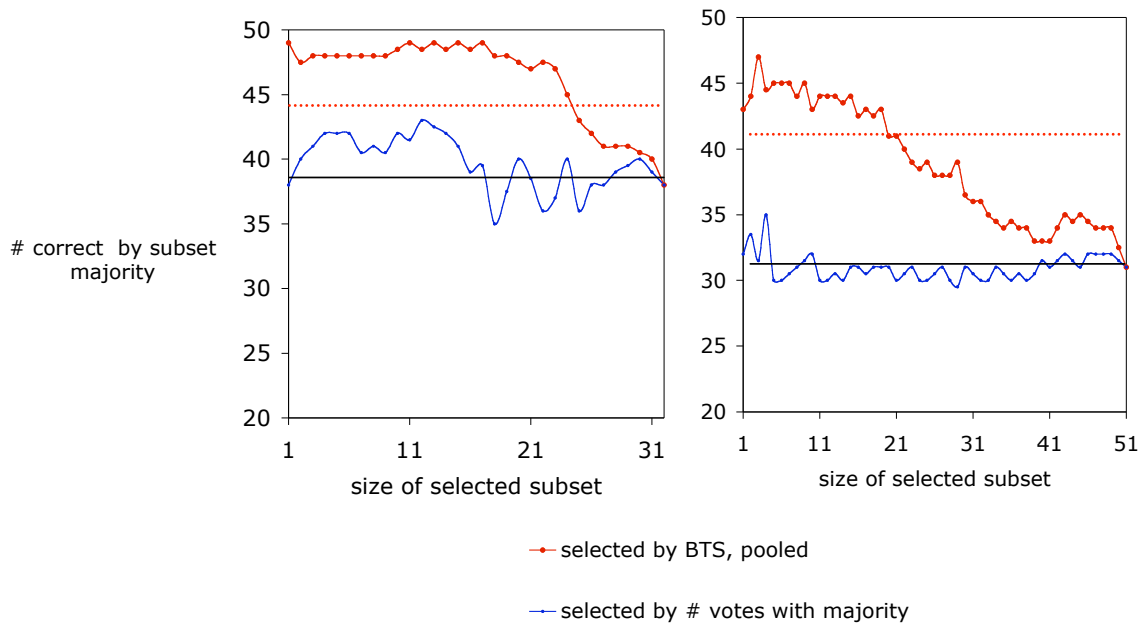


Figure 2 High accuracy of BTS selected expert subsets. Performance of expert subsets, selected by either BTS scores (red lines) or by Conventional Wisdom scores (blue lines). In each case the top ranked respondents by given criterion are retained in the subset. Right panel is MIT study, left panel the Princeton study. The y-axis is the number of correct answers for decisions reached by majority rule within a respondent subset (ties are credited 0.5). The x-axis is the number of respondents in the subset. The majority opinion of small BTS selected subsets is more accurate than majority decision (solid black horizontal line) and also more accurate than the BTS criterion applied to each question individually (dashed red horizontal line).