



## A Bayesian Truth Serum for Subjective Data

Drazen Prelec, *et al.*

*Science* **306**, 462 (2004);

DOI: 10.1126/science.1102081

**The following resources related to this article are available online at [www.sciencemag.org](http://www.sciencemag.org) (this information is current as of October 13, 2008 ):**

**Updated information and services**, including high-resolution figures, can be found in the online version of this article at:

<http://www.sciencemag.org/cgi/content/full/306/5695/462>

**Supporting Online Material** can be found at:

<http://www.sciencemag.org/cgi/content/full/306/5695/462/DC1>

This article has been **cited by** 4 article(s) on the ISI Web of Science.

This article has been **cited by** 2 articles hosted by HighWire Press; see:

<http://www.sciencemag.org/cgi/content/full/306/5695/462#otherarticles>

This article appears in the following **subject collections**:

Computers, Mathematics

[http://www.sciencemag.org/cgi/collection/comp\\_math](http://www.sciencemag.org/cgi/collection/comp_math)

Information about obtaining **reprints** of this article or about obtaining **permission to reproduce this article** in whole or in part can be found at:

<http://www.sciencemag.org/about/permissions.dtl>

## A Bayesian Truth Serum for Subjective Data

Dražen Prelec

Subjective judgments, an essential information source for science and policy, are problematic because there are no public criteria for assessing judgmental truthfulness. I present a scoring method for eliciting truthful subjective data in situations where objective truth is unknowable. The method assigns high scores not to the most common answers but to the answers that are more common than collectively predicted, with predictions drawn from the same population. This simple adjustment in the scoring criterion removes all bias in favor of consensus: Truthful answers maximize expected score even for respondents who believe that their answer represents a minority view.

Subjective judgment from expert and lay sources is woven into all human knowledge. Surveys of behaviors, attitudes, and intentions are a research staple in political science, psychology, sociology, and economics (1). Subjective expert judgment drives environmental risk analysis, business forecasts, historical inferences, and artistic and legal interpretations (2).

The value of subjective data is limited by its quality at the source—the thought process of an individual respondent or expert. Quality would plausibly be enhanced if respondents felt as if their answers were being evaluated by an omniscient scorer who knew the truth (3). This is the situation with tests of objective knowledge, where success is defined as agreement with the scorer's answer key, or in the case of forecasts, an observable outcome (4). Such evaluations are rarely appropriate in social science, because the scientist is reluctant to impose a particular definition of truth, even if one were available (5).

Here, I present a method of eliciting subjective information, designed for situations where objective truth is intrinsically or practically unknowable (6). The method consists of an “information-scoring” system that induces truthful answers from a sample of rational (i.e., Bayesian) expected value-maximizing respondents. Unlike other Bayesian elicitation mechanisms (7–9), the method does not assume that the researcher knows the probabilistic relationship between different responses. Hence, it can be applied to previously unasked questions, by a researcher who is a complete outsider for the domain. Unlike earlier approaches to “test theory without an answer key” (5), or the

Delphi method (10), it does not privilege the consensus answer. Hence, there is no reason for respondents to bias their answer toward the likely group mean. Truthful responding remains the correct strategy even for someone who is sure that their answer represents a minority view.

Instead of using consensus as a truth criterion, my method assigns high scores to answers that are more common than collectively predicted, with predictions drawn from the same population that generates the answers. Such responses are “surprisingly common,” and the associated numerical index is called an information score. This adjustment in the target criterion removes the bias inherent in consensus-based methods and levels the playing field between typical and unusual opinions.

The scoring works at the level of a single question. For example, we might ask: (i) What is your probability estimate that humanity will survive past the year 2100 (100-point probability scale)? (ii) Will you vote in the next presidential election (Definitely/Probably/Probably Not/Definitely Not)? (iii) Have you had more than 20 sexual partners over the past year (Yes/No)? (iv) Is Picasso your favorite 20th-century painter (Yes/No)?

Each respondent provides a personal answer and also a prediction of the empirical distribution of answers (i.e., the fraction of people endorsing each answer). Predictions are scored for accuracy, that is, for how well they match the empirical frequencies. The personal answers, which are the main object of interest, are scored for being surprisingly common. An answer endorsed by 10% of the population against a predicted frequency of 5% would be surprisingly common and would receive a high information score; if predictions averaged 25%, it would be a surprisingly uncommon answer, and hence receive a low score.

The surprisingly common criterion exploits an overlooked implication of Bayesian reasoning about population frequencies. Namely, in most situations, one should expect that others will underestimate the true frequency of one's own opinion or personal characteristic. This implication is a corollary to the more usual Bayesian argument that the highest predictions of the frequency of a given opinion or characteristic in the population should come from individuals who hold that opinion or characteristic, because holding the opinion constitutes a valid and favorable signal about its general popularity (11, 12). People who, for example, rate Picasso as their favorite should—and usually do (13)—give higher estimates of the percentage of the population who shares that opinion, because their own feelings are an informative “sample of one” (14). It follows, then, that Picasso lovers, who have reason to believe that their best estimate of Picasso popularity is high compared with others' estimates, should conclude that the true popularity of Picasso is underestimated by the population. Hence, one's true opinion is also the opinion that has the best chance of being surprisingly common.

The validity of this conclusion does not depend on whether the personally truthful answer is believed to be rare or widely shared. For example, a male who has had more than 20 sexual partners [answering question (iii)] may feel that few people fall in this promiscuous category. Nevertheless, according to Bayesian reasoning, he should expect that his personal estimate of the percentage (e.g., 5%) will be somewhat higher than the average of estimates collected from the population as a whole (e.g., 2%). The fact that he has had more than 20 sexual partners is evidence that the general population, which includes persons with fewer partners, will underestimate the prevalence of this profile.

Truth-telling is individually rational in the sense that a truthful answer maximizes expected information score, assuming that everyone is responding truthfully [hence, it is a Bayesian Nash equilibrium (15)]. It is also collectively rational in the sense that no other equilibrium provides a higher expected information score, for any respondent. In actual applications of the method, one would not teach respondents the mathematics of scoring or explain the notion of equilibrium. Rather, one would like to be able to tell them that truthful answers will maximize their expected scores, and that in arriving at their personal true answer they are free to ignore what other respondents might say. The equilibrium analysis confirms that under cer-

Massachusetts Institute of Technology, Sloan School of Management, E56-320, 38 Memorial Drive, Cambridge, MA 02139, USA. E-mail: dprelec@mit.edu

tain conditions one can make such a claim honestly.

The equilibrium results rest on two assumptions. First, the sample of respondents must be sufficiently large so that a single answer cannot appreciably affect empirical frequencies (16). The results do hold for large finite populations but are simpler to state for a countably infinite population, as is done here. Respondents are indexed by  $r \in \{1, 2, \dots\}$ , and their truthful answer to a  $m$  multiple-choice question by  $t^r = (t_1^r, \dots, t_m^r)$  ( $t_k^r \in \{0, 1\}$ ,  $\sum_k x_k^r = 1$ ).  $t_k^r$  is thus an indicator variable that has a value of one or zero depending on whether answer  $k$  is or is not the truthful answer of respondent  $r$ . The truthful answer is also called a personal opinion or characteristic.

Second, respondents treat personal opinions as an “impersonally informative” signal about the population distribution, which is an unknown parameter,  $\omega = (\omega_1, \dots, \omega_m) \in \Omega$  (17). Formally, I assume common knowledge (18) by respondents that all posterior beliefs,  $p(\omega|t^r)$ , are consistent with Bayesian updating from a single distribution over  $\omega$ , also called a common prior,  $p(\omega)$ , and that:  $p(\omega|t^r) = p(\omega|t^s)$  if and only if  $t^r = t^s$ . Opinions thus provide evidence about  $\omega$ , but the inference is impersonal: Respondents believe that others sharing their opinion will draw the same inference about population frequencies (19). One can therefore denote a generic respondent with opinion  $j$  by  $t_j$  and suppress the respondent superscript from joint and conditional probabilities:  $Prob\{t_j^r = 1 \mid t_i^s = 1\}$  becomes  $p(t_j|t_i)$ , and so on.

For a binary question, one may interpret the model as follows. Each respondent privately and independently conducts one toss of a biased coin, with unknown probability  $\omega_H$  of heads. The result of the toss represents his opinion. Using this datum, he forms a posterior distribution,  $p(\omega_H|t^r)$ , whose expectation is the predicted frequency of heads. For example, if the prior is uniform, then the posterior distribution following the toss will be triangular on  $[0, 1]$ , skewed toward heads or tails depending on the result of the toss, with an expected value of one-third or two-thirds. However, if the prior is not uniform but strongly biased toward the opposite result (i.e., tails), then the expected frequency of heads following a heads toss might still be quite low. This would correspond to a *prima facie* unusual characteristic, such as having more than 20 sexual partners within the previous year.

An important simplification in the method is that I never elicit prior or posterior distributions, only answers and predicted frequencies. Denoting answers and predictions by  $x^r = (x_1^r, \dots, x_m^r)$  ( $x_k^r \in \{0, 1\}$ ,  $\sum_k x_k^r = 1$ ) and  $y^r = (y_1^r, \dots, y_m^r)$  ( $y_k^r \geq 0$ ,  $\sum_k y_k^r = 1$ ), respectively,

I calculate the population endorsement frequencies,  $\bar{x}_k$ , and the (geometric) average,  $\bar{y}_k$ , of predicted frequencies,

$$\bar{x}_k = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{r=1}^n x_k^r,$$

$$\log \bar{y}_k = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{r=1}^n \log y_k^r$$

Instead of applying a preset answer key, we evaluate answers according to their information score, which is the log-ratio of actual-to-predicted endorsement frequencies. The information score for answer  $k$  is

$$\log \frac{\bar{x}_k}{\bar{y}_k} \tag{1}$$

At least one answer will have a nonnegative information score. Variance in predictions tends to lower all  $\bar{y}_k$  values and hence raises information scores.

The total score for a respondent combines the information score with a separate score for the accuracy of predictions (20):

$$\begin{aligned} \text{score for respondent } r = \\ \text{information score} + \text{prediction score} = \\ \sum_k x_k^r \log \frac{\bar{x}_k}{\bar{y}_k} + \alpha \sum_k \bar{x}_k \log \frac{y_k^r}{\bar{x}_k}, 0 < \alpha \end{aligned} \tag{2}$$

Equation 2 is the complete payoff equation for the game. It is symmetric, and zero-sum if  $\alpha = 1$ . The first part of the equation selects a single information-score value, given that  $x_k^r = 0$  for all answers except the one endorsed by  $r$ . The second part is a penalty proportional to the relative entropy (or Kullback-Leibler divergence) between the empirical distribution and  $r$ 's prediction of that distribution (21, 22). The best prediction score is zero, attained when prediction exactly matches reality,  $y_k^r = \bar{x}_k$ . **Expected prediction score is maximized by reporting expected frequencies,  $y_k^r = E\{\bar{x}_k|t^r\}$**  (2). The constant  $\alpha$  fine-tunes the weight given to prediction error.

To see how this works in the simple coin toss setting, imagine that there are only two equally likely possibilities: Either the coin is fair, or it is unfair, in which case it always comes up heads. A respondent who privately observes a single toss of tails knows that the coin is fair, and predicts a 50-50 split of observations. A respondent observing heads lowers the probability of fairness from the prior 1/2 to a posterior of 1/3, in accord with Bayes' rule, which in turn yields a predicted (i.e., expected) frequency of 1/6 for tails (multiplying 1/3 by 1/2). From the perspective of someone observing tails, the expectation of others' predictions of the frequency of tails will be a mix of predictions of 1/2 (from those

tossing tails) and 1/6 (from those tossing heads), yielding a geometric mean clearly lower than his or her predicted frequency of 1/2. Hence, he or she expects that tails will prove to be more common than predicted and receive a positive information score. By contrast, heads is expected to be a surprisingly uncommon toss, because the predicted frequency of 1/2 is lower than the expectation of others' predictions, which is a mix of 1/2 and 5/6 predictions. A similar argument would show that those who draw heads should expect that heads will prove to be the answer with the high information score.

The example illustrates a general property of information scores. Namely, a truthful answer constitutes the best guess about the most surprisingly common answer, if “best” is defined precisely by expected information score and if other respondents are answering truthfully and giving truthful predicted frequencies. This property does not depend on the number of possible answers or on the prior (23). It leads directly to the equilibrium result [proof in the supporting online material (SOM) text].

For this theorem, assume that (i) every respondent  $r$  with opinion  $t^r$  forms a posterior over the population distribution of opinions,  $p(\omega|t^r)$ , by applying Bayes' rule to a common prior  $p(\omega)$ ; (ii)  **$p(\omega|t^r) = p(\omega|t^s)$  if and only if  $t^r = t^s$** ; and (iii) scores are computed according to Eq. 2. Then, (T1) truth-telling is a Nash equilibrium for any  $\alpha > 0$ ; Truth-telling maximizes expected total score of every respondent who believes that others are responding truthfully; (T2) expected equilibrium information scores are nonnegative and attain a maximum for all respondents in the truth-telling equilibrium; (T3) for  $\alpha = 1$ , the game is zero-sum, and the total scores in the truth-telling equilibrium equal  $\log p(\omega|t^r) + K$ , with  $K$  set by the zero-sum constraint.

Truth-telling is defined as truthful answers,  $x^r = t^r$ , and truthful predictions,  $y^r = E\{\omega|t^r\}$ . T2 states that although there are other equilibria, constructed by mapping multiple true opinions into a single response category or by randomization, these less revealing equilibria result in lower information scores for all respondents. If needed, one can enhance the strategic advantage of truth-telling by giving relatively more weight to information score in Eq. 2 (24). For sufficiently small  $\alpha$ , the expected total scores in the truth-telling equilibrium will Pareto-dominate expected scores in any other equilibrium. T3 shows that by setting  $\alpha = 1$  we also have the option of presenting the survey as a purely competitive, zero-sum contest. Total scores then rank respondents according to how well they anticipate the true distribution of answers. Note that the scoring system asks only for

the expected distribution of true answers,  $E\{\omega|t^r\}$  and not for the posterior distribution  $p(\omega|t^r)$ , which is an  $m$ -dimensional probability density function. Remarkably, one can infer which respondents assign more probability to the actual value of  $\omega$  by means of a procedure that does not elicit these probabilities directly.

In previous economic research on incentive mechanisms, it has been standard to assume that the scorer (or the “center”) knows the prior and posteriors and incorporates this knowledge into the scoring function (7–9, 25). In principle, any change in the prior, whether caused by a change in question wording, in the composition of the sample, or by new public information, would require a recalculation of the scoring functions. By contrast, my method employs a universal “one-size-fits-all” scoring equation, which makes no mention of prior or posterior probabilities. This has three benefits for practical application. First, questions do not need to be limited to some pretested set for which empirically estimated base rates and conditional probabilities are available; instead, one can use the full resources of natural language to tailor a new set of questions for each application. Second, it is possible to apply the same survey to different populations, or in a dynamic setting (which is relevant to political polling). Third, one can honestly instruct respondents to refrain from speculating about the answers of others while formulating their own answer. Truthful answers are optimal for any prior, and there are no posted probabilities for them to consider, and perhaps reject.

These are decisive advantages when it comes to scoring complex, unique questions. In particular, one can apply the method to elicit honest probabilistic judgments about the truth value of any clearly stated proposition, even if actual truth is beyond reach and no prior is available. For example, a recent book, *Our Final Century*, by a noted British astronomer, gives the chances of human survival beyond the year 2100 at no better than 50:50 (26). It is a provocative assessment, which will not be put to the test anytime soon. With the present method, one could take the question: “Is this our final century?” and submit it to a sample of experts, who would each provide a subjective probability and also estimate probability distributions over others’ probabilities. T1 implies that honest reporting of subjective probabilities would maximize expected information score. Experts would face comparable truth-telling incentives as if they were betting on the actual outcome [e.g., as in a futures market (27)] and that outcome could be determined in time for scoring.

I illustrate this with a discrete computation, which assumes that probabilities are elicited at 1% precision by means of a 100-

point multiple-choice question (in practice, one would have fewer categories and smooth out the empirical frequencies). The population vector  $\omega = (\omega_{00}, \dots, \omega_{99})$  indexes the unknown distribution of such probabilities among experts. Given any prior,  $p(\omega)$ , it is a laborious but straightforward exercise to calculate expected information score as function of true personal probability and endorsed probability. Figure 1, lines A90 and B90, present the result of such calculations, with two different priors,  $p_A(\omega)$  and  $p_B(\omega)$ , for experts who happen to agree that the probability of disaster striking before 2100 is 90%. The experts thus share the same assessment but have different theories about how their assessment is related to the assessment of others. Although lines A90 and B90 differ, the expected information score is in both cases maximized by a truthful endorsement of 90%. This confirms T1. In both cases, each expert believes that his subjective probability is pessimistic relative to the population: The expectation of others’ probabilities, conditioned on a personal estimate of 90%, is only 65% with  $p_A(\omega)$  and 54% with  $p_B(\omega)$ .

If the subjective probability shifts to 50%, the lines move to A50, B50, and the optimum, in both cases, relocates to 50%. Hence, the optimum automatically tracks changes in subjective belief, in this case the subjective probability of an unknown future event, but is invariant with respect to assumptions about how that belief is related to beliefs of other individuals. Changing these assumptions will simply lead back to the same recommendation: Truthfully report subjective probability.

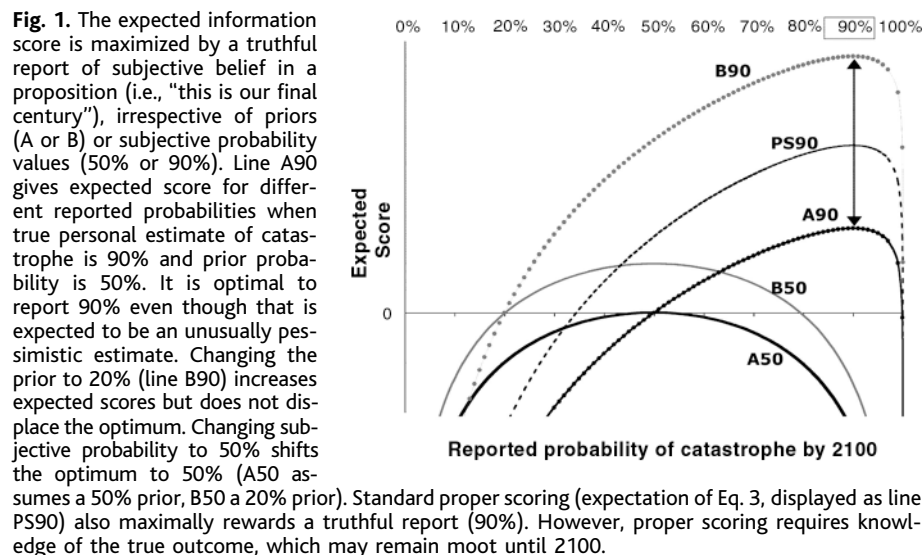
Respondents are thus free to concentrate on their personal answer and need not worry about formulating an adequate prior. Any model of the prior is likely to be complex and involve strong assumptions. For example, in the calculations in Fig. 1, I assumed that

experts’ estimates are based on a private signal, distributed between zero and one, representing a personal assessment of the credibility of evidence supporting the bad outcome. The “credibility signal” is a valid but stochastic indicator of the true state of affairs: On the bad scenario, credibility signals are independent draws from a uniform distribution, so that some experts “get the message” and some do not; on the good scenario, they are independent draws from a triangular distribution, peaking at zero (no credibility) and declining linearly to one (full credibility). A prior probability of catastrophe then induces a monotonic mapping from credibility signals to posterior probabilities of catastrophe, as well as a prior over experts’ probability estimates,  $p(\omega)$ .

Lines A and B differ in that the prior probability of catastrophe is presumed to be 50% for line A and 20% for line B. Expected scores are higher for B, because the 90% estimate is more surprising in that case.

One could question any of the assumptions of this model (28). However, changing the assumptions would not move the optimum, as long as the impersonally informative requirement is preserved. (The impersonally informative requirement means that two experts will estimate the same probability of catastrophe if and only if they share the same posterior distribution over other experts’ probabilities). Thus, even though information scoring conditions success on the answers of other people, the respondent does not need to develop a theory of other people’s answers; the most popular answer has no advantage of “winning,” and the entire structure of mutual beliefs, as embodied in the prior, is irrelevant.

It is instructive to compare information scores with scores that would be computed if the scorer had a crystal ball and could score



estimates for accuracy. The standard instrument for eliciting honest probabilities about publicly verifiable events is the logarithmic proper scoring rule (2, 4, 29). With the rule, an expert who announces a probability distribution  $z = (z_1, \dots, z_n)$  over  $n$  mutually exclusive events would receive a score of

$$K + \log z_i \tag{3}$$

if event  $i$  is realized. For instance, an expert whose true subjective probability estimate that humanity will perish by 2100 is 90%, but who announced a possibly different probability  $z$ , would calculate an expected score of  $0.9 \log z + 0.1 \log(1 - z)$ , assuming, again, that there was some way to establish the true outcome. This expectation is maximized at the true value,  $z = 0.90$ , as shown by line PS90 in Fig. 1 (elevation is arbitrary). It is hard to distinguish proper scoring, which requires knowledge of the true outcome, from information scoring, which does not require such knowledge (30).

There are two generic ways in which the assumption of an impersonally informative prior might fail. First, a true answer might not be informative about population frequencies in the presence of public information about these frequencies (inducing a sharp prior). For instance, a person's gender would have minimal impact on their judgment of the proportion of men and women in the population. This would be a case of  $t^r \neq t^s$  but  $p(\omega|t^r) \equiv p(\omega|t^s)$ , and the difference between expected information scores for honest and deceptive answers would be virtually zero (though still positive). As shown below, the remedy is to combine the gender question with an opinion question that interacts with gender.

Second, respondents with different tastes or characteristics might choose the same answer for different reasons and hence form different posteriors. For example, someone with

nonstandard political views might treat his or her liking for a candidate as evidence that most people will prefer someone else. This would be a case of:  $p(\omega|t^r) \neq p(\omega|t^s)$  although  $t^r = t^s$ . Here, too, the remedy is to expand the questionnaire, allowing the person to reveal both the opinion and characteristic.

A last example, an art evaluation, illustrates both remedies. The example assumes existence of experts and laymen, and a binary state-of-nature: a question of whether a particular artist either does or does not represent an original talent. By hypothesis, art experts recognize this distinction quite well, but laymen discriminate poorly and, indeed, have a higher chance of enjoying a derivative artist than an original one. The fraction of experts is common knowledge, as are the other probabilities (Table 1).

In the short version of the survey, respondents only state their opinion; in the long version, they also report their expertise. Table 1 displays expected information scores for all possible answers, as a function of opinion and expertise. With the short version, truth-telling is optimal for experts but not for laymen, who do have a slight incentive to deceive if they happen to like the exhibition. With the long version, however, the diagonal, truth-telling entries have highest expected score. In particular, respondents will do better if they reveal their true expertise even though the distribution of expertise in the surveyed population is common knowledge.

Expected information scores in this and other examples reflect the amount of information associated with a particular opinion or characteristic. In Table 1, experts have a clear advantage even though they comprise a minority of the sample, because their opinion is more informative about population frequencies. In general, the expected information score for opinion  $i$  equals the expected relative entropy between distribution  $p(\omega|t_k, t_i)$  and  $p(\omega|t_k)$ , averaged over all  $t_k$ . In words,

the expected score for  $i$  is the information-theoretic measure of how much endorsing opinion  $i$  shifts others' posterior beliefs about the population distribution. An expert endorsement will cause greater shift in beliefs, because it is more informative about the underlying variables that drive opinions for both segments (31). This measure of impact is quite insensitive to the size of the expert segment or to the direction of association between expert and nonexpert opinion.

By establishing truth-telling incentives, I do not suggest that people are deceitful or unwilling to provide information without explicit financial payoffs. The concern, rather, is that the absence of external criteria can promote self-deception and false confidence even among the well-intentioned. A futurist, or an art critic, can comfortably spend a lifetime making judgments without the reality checks that confront a doctor, scientist, or business investor. In the absence of reality checks, it is tempting to grant special status to the prevailing consensus. The benefit of explicit scoring is precisely to counteract informal pressures to agree (or perhaps to "stand out" and disagree). Indeed, the mere existence of a truth-inducing scoring system provides methodological reassurance for social science, showing that subjective data can, if needed, be elicited by means of a process that is neither faith-based ("all answers are equally good") nor biased against the exceptional view.

References and Notes

1. C. F. Turner, E. Martin, Eds., *Surveying Subjective Phenomena* (Russell Sage Foundation, New York, 1984), vols. I and II.
2. R. M. Cooke, *Experts in Uncertainty* (Oxford Univ. Press, New York, 1991).
3. A formalized scoring rule has diverse uses: training, as in psychophysical experiments (32); communicating desired performance (33); enhancing motivation and effort (34); encouraging advance preparation, as in educational testing; attracting a larger and more representative pool of respondents; diagnosing suboptimal judgments (4); and identifying superior respondents.
4. R. Winkler, *J. Am. Stat. Assoc.* **64**, 1073 (1969).
5. W. H. Batchelder, A. K. Romney, *Psychometrika* **53**, 71 (1988).
6. In particular, this precludes the application of a futures markets (27) or a proper scoring rule (29).
7. C. d'Aspremont, L.-A. Gerard-Varet, *J. Public Econ.* **11**, 25 (1979).
8. S. J. Johnson, J. Pratt, R. J. Zeckhauser, *Econometrica* **58**, 873 (1990).
9. P. McAfee, P. Remy, *Econometrica* **60**, 395 (1992).
10. H. A. Linstone, M. Turoff, *The Delphi Method: Techniques and Applications* (Addison-Wesley, Reading, MA, 1975).
11. R. M. Dawes, in *Insights in Decision Making*, R. Hogarth, Ed. (Univ. of Chicago Press, Chicago, IL, 1990), pp. 179-199.
12. S. J. Hoch, *J. Pers. Soc. Psychol.* **53**, 221 (1987).
13. It is one of the most robust findings in experimental psychology that participants' self-reported characteristics—behavioral intentions, preferences, and beliefs—are positively correlated with their estimates of the relative frequency of these characteristics (35). The psychological literature initially regarded this as an egocentric error of judgment (a "false consensus") (36) and did not consider the Bayesian

**Table 1.** An incomplete question can create incentives for misrepresentation. The first pair of columns gives the conditional probabilities of liking the exhibition as function of originality (so that, for example, experts have a 70% chance of liking an original artist). It is common knowledge that 25% of the sample are experts, and that the prior probability of an original exhibition is 25%. The remaining columns display expected information scores. Answers with highest expected information score are shown by bold numbers. Truth-telling is optimal in the long version but not in the short version of the survey.

Opinion	Probability of opinion conditional on quality of exhibition		Expected score						
	Original	Derivative	Long version				Short version		
			Expert claim		Layman claim		Like	Dislike	
		Like	Dislike	Like	Dislike				
Expert									
Like	70%	10%	<b>+575</b>	-776	-462	+67	<b>+191</b>	-57	
Dislike	30%	90%	-934	<b>+95</b>	+84	-24	-86	<b>+18</b>	
Layman									
Like	10%	20%	-826	+32	<b>+45</b>	-18	-66	<b>+12</b>	
Dislike	90%	80%	-499	-156	-73	<b>+2</b>	-6	<b>-4</b>	

explanation, as was pointed out by Dawes (11, 14). There is still some dispute over whether the relationship is entirely consistent with Bayesian updating (37).

14. R. M. Dawes, *J. Exp. Soc. Psychol.* **25**, 1 (1989).
15. D. Fudenberg, J. Tirole, *Game Theory* (MIT Press, Cambridge, MA, 2000).
16. With finite players, the truth-telling result holds provided that the number of players exceeds some finite  $n$ , which in turn depends on  $p(\omega)$ .
17. J. M. Bernardo, A. F. M. Smith, *Bayesian Theory*, Wiley Series in Probability and Statistics (Wiley, New York, 2000).
18. R. J. Aumann, *Econometrica* **55**, 1 (1987).
19. More precisely, I assume in the SOM text that for any finite subset of respondents, there is a common and exchangeable prior over their opinions (hence, the prior is invariant under permutation of respondents). By de Finetti's representation theorem (25), this implies the existence of a probability distribution,  $p(\omega)$ , such that opinions are independent conditional on  $\omega$ . Conditional independence ensures  $\tau^r = \tau^s \Rightarrow p(\omega|\tau^r) = p(\omega|\tau^s)$ . The reverse implication (i.e., that different opinions imply different posteriors) is also called stochastic relevance (8).
20. The finite  $n$ -player scoring formula ( $n \geq 3$ ), for respondent  $r$ , is

$$\sum_{s \neq r} \sum_k \bar{x}_k \log \frac{\bar{x}_k^{-rs}}{\bar{y}_k^{-rs}} + \alpha \sum_{s \neq r} \sum_k \bar{x}_k^{-rs} \log \frac{y_k^r}{\bar{x}_k^{-rs}}$$

where  $\bar{x}_k^{-rs} = \left( \sum_{q \neq r, s} x_k^q + 1 \right) / (n + m - 2)$  and  $\log \bar{y}_k^{-rs} = \sum_{q \neq r, s} \log y_k^q / (n - 2)$ . The score for  $r$  is built up from pairwise comparisons of  $r$  against all other respondents  $s$ , excluding from the pairwise calculations the answers and predictions of respondents  $r$  and  $s$ . To prevent infinite scores associated with zero frequencies, I replace the empirical frequencies with Laplace estimates derived from these frequencies. This is equivalent to "seeding" the empirical sample with one extra answer for each possible choice. Any distortion in incentives can be made arbitrarily small by increasing the number of respondents,  $n$ . The scoring is zero-sum when  $\alpha = 1$ .

21. T. M. Cover, J. A. Thomas, *Elements of Information Theory* (Wiley, New York, 1991).
22. S. Kullback, *Information Theory and Statistics* (Wiley, New York, 1954).
23. The key step in the proof involves calculation of expected information score for someone with per-

sonal opinion  $i$  but endorsing a possibly different answer  $j$ ,

$$E\{\log \frac{\bar{x}_j}{\bar{y}_j} | t_i\} = \int_{\Omega} p(\omega | t_i) E\{\log \frac{\bar{x}_j}{\bar{y}_j} | \omega\} d\omega \quad (a)$$

$$= \int_{\Omega} p(\omega | t_i) \sum_{k=1}^m \omega_k \log \frac{\omega_j}{p(t_j | t_k)} d\omega \quad (b)$$

$$= \sum_{k=1}^m p(t_k | t_i) \int_{\Omega} p(\omega | t_k, t_i) \times \log \frac{p(t_j | \omega) p(t_k | t_j, \omega)}{p(t_j | t_k) p(t_k | \omega)} d\omega \quad (c)$$

$$= \sum_{k=1}^m p(t_k | t_i) \int_{\Omega} p(\omega | t_k, t_i) \times \log \frac{p(\omega | t_k, t_j)}{p(\omega | t_k)} d\omega. \quad (d)$$

Once we reach (d), we can use the fact that the integral,

$$\int_{\Omega} p(\omega | t_k, t_i) \log p(\omega | t_k, t_j) d\omega,$$

is maximized when:  $p(\omega | t_k, t_i) = p(\omega | t_k, t_j)$ , to conclude that a truthful answer,  $i$ , will have higher expected information score than any other answer  $j$ . To derive (d), we first compute expected information score (a) with respect to the posterior distribution,  $p(\omega | t_i)$ , and use the assumption that others are responding truthfully to derive (b). For an infinite sample, truthful answers imply:  $\bar{x}_j = \omega_j$ , and truthful predictions:  $\log \bar{y}_j = \sum_k \omega_k \log p(t_j | t_k)$ , because the fraction  $\omega_k$  of respondents who draw  $k$  will predict  $p(t_j | t_k)$  for answer  $j$ . To derive (c) from (b), we apply conditional independence to write  $\omega_k p(\omega | t_i)$  as  $p(t_k | t_i) p(\omega | t_k, t_i)$ ,  $\omega_j$  as  $p(t_j | \omega)$ , and 1 as  $p(t_k | t_j, \omega) / p(t_k | \omega)$ , which is inserted into the fraction. (d) follows from (c) by Bayes' rule.

24. The scoring system is not easy to circumvent by collective collusion, because if everyone agrees to give the same response then that response will no longer be surprisingly common, and will receive a zero information score. The prediction scores will also be zero in that case.
25. J. Cremer, R. P. McLean, *Econometrica* **56**, 1247 (1988).
26. M. Rees, *Our Final Century* (Heinemann, London, 2003).
27. J. Berg, R. Forsythe, T. A. Rietz, in *Understanding*

*Strategic Interaction: Essays in the Honor of Reinhard Selten*, W. Albers, W. Guth, B. Hammerstein, B. Moldovanu, E. van Damme, Eds. (Springer, New York, 1997), pp. 441–463.

28. Certainly, the assumption of independent credibility signals is unrealistic in that it implies that expert opinion can in aggregate predict the true outcome perfectly; a more realistic model would have to interpose some uncertainty between the outcome and the totality of expert opinion.
29. L. J. Savage, *J. Am. Stat. Assoc.* **66**, 783 (1971).
30. Information scoring is nonmetric, and the 100 probability levels are treated as simply 100 distinct response categories. The smooth lines in Fig. 1 reflect smooth underlying priors,  $p_x(\omega)$  and  $p_y(\omega)$ . Unlike proper scoring, information scoring could be applied to verbal expressions of probability ("likely," "impossible," etc.).
31. Precisely, if the opinions of one type of respondent are a statistical "garbling" of the opinions of a second type, then the first type will receive a lower score in the truth-telling equilibrium. Garbling means that the more informed individual could replicate the statistical properties of the signal received by the less informed individual, simply by applying a randomization device to his own signal (38).
32. D. M. Green, J. A. Swets, *Signal Detection Theory and Psychophysics* (Peninsula Publishing, Los Altos, CA, 1989).
33. W. Edwards, *Psychol. Rev.* **68**, 275 (1961).
34. C. F. Camerer, R. Hogarth, *J. Risk Uncert.* **18**, 7 (1999).
35. G. Marks, N. Miller, *Psychol. Bull.* **102**, 72 (1987).
36. L. Ross, D. Greene, P. House, *J. Exp. Soc. Psychol.* **13**, 279 (1977).
37. J. Krueger, R. W. Clement, *J. Pers. Soc. Psychol.* **67**, 596 (1994).
38. D. Blackwell, *Ann. Math. Stat.* **24**, 265 (1953).
39. I thank D. Mijovic-Prelec, S. Frederick, D. Fudenberg, J. R. Hauser, M. Kearns, E. Kugelberg, R. D. Luce, D. McAdams, S. Seung, R. Weaver, and B. Wernerfelt for comments and criticism. I acknowledge early support for this research direction by Harvard Society of Fellows, MIT E-Business Center, and the MIT Center for Innovation in Product Development.

**Supporting Online Material**

www.sciencemag.org/cgi/content/full/306/5695/462/DC1  
SOM Text

28 June 2004; accepted 15 September 2004

## Single-Atom Spin-Flip Spectroscopy

A. J. Heinrich,\* J. A. Gupta, C. P. Lutz, D. M. Eigler

We demonstrate the ability to measure the energy required to flip the spin of single adsorbed atoms. A low-temperature, high-magnetic field scanning tunneling microscope was used to measure the spin excitation spectra of individual manganese atoms adsorbed on Al<sub>2</sub>O<sub>3</sub> islands on a NiAl surface. We find pronounced variations of the spin-flip spectra for manganese atoms in different local environments.

The magnetic properties of nanometer-scale structures are of fundamental interest and may play a role in future technologies, including classical and quantum computation. Such mag-

netic structures are composed of magnetic atoms in precise arrangements. The magnetic properties of each atom are profoundly influenced by its local environment. Magnetic properties of atoms in a solid can be probed by placing the atoms in tunnel junctions. Early experiments with planar metal-oxide-metal tunnel junctions doped with paramagnetic impurities exhibited surprisingly complex con-

ductance spectra described as "zero-bias anomalies" (1–4). Such anomalies were shown to reflect both spin-flips driven by inelastic electron scattering and Kondo interactions of magnetic impurities with tunneling electrons (5–7). Single, albeit unknown, magnetic impurities were later studied in nanoscopic tunnel junctions (8, 9). Recently, magnetic properties of single-molecule transistors that incorporated either one or two magnetic atoms were probed by means of their elastic conductance spectra (10, 11). These measurements determined  $g$  values and showed field-split Kondo resonances due to the embedded magnetic atoms.

The scanning tunneling microscope (STM) offers the ability to study single magnetic moments in a precisely characterized local environment and to probe the variations in magnetic properties with atomic-scale spatial resolution. Previous STM studies of atomic-scale magnetism include Kondo resonances of magnetic atoms on metal surfaces (12, 13), increased noise at the Larmor frequency (14, 15),

IBM Research Division, Almaden Research Center, 650 Harry Road, San Jose, CA 95120, USA.

\*To whom correspondence should be addressed. E-mail: heinrich@almaden.ibm.com