

## **Creating Truthtelling Incentives with the Bayesian Truth Serum**

The “Bayesian Truth Serum” (BTS) is a survey scoring method that provides truthtelling incentives for respondents answering multiple-choice questions about intrinsically private matters: opinions, tastes, past behavior. The method requires respondents to supply not only their own answers, but also percentage estimates of others’ answers. The formula then assigns high scores to answers that are surprisingly common, i.e. whose actual frequency exceeds their predicted frequency. Two studies demonstrate that this method both encourages and rewards truthful responses. First, we conducted a general knowledge questionnaire in which we listed items such as brand names, famous people, and scientific terms. One-third of the items were nonexistent foils. Respondents who were paid for higher BTS scores claimed to recognize fewer foils, even when given competing incentives to overclaim their knowledge. Participants also earned more money when they denied knowledge of foils. Our second study explored whether survey takers could exploit the system using strategic deception. We found that they could not: on four surveys with diverse content, truthtelling outscored the wide variety of deception strategies we tested.

Keywords: Truthtelling incentives, survey design, Bayesian inference, scoring rules, false consensus effect.

## *INTRODUCTION*

In opinion research as traditionally conducted, respondents are given no incentives for performance — for the quality or usefulness of their answers. They may, of course, be compensated for time and effort, but the level of compensation does not hinge on the particular answers that they provide. There is, in other words, no hidden answer key by which the survey administrator judges some answers to a question as more worthy of compensation than others.

The reason for this is straightforward: when questions deal with intrinsically private matters — a respondent's opinions, preferences, intentions, or past behaviors — then the correct answer for a particular individual is simply the answer that best matches his private opinions or preferences, and the survey administrator is in no position to judge whether any given answer really does reflect these. To evaluate answers, the administrator would apparently need to know which answer is personally correct for each respondent. But such an omniscient administrator would not need to conduct a survey in the first place.

Prelec (2004) proposed a “Bayesian Truth Serum” (BTS) scoring method that provides incentives for providing truthful — in the dual sense of honest and carefully considered — answers to questions dealing with personal information. The key idea behind BTS is to assign a high score to an answer whose actual frequency is greater than its predicted frequency, with predictions drawn from the same population that supplies the answers. To implement this method, the administrator asks each respondent to provide not only a personal answer, but also estimates in percentage terms of how other respondents will answer the same question. With this additional input, a numerical answer key can be generated in which the “information score” for any answer is the log-ratio:

$$(1) \quad \text{Information score for an answer} = \log \frac{\text{actual relative frequency of the answer}}{\text{(geometric) mean predicted frequency of the answer}}$$

Although (1) identifies some answers as “winners” and other answers as “losers” after the survey is analyzed, each respondent has reason to believe that the answer that matches his own private opinion has the best chance of achieving a high score. Specifically, the BTS theorem states that under certain general conditions, personally truthful answers maximize the expected information score for any respondent who believes that others are answering truthfully — i.e., that they are giving truthful answers and optimal Bayesian predictions of the distribution of answers. The scoring system transforms a survey into a competitive, zero-sum contest, in which truth-telling is a strict Bayesian Nash equilibrium (Prelec 2004). Like other Bayesian mechanisms, BTS exploits the subjective correlation between one’s opinion and the opinions of others (Cremer and McLean 1988; d’Aspremont and Gerard-Varet 1979; Johnson, Pratt and Zeckhauser 1990; McAfee and Reny 1992; McLean and Postlewaite 2002; Miller, Resnick and Zeckhauser 2005). However, unlike previous mechanisms, BTS does not incorporate assumptions about this correlation into the scoring function. Here, the function is generic, not requiring any input from the survey administrator.

We may contrast (1) with consensus scoring, which would assign a high score to the most popular answer. Consensus scoring creates incentives for deception by respondents who suspect that their opinion is in the minority. The information scoring criterion, however, creates no such incentives: untruthful answers have lower expected scores, irrespective of whether the respondent believes that his opinion is common or rare. In this sense, the BTS system levels the playing field between typical and atypical opinions.

Here we provide the first experimental evidence that BTS both encourages and rewards truth-telling. Our first study demonstrates that the prospect of payment based on one's information score creates a credible and persuasive incentive to tell the truth, even when there are competing incentives to deceive. We conduct a general knowledge questionnaire in which we ask respondents if they recognize various items: electronics brand names, historical figures, philosophy terms, etc. Paulhus et al. (2003) find that people with a need for self-enhancement tend to overclaim their knowledge on such questionnaires. By including nonexistent foils alongside the real items, we can measure the degree of deception. We also give some respondents an extra incentive to overclaim by promising to pay them for each item they claim to know. When significant bonus payments are awarded to the survey takers with the highest BTS scores, people claim to recognize fewer foils than when bonuses are awarded randomly. The study also validates our claim that truth-telling is in the respondents' interest: people do in fact achieve higher scores, and earn more money, when they deny knowledge of foils.

In our second study, we investigate whether it is possible for survey takers to exploit the BTS system by engaging in strategic deception that they hope will be more profitable than answering truthfully. In four surveys, with content chosen to be neutral enough that we can plausibly treat actual answers as truthful, we compare information scores from actual responses to those resulting from various deception strategies. For example, we test whether respondents can score higher by giving the answers they believe will be most popular, rather than their true opinions. We also examine whether respondents do better by misrepresenting their demographic characteristics (gender), or by simulating the answers of some other person they know well. We find that genuine answers reliably outperform every deception policy we test, and that no identifiable subgroup of respondents can expect to benefit from deception.

We begin by providing a simple example of how the BTS method rewards truth-telling when the truth is unknowable to the administrator, and by briefly reviewing Prelec's (2004) result. This is followed by descriptions of our two experimental studies and a concluding section.

### ***BTS SCORING THEORY***

#### ***Intuition Behind the "Surprisingly Common" Criterion***

BTS scoring works at the level of a single question. For example, we might ask: "Imagine that your host offers you a glass of wine before dinner. Would you prefer red or white?" To implement BTS scoring, each respondent must both give his own answer and predict the fraction of people who will endorse each answer. *Both* components of his response are scored: predictions for accuracy, and personal answers for being "surprisingly common," i.e. more common than collectively predicted.

Prelec (2004) proves that truthful answers can be expected to be surprisingly common (and therefore high scoring). Developing the wine example will provide the intuition behind this result. Consider Sarah, an imagined survey respondent asked to choose between red and white wine. Suppose that Sarah personally prefers red, and predicts that a majority of wine drinkers share her preference. We can think of her predictions, shown in panel (a) of Figure 1, as her estimates of the numerator of equation (1) for the answers red and white, respectively.

\*\*\* INSERT FIGURE 1 ABOUT HERE \*\*\*

To determine which answer is most likely to be surprisingly common, we must also estimate the denominator of (1) — the sample group's collective predictions. To do so, we can usefully divide the survey group into people who prefer red as Sarah does, and those who prefer white. Suppose that in general, a person's own preference positively influences his estimate of the overall popularity of that preference. On the whole, other red drinkers will then have predictions roughly similar to Sarah's, as in panel (b), because their predictions are conditioned on the same private signal. Conversely, white drinkers are likely to estimate a smaller degree of preference for red (panel c). Because the overall group's prediction is a weighted average of these groups, we can expect the denominator of (1) to fall somewhere between the two. Comparing panels (a) and (d) of the figure, we see that by this reasoning, Sarah's best guess should be that red — her true preference — will be surprisingly common, and that white will be surprisingly uncommon.

We emphasize three points about this example. First, its logic holds regardless of whether one's opinion is in the majority, as in our example, or in the minority. Second, the same reasoning that leads Sarah to expect that her true answer will be surprisingly common applies to white wine drinkers, and more generally to any personally true answer on a multiple choice opinion survey. In *outcome*, of course, both answers cannot be correct. But in *expectation*, one's own preference is his best guess of the most surprisingly common response. Third, while the rather elaborate reasoning above provides insight into equation (1), the result does not depend on real respondents mimicking this thought process.

The method's validity does, however, depend on the assumption that respondents (consciously or unconsciously) use their own tastes as information about the popularity of those tastes among others. A great deal of experimental evidence supports this proposition. The seminal experiment was done by Ross, Greene and House (1977), who asked students whether

they would be willing to walk around campus wearing a sign that read “Repent.” Students who were themselves willing to wear the sign tended to give higher estimates of the proportion of others who would also oblige. This result has been replicated in dozens of studies (see Marks and Miller 1987 for a review), but some time passed before its normative status was properly addressed. Ross, Greene and House declared the effect a “false” consensus: an egocentric assumption that others are similar to ourselves. Dawes (1989; 1990), however, argued convincingly in favor of a Bayesian interpretation of the finding: predictions of behavior are correlated with one’s own behavior because people rationally update a prior belief based on a “sample of one.” Some debate remains about whether experimental subjects use sample information efficiently (Engelmann and Strobel 2000; Krueger and Clement 1994), but the evidence is overwhelmingly clear that people who hold a particular opinion or preference give higher than average estimates of the prevalence of that opinion or preference.<sup>1</sup>

### ***The BTS Scoring Formula***

Index questionnaire respondents by  $r \in \{1, 2, \dots\}$ , and their answer and predictions for an  $m$ -multiple choice question as  $x^r$  and  $y^r$ , respectively:  $x^r \in \{1, \dots, m\}$ , and  $y^r = (y_1^r, \dots, y_m^r)$ , ( $y_k^s \geq 0$ ,  $\sum_k y_k^s = 1$ ). (In the wine example above,  $m=2$ .) We can then calculate the sample frequencies,  $\bar{x}_k$ , and the (geometric) average of predicted frequencies,  $\bar{y}_k$ ,

$$(2) \quad \begin{aligned} \bar{x}_k &= \frac{1}{n} \sum_{r=1}^n I(x^r = k) \\ \log \bar{y}_k &= \frac{1}{n} \sum_{r=1}^n \log y_k^r, \end{aligned}$$

where  $I(\cdot)$  is the zero-one indicator function, and  $n$  is the sample size. Answers are evaluated according to their *information score*, which is the log-ratio of actual  $\bar{x}_k$  to predicted  $\bar{y}_k$

endorsement frequencies. The total BTS score then combines the information score with a separate score for the accuracy of predictions:

$$(3) \quad \text{BTS score for } r \equiv u(x^r = j, y^r) = \log \frac{\bar{x}_j}{\bar{y}_j} + \sum_{k=1}^m \bar{x}_k \log \frac{y_k^r}{\bar{x}_k}$$

$$= \text{Information score} + \text{Prediction score.}$$

Prelec (2004) proves that for the game defined by (3), truthtelling is a strict Bayesian Nash equilibrium. We do not repeat the details of this proof here, but rather highlight the assumptions on which it relies.

The proof first assumes that the number of survey takers is sufficiently large that each individual response has a negligible impact on the sample frequencies. Then truthtelling is a Nash equilibrium if truthful predictions and truthful answers maximize expected score. In the case of the prediction score component of (3), it is a standard result that truthful predictions are optimal with a log rule (Cooke 1991). The result that truthful *answers* are optimal rests on the assumption that people reason like Bayesian statisticians to construct their guesses about the distribution of responses. Let each respondent's truthful answer to a question with  $m$  possible answers be indexed by a random variable  $t^r \in \{1, \dots, m\}$  (as distinguished from his *reported* answer  $x^r$ ). The distribution of opinions in the infinite population is given by an  $m$ -dimensional vector,  $w = (w_1, \dots, w_m) \in W = D^m$ . We distinguish between respondents' posterior, which is to say their actual beliefs about this parameter,  $p(w | t^r)$ , and the hypothetical common prior distribution  $p(w)$ . We further assume that these densities have three characteristics. *Common prior*: it is common knowledge the posterior beliefs,  $p(w | t^r)$ , are consistent with Bayesian updating from a common prior distribution,  $p(w)$ . *Conditional independence*: opinions are independent, conditional on the actual distribution:  $p(t^r = k, t^s = i | w) = p(t^r = k | w) p(t^s = i | w)$ . *Stochastic*

*relevance*: respondents with different opinions have different posterior beliefs:  $k \neq i$  implies  $p(w | t^r = k) \neq p(w | t^s = i)$ .

Essentially, the proof assumes that in constructing their guesses people begin with some “prior belief” about the distribution of answers to the question under consideration, then update this prior with their own opinion, which they regard as a draw of a ball from an urn containing an unknown mixture  $w$  of differently colored balls. Before drawing the ball, they share the same beliefs  $p(w)$  about possible mixtures. After drawing, they update these beliefs using Bayes’ rule:  $p(w | t^r = k) = p(t^r = k | w)p(w)/p(t^r = k)$ , where  $p(t^r = k)$  is obtained by taking the expectation of  $p(w_1, \dots, w_m)$  on the  $k$ -th coordinate. Conditional independence implies that respondents who draw the same color also form the same estimate of the proportions in the urn (hence the drawn ball represents the only source of information about these proportions, apart from the common prior). Under these assumptions, the answer that maximizes a respondent’s expected information score is his true opinion:  $x^r = t^r$ .

We note that there is little evidence that a “prior” belief — uninformed by one’s own preference — has any psychological reality. Nevertheless, as the empirical literature on false consensus attests, the Bayesian model seems to be a good as-if description of judgment. Second, for the common prior and conditional independence assumptions to hold strictly, all respondents who share a given opinion must make the same predictions, assuming they truthfully report expected frequencies. Of course we don’t observe such unanimity in practice. Instead, there is considerable variation in predictions, which can be interpreted as a shared Bayesian posterior among like-minded people, plus a noise term. One purpose of our studies is to test whether information scoring is robust to this noise.

Although there are other equilibria, in which respondents either randomize or give the same answer irrespective of opinion, these tend to be strategically implausible. For example, if the population contains two subgroups, experts and nonexperts, such that experts know the nonexperts' opinions but not vice versa, then the nonexperts would receive a negative total BTS score in the truthtelling equilibrium, and would consequently prefer the uninformative equilibrium, where everyone chooses answers at random and receives zero total score. However, there is nothing the nonexperts can do strategically to induce the experts to join them in randomizing their responses. In a "struggle of wills," if the nonexperts randomize and the experts respond truthfully, this will only increase the advantage of the experts in total score. Hence, the experts — who prefer the truthtelling equilibrium — have nothing to fear by answering truthfully.

### ***USING BTS TO ELICIT TRUTHTELLING***

The objective of our first study is two-fold. First, we aim to verify that the truth serum is credible — that survey takers find it believable enough that they respond to incentives based on BTS scores, even when facing competing incentives to deceive. Second, we want to demonstrate that the method does in fact reward truthtelling by assigning higher BTS scores to truthful answers than deceptive ones.

#### ***The Overclaiming Questionnaire***

Evaluating the quality of subjective data presents the obvious challenge of distinguishing truthful responses from untruthful ones. We address this problem by making use of *over-claiming*, the

tendency to claim knowledge about or awareness of non-existent items. Phillips and Clancy (1972) developed an index of over-claiming for use in consumer surveys by asking respondents to rate their familiarity with various consumer goods. None of the goods on the survey actually existed, so any claims of familiarity suggested a predisposition towards exaggerating one's knowledge.

The over-claiming technique provides a useful test of the Bayesian Truth Serum because it provides an objective criterion for measuring truth-telling: if survey takers are completely truthful, they should not claim recognition of any bogus items. We emphasize that by "truthful," we mean not only that respondents honestly report their knowledge, but also that they consider their answers thoughtfully and take care to avoid mistakes. The technique also has a built-in incentive to deceive. Paulhus et al. (2003) showed that individuals with a psychological need for self-enhancement tend to over-claim. We can therefore test whether compensation tied to BTS scores can overcome this tendency.

For our survey we used a subset of the over-claiming questionnaire developed by Paulhus and Bruce (1990), presenting 12 items from each of six categories: arts, historic names, authors and characters, computers and electronics, life sciences, and philosophy. We asked subjects to indicate whether they personally recognized each item, and to estimate the percentage of others taking the questionnaire who would report recognizing it. Although Paulhus and Bruce's original questionnaire asks respondents to indicate their degree of familiarity with each item on a 7-point scale, we used a binary measure to make it easier for subjects to estimate the distribution of responses. Within each category, one-third of the items were non-existent foils. Subjects were not warned that foils were present.

### ***Experimental Design and Predictions***

To test for the effects of BTS-based truth-telling incentives vs. competing incentives, we used a 2 (BTS-based truth-telling incentive: yes, no)  $\times$  2 (explicit deception incentive: yes, no) between subjects experimental design. Subjects assigned to the BTS incentives condition were given the following instructions:

If you answer every question on the questionnaire, you will be eligible for a \$25 bonus. One-third of the participants in your session will receive the bonus. The winners will be the 1/3 of people with the highest “BTS Score,” and will be paid at the end of your session.

BTS Scoring is a method recently invented by an MIT professor, and published in the academic journal *Science*. The method scores responses to surveys and questionnaires, similar to the way tests like the SAT are scored. The difference is that while SAT scoring rewards you for choosing the objectively correct answer, there is no “objectively correct” answer on opinion surveys. Instead, BTS scoring rewards you for answering honestly. Even though there is no way for anyone to know if your answers are truthful — they're your personal opinions and beliefs — your score will be higher on average if you tell the truth.

Because only the top one-third of BTS scorers in your session will receive the bonus, you are most likely to earn the \$25 bonus if you answer as truthfully as you can. By “truthfully,” we mean not only that you are honest, but also that you consider each question thoroughly before deciding your answer, and that you take care to avoid mistakes (such as clicking option A when you meant to click option B).

Subjects in the control (no truth-telling incentives) condition were told:

If you answer every question on the questionnaire, you will be eligible for a \$25 bonus.

One-third of the participants in your session will receive the bonus. The winners will be chosen randomly and paid when the experiment is complete.

Please answer as honestly as you can.

Although we expected that some survey takers would be predisposed to self-enhancement and would therefore be motivated to over-claim, we also wanted to introduce an explicit incentive to deceive in order to clearly compare the impact of competing rewards for honesty and dishonesty. Accordingly, we promised some participants an extra payment in proportion to the number of survey items they claimed to recognize. These respondents were given additional instructions that varied depending on the condition to which they were assigned. For the BTS condition, they were:

We will pay you an extra 10 cents for each item that you recognize. However, remember that your BTS score will be lower if you do not respond truthfully, so you will be more likely to earn the \$25 bonus if you answer as honestly as possible, and only claim recognition of the items you actually recognize.

For the control condition:

We will pay you an extra 10 cents for each item that you recognize. However, we still want you to answer as honestly as possible, and only claim recognition of the items you actually recognize.

The remaining respondents (i.e., those in the condition with no explicit deception incentive) were given no additional instructions.

A total of N=133 people participated in the study, so there were about 33 subjects in each of the four conditions. The study was conducted in four separate sessions with students recruited at Harvard and MIT. The questionnaire was administered by computer, which enabled us to compute the respondents' information scores as they finished the survey, and to pay the subjects according to their experimental conditions at the end of each session as promised.<sup>2</sup> A random number generator was used to select the \$25 bonus recipients in the control condition.

If, based on the instructions provided, participants believe that BTS scoring really rewards truth-telling, we expect to them to claim recognition of fewer foils in the BTS condition than in the control condition. The experimental cell in which respondents faced competing incentives — a BTS truth-telling incentive as well as an explicit deception incentive — provides a stronger test. If belief in BTS scoring is strong, respondents should ignore the offer of 10¢ per item recognized and respond honestly. But if respondents are skeptical about our ability to reward honesty, they will reject the BTS incentives and over-claim their recognition of items on the questionnaire.

## **Results**

**Truth-inducing property of BTS.** We used signal detection analysis (Swets 1964) to analyze whether BTS induced survey takers to answer the questionnaire more truthfully. Accordingly, responses fall into one of four categories: *hits*, or claiming to recognize items that exist (i.e. reals); *misses*, failure to recognize items that exist; *false alarms*, claiming to recognize non-existent items (i.e. foils); and *correct rejections*, claiming to not recognize non-existent items. Table 1 shows the average proportion across subjects of hits and false alarms. It also

shows *accuracy*, the degree of discrimination between real and foil items, which we define as hits minus false alarms; and *bias*, a measure of the general tendency to over-claim, defined as hits plus false alarms.

\*\*\* INSERT TABLE 1 ABOUT HERE \*\*\*

Of greatest interest are the false alarms — the tendency to claim knowledge about items that are not real. An ANOVA test shows a main effect of the truth-telling incentive ( $F_{1,129} = 24.2$ ,  $p < .0001$ ), indicating fewer false alarms in the presence of BTS incentives; and of the deception incentive ( $F_{1,129} = 9.4$ ,  $p < .003$ ), indicating more false alarms when subjects were paid 10¢ per item recognized. The interaction was also significant ( $F_{1,129} = 9.8$ ,  $p < .003$ ): the deception incentive had greater influence in the absence of BTS incentives. Moreover, pooling the two BTS conditions and disregarding the Control+10¢ group, we find that subjects claim to recognize fewer foils under BTS as compared to the control condition absent imposed inducements to over-claim ( $t_{98} = -2.1$ ,  $p < .04$ ). The bias results are similar: an ANOVA test confirms that bias is lower under the BTS truth-telling incentive, higher under the deception incentive; their interaction is again significant.

The results for hits are analogous, but driven entirely by a higher hit rate in the Control+10¢ group than the other groups ( $p < .0001$ ). This is not surprising, as the payment per recognized item creates an incentive to over-claim on reals as well as foils, and of course foils and unfamiliar reals are indistinguishable. Finally, an ANOVA for accuracy shows a main effect of the truth-telling incentive, but not for the deception incentive or the interaction term. Together, these results suggest that respondents were more truthful when facing BTS incentives, even in

the presence of competing incentives to over-claim. Subjects in the BTS condition also spent nearly a minute longer on the questionnaire (7:23 vs. 6:28,  $t_{131} = 2.87, p < .005$ ), suggesting that they responded more carefully and thoughtfully than did those in the control condition.

***Truth-rewarding property of BTS.*** With 12 items in each of six different categories, our questionnaire contained 72 total items. One of the authentic items was mislabeled and subsequently excluded, leaving 47 reals and 24 foils for analysis. We test the prediction that, under conditions when we can reliably identify which responses are “truthful,” that information scores are higher for these responses. Table 2 shows the average information score per item for each response option: *I recognize* or *I do not recognize* the item.

\*\*\* INSERT TABLE 2 ABOUT HERE \*\*\*

The most important results are those for the two BTS conditions, because only under BTS incentives does Prelec’s (2004) theorem predict that truth will necessarily outscore deception; and for foils, because only for these items is there an objectively truthful response. These results, highlighted in the table, show that the truthful response of *don’t recognize* convincingly outscores *recognize* in both conditions where such honesty is incentive-compatible (BTS:  $t_{23} = -7.52, p < .0001$ ; BTS+10¢:  $t_{23} = -6.62, p < .0001$ ). *Don’t recognize* scored higher for 21 of the 24 foils in the BTS condition, and 22 of 24 in the BTS+10¢ condition. These results confirm truthful answers are in fact surprisingly common. Moreover, information scores for the two BTS conditions are very similar; evidently, the presence of a countervailing incentive to over-claim does not significantly distort the truth-rewarding property of information scoring.

The remaining information scores are more difficult to interpret. On real items, recognition mildly outscores non-recognition in the BTS conditions (significantly so only for BTS+10¢), suggesting that respondents know somewhat more about the questionnaire topics than the group expected. There is no truth-telling equilibrium in the control conditions. Overall, however, these broader results indicate that the truth-telling rewards are very decisive, as the difference between scores for don't recognize and recognize is by far the largest in the table's shaded area. The results also show that higher information scores for non-recognition are not simply an artifact of the domain, because the scores in other conditions mostly deviate from this pattern.

### ***Discussion***

These results of Study 1 confirm that compensating people according to their BTS scores creates a compelling incentive to respond more truthfully. Evidently, respondents found our explanation of the scoring system credible, even though the idea of a “truth serum” might reasonably arouse skepticism. The signal detection analysis results are nearly identical for the BTS groups with and without the additional payment per recognized item, suggesting that survey takers had enough faith that veracity would be rewarded that they ignored the financial inducement to exaggerate their knowledge. This assessment of the relative incentives is rational, as expected earnings were higher for those who told the truth. Respondents in the BTS+10¢ group claimed to recognize 30.3 items (virtually the same as in the Control and BTS conditions, so presumably reflecting their actual knowledge), earning \$3.03 per person on average. Had they claimed to recognize all 72 items, they would have earned  $\$7.20 - \$3.03 = \$4.17$  more. However, in expectation their reward for truth-telling was  $1/3$  of \$25, or \$8.33.<sup>3</sup>

It is noteworthy that even in the BTS condition, respondents claimed to recognize 14% of the nonexistent items. This may be a floor effect, and that even when facing incentives survey takers will inevitably make some mistakes, or that larger rewards are necessary to further reduce “recognition” of foils. Another possibility is that these residual false alarms are the result of psychological processes that do not respond to external incentives. Paulhus (1984) distinguishes two types of socially desirable responding: *impression management*, the purposeful manipulation of answers in order to create a positive social image; and *self-deceptive positivity*, which may be unconscious and is used to help maintain self-esteem, optimism, and related personality constructs. Although variations in demand for social desirability (such as an expectation that survey responses will be made public) moderate impression management, they have no effect on self-deception. In fact, people who score high on measures of narcissism appear unfazed even by attempts to confront or embarrass them with evidence of their exaggerated self-presentation. Financial incentives for candor, even when credible and meaningful, may have limited influence on self-deception.

### ***TESTING THE ROBUSTNESS OF BTS SCORING TO EXPLOITATION ATTEMPTS***

#### ***Approach***

The over-claiming study demonstrates not only that survey takers *do* respond to BTS-based incentives, but that they *should* — truthful answers were shown to be surprisingly common and therefore high-scoring. The purpose of Study 2 is to test the robustness of this second finding: is it possible for individuals to “game” the system in such away that they can expect to achieve higher truth serum scores by deviating from the truth, perhaps in some systematic way?

To answer this question, we administered a series of opinion and market research surveys to various sample groups. The content of these surveys was sufficiently benign that we could reasonably take the respondents' actual answers as truthful. Next, for each survey we defined a broad set of deception strategies that an individual might employ in an attempt to exploit the scoring system. We then computed BTS scores for the survey takers' actual responses, and for the responses they would have given if each individual had deceived while everyone else maintained truth-telling. We assess the method's robustness to exploitation attempts by comparing these results — i.e., by testing whether the respondent group, or some identifiable segment thereof, would have attained a higher score by applying some systematic non-truthful policy. Because truthful predictions are well-known to be optimal with a logarithmic scoring rule, we discuss prediction scores no further and confine our attention to verifying the information scoring component of the BTS scoring rule.

Our assumption that in these data sets actual answers are in fact truthful is worth examining. Indeed, although the questions in our surveys do not give respondents strong reasons to deceive, it is likely that some fraction of answers do not correspond to true opinions, because of carelessness or self-impression management. Even if so, however, the logic of testing actual responses against deception policies remains valid. Since we do not know which actual answers are truthful and which are not, superimposing a directed strategy on actual data will bear equally on both types of answers and will only further reduce the fraction of truthful ones. Essentially, our methodology tests whether we can improve information scores by corrupting answers that may already be somewhat corrupted.

### ***Content of the Four Surveys***

We conducted four separate surveys with varied content to test the effectiveness of BTS across multiple domains and respondent pools. All surveys were anonymous and administered with paper and pencil. Brief descriptions of each survey are below, and details summarized in Table 3:

***Personality.*** This survey contained 13 statements from an assessment exercise used by an executive recruitment (“headhunter”) firm to evaluate personality traits of job seekers. The items are “difficult” in the sense that the answer potential employers would prefer is not obvious. One example: *When I’m under a lot of stress I would rather relax by myself than relax with my family.* Respondents indicated whether they personally agreed with each statement, and also estimated the percentage of their peers who would agree. Survey takers were MBA student volunteers who were encouraged to respond truthfully and accurately, and were told that they would receive (anonymous) feedback in the form of a factor analytic interpretation of their answers.

***Faces.*** We showed participants a set of 24 color photographs of young adults (13 women and 11 men). As with the Personality task, they made a binary judgment — in this case, whether they regarded the person in each photo as attractive. Here, however, we extended the survey design in two ways. First, here we computed information scores separately for men and women to account for the possibility that preferences vary systematically by gender. Accordingly, respondents reported their own sex, and gave separate predictions about the percentage of male and female college students who would find each photo attractive. Second, to test the performance of “impersonation” as a deception strategy, we also asked each respondent to consider the tastes of someone else of their choosing, and to rate each photo as if in the shoes and

mind of that other person. To make the impersonation target more concrete, we encouraged subjects to choose someone they know well, and asked them to report his or her first name and gender. The participants were MIT students approached in a student center, who were given a \$2 coupon for a local convenience store as compensation.

***Humor.*** This survey is identical in design to Faces: two answer choices for each item, segmentation by gender, and the elicitation of impersonated responses in addition to the subjects' own opinions. Again participants were recruited at the MIT student center and compensated with a convenience store coupon. In this case, however, we sought a domain with greater ambiguity and variety in tastes: a series of 13 "Deep Thoughts by Jack Handey," which are mock-profound observations that originated as a recurring segment on the television show *Saturday Night Live*. Deep Thoughts can be hilarious, offensive, or utterly obscure, depending on one's taste in humor and counter-cultural sophistication. One example: "If you go to a costume party at your boss's house, wouldn't you think a good costume would be to dress up like the boss's wife? Trust me, it's not." Respondents read the thoughts and judged them as funny or unfunny.

***Purchase Intentions.*** This survey uses content directly relevant to product development: estimates of purchase intent for new products. We presented to participants photographs, descriptions, and retail prices of six novel products from a Sharper Image catalog: a portable exercise cycle, a motorized DVD storage tower, electronic drum sticks, a wearable air purifier, and an automatic eyeglass cleaner. As for Faces and Humor, we segmented the respondents by sex. Moreover, whereas previously the response options were binary, here we offered four choices: definitely will buy, probably will buy, probably will not buy, and definitely will not buy. Respondents, who were MBA student volunteers, indicated their own purchase intent and

the estimated percentage of their peers who would select each category. Because we segmented the sample pool, it was therefore necessary to elicit *eight* separate predictions for each product: four response options  $\times$  two genders.

\*\*\* INSERT TABLE 3 ABOUT HERE \*\*\*

### *Development of Deception Strategies*

The space of possible deception strategies is in principle very large; we evaluated strategies that bear on the following questions:

1. Is the scoring system neutral (as theory predicts) between opinions that are expected to be typical or atypical? Thinking informally, a respondent might be tempted to game the scoring system in one of two ways: either by boosting the numerator of the information score ratio by choosing the most “typical-looking” answers, or by reducing the denominator by choice of unusual, off responses. Neither strategy should pay off.
2. Does the scoring system favor people who believe they are typical or atypical across an entire battery of items? More generally, are there personality traits or demographic characteristics that might justify the use of some deception strategy?
3. Are the penalties for deception “sufficiently large,” and do they extend to false reporting of demographic characteristics?
4. Can the scoring system discriminate between authentic personal answers and answers contrived to mimic the opinions of another individual? In other words, can a person expect to achieve a higher score by pretending to be someone else?

In all we studied twenty different deception strategies, which are described in the Appendix. Some are very simple, such as *always affirm*: rate all items as funny, attractive, etc. depending on the survey content. Others are more complex and depend on one's predictions about others, like *consensus for other sex*: pretend to be a member of the other sex, and give the answer you expect to be in the majority for that sex. In general we obtained deceptive responses not by asking subjects to lie, but rather by "lying on their behalf": transforming their original responses according to algorithms that mimicked each deception strategy. This approach has two advantages. First, it permits us to cast a much wider net — subjects could not have reliably completed their surveys twenty times, adopting a different mindset for each iteration. Second, it eliminates ambiguity about what deception strategy has been employed. The exception to this procedure is the *impersonation* strategy; as explained above, we directed participants on the Faces and Humor surveys to try to mimic the responses of someone well-known to them. Note that not every strategy is applicable to all surveys. For example, the strategies that use gender as an input were not tested for Personality, on which no gender data were collected.

### ***Results and Discussion***

Using equation (1), we computed total information scores for all respondents based on their actual responses, and then on the data sets resulting from strategic deception. We averaged these scores across respondents, and to facilitate comparisons across surveys, we divided these averages by the number of survey questions, yielding average information scores per item. Table 4 summarizes our findings. The results show that BTS scoring is extremely robust to exploitation attempts. Of the 58 total strategies tested across the four surveys, 53 score significantly lower than actual answers, three are not significantly different, and two score

significantly higher. On three of the four surveys, truthful responses outscored every deception strategy we devised.

\*\*\* INSERT TABLE 4 ABOUT HERE \*\*\*

On the Humor survey, however, five strategies were more successful (two significantly so) than the respondents' actual answers. As the table shows, these strategies were *always funny* and various forms of *contrarian*. Why did these policies outperform actual responses? The explanation is straightforward: it turns out that Deep Thoughts are surprisingly funny as a category. For nearly every Deep Thought — 12 out of 13 among both men and women — more people found it funny than the group collectively expected. All of the winning strategies changed many responses from *not funny* to *funny*. These strategies' success, however, should be viewed with some skepticism. When a series of questions tap similar content, as was the case on the Humor survey, it is possible that the same answer will emerge as the high-scoring answer for all, or most, of the items. But to exploit this regularity, respondents would have to know the direction of surprise, and we cannot infer that they are able to do so here. It might have turned out that Deep Thoughts are surprisingly *unfunny*, or as in the Faces survey, which has an identical design, that there is no directional regularity to exploit.

In situations like this, when the questionnaire presents a homogeneous set of items, a respondent's opinion combines two stochastic signals. The first is the overall appeal of Deep Thoughts for that respondent, and does not change across the 13 items. The second signal captures how funny he finds a particular Deep Thought relative to the others presented. This signal is resampled for each item. We can eliminate the effect of the common signal by

constraining strategies to the same number of funny ratings per respondent as in the actual data. For example, a *constrained consensus* strategy for a survey taker who rated five Deep Thoughts as funny would first sort the items by predicted popularity (i.e., by the survey takers' predictions of how many others would find them funny), then assign the rating *funny* to the top five. As expected, this quota system eliminates the advantage of the contrarian strategies. Under such constraints, both *consensus* (average information score +.27,  $p < .001$ ) and *contrarian* (+.20,  $p < .0001$ ) are outscored by actual answers (+.31), and consensus outperforms contrarian, fully replicating the pattern of the other three surveys.

***Can some subgroup of people benefit from deception?*** Respondents on the whole are substantially penalized for deception. But it is possible that averaging across respondents conceals the existence of segments that might profitably deceive. To so profit, people in the relevant segment would need to be able to identify themselves beforehand. We test two segmentation variables: gender and “subjective typicality,” which we will define momentarily.

Conducting the analysis of Table 4 separately for men and women broadly confirms the results for the pooled samples. Eliminating the Personality survey, for which we did not collect gender data, we tested 48 deception strategies across the three remaining surveys. For men, actual responses outscored deception 43 times (40 of which are significantly higher at  $p < .05$  by paired two-tailed *t*-tests), and deception was higher in only 5 cases (2 significant at  $p < .05$ ). The results are similar among women: actual answers scored higher in 44 comparisons (37 significantly so) vs. four cases where deception scored higher (one significantly so). Note that these convincing results among women are despite very small sample sizes on the Faces (12 female respondents) and Humor (14) surveys.

As in the grouped-gender analysis, eight of the nine cases in which a deception strategy outperformed actual responses were policies resulting in rating more Deep Thoughts funny on the Humor survey: *always funny* and variants of *contrarian*. But among both men and women, actual responses again significantly ( $p < .03$ ) outscore all of these approaches when the number of funny ratings is held constant for the *constrained consensus* and *constrained contrarian* strategies. The only other instance of a successful deception strategy was *always won't buy* on the Purchase Intent survey among men, which yielded a trivial information score improvement of .004 per item over actual responses ( $p > .80$ ). In summary, then, it is very unlikely that respondents could expect lying in some gender-tailored way would pay off.

Another plausible way of discriminating individuals who might gain from deception is subjective typicality: the degree to which a respondent expects his judgments to coincide with others'. We measured subjective typicality as the percentage of others that a respondent expects to concur with his answer, averaged across all questions (i.e., it is the average estimated consensus).<sup>4</sup> To test the usefulness of this screening variable, we selected the ten most subjectively typical respondents on each survey and again tested our deception strategies against actual responses. We found little reason to believe that even these selected few can reliably score higher by lying than telling the truth: among these ten people, actual answers outscored deception on 46 out of 58 tests across the four surveys, with deception winning 11 times and one tie. Nine of the cases in which deception won were for the Humor survey, and again can be explained by the main effect of switching answers to the surprisingly common (and hence high-scoring) response *funny*. Even when using the consensus strategy, for which the correlation between subjective typicality and the advantage of deception is highest (about .50), these "top ten" respondents *lose* an average of .02 points per item.<sup>5</sup>

Segmenting on gender or subjectively typicality fails for two reasons. First, the discriminatory power of these segmenting variables is weak. Correlations between typicality and the gains from deception over truth, for example, are generally modest, averaging  $r = .21$  in our data. Second, actual answers strongly outperform deception *in general*. Therefore, even if special individuals could reliably identify themselves, the expected improvement in their information scores would still be little or none, and there is considerable risk of a large penalty. While we cannot rule out the possibility that respondents who might profit from deception have some other criterion for self-identification, it seems that the two segmentation variables available to us — gender and a person's belief about how closely her opinions conform to the group's — are not useful in gaming the system.

***Information score penalty as a function of information loss.*** Some deception strategies appear *ex ante* more likely to be successful than others. For example, trying to anticipate and join the majority opinion seems a more promising approach than responding at random. And a few of the strategies we developed, such as generally trying to be contrarian but retaining answers expected to be very typical, are almost bizarre. Despite these differences in face validity, however, we discovered that all deception strategies are equally bad in the sense that the size of the penalty incurred for lying depends almost entirely on the extent to which truthful responses are changed. The strong linear relationship between the deviation from actual responses and resulting information score is shown in Figure 2. In fact, the correlations are .93, .94, and .98 for the Personality, Faces, and Purchase Intent surveys, respectively. For Humor, this correlation is only .10, reflecting the disruption in scores caused by the fact that nearly all Deep Thoughts are surprisingly funny.

\*\*\* INSERT FIGURE 2 ABOUT HERE \*\*\*

The finding that the reduction in information scores is a linear function of the “information loss” relative to actual responses is striking, particularly when considering the impersonation strategy respondents used on the Faces and Humor surveys. Apparently, we understand our own preferences in a much richer way than those of even well-known others, and our mimicry is somehow one-dimensional in a way that information scoring can easily discern as less than an authentic person.

### ***CONCLUDING DISCUSSION***

This article has had three objectives. The first was to test whether the Bayesian Truth Serum creates legitimate incentives for truth-telling in the context of real market research and opinion surveys. The second was to confirm that survey takers respond to these incentives, even when facing competing incentives to dissemble. The over-claiming questionnaire of Study 1 confirmed that BTS does indeed reward and encourage truthful responses. Finally, we explored whether the BTS system might be exploited by individuals who hope to improve their scores through strategic deception. Study 2 showed that any deviation from the truth, no matter how carefully conceived, is expected to reduce information scores.

We emphasize that someone who supplies a “losing” answer on a BTS-scored questionnaire is not thereby convicted of dishonesty or incompetence. A low score does not automatically indicate deception at the level of an individual answer; rather, BTS scoring penalizes the *intention* to deceive or respond carelessly by linking such intentions with a lower *expected* score.

The situation is analogous to tests like the SAT, except that for aptitude tests the incentives for truthfulness are self-evident, and the “winning” answer is also defined as objectively correct. For both aptitude tests scored against objective knowledge and opinion surveys scored using BTS, one cannot determine whether a false answer reflects the test-taker’s honest opinion or is an instance of deception. What the answer key does ensure is that the test-takers’ incentives are properly aligned: giving an answer that one believes is incorrect always reduces expected score. We now conclude with a discussion of the ability of BTS to reduce different kinds of deviation from the truth, and of some issues of practical importance when applying the method.

### ***Different Types of Untruthfulness***

Broadly, we can divide untrue responses into three categories: intentional deception; carelessness; and inauthentic responding, where, for reasons that may not be fully conscious, a respondent gives answers that are biased by social norms or the opinions of others. It is useful to consider the effect of BTS on these different types of deviations from the truth.

It is clear from our results that the truth serum method *penalizes* all three kinds of untruthfulness, which were represented in Study 2 by various deception strategies. Examples of intentional deception include reversing truthful responses and misrepresenting gender. Carelessness was represented by randomized responses. Randomization also captures the notion of stereotyped responding, where the same answer is given across a block of questions, except that in this case the randomizing “coin-toss” is performed only once, at the start of the block. Inauthentic response strategies included trying to join the majority (or minority), and directed impersonation — policies that mimic biases that might degrade truthfulness. As we have seen, all of these departures from truthfulness reduced information scores, and were equally bad in the

sense that the reduction in score was a linear function of the number of responses changed, without regard for the style or apparent sophistication of the deception policy applied.

Although we also found convincing evidence that BTS *motivates* respondents to tell the truth, whether it reduces all three types of untruthfulness equally well is less clear. The results of Study 1 show that BTS can certainly discourage intentional deception. When participants were paid to exaggerate their knowledge (10¢ per item recognized), they claimed knowledge of two-thirds fewer foil items when also given BTS incentives as when not (3.4 vs. 10.1 foils out of 24 total on the questionnaire). They also claimed to recognize fewer real items. Disentangling carelessness and inauthenticity, however, is harder. Subjects in the control condition with no financial incentive to deceive claimed to recognize 20% of the foils. BTS reduced this rate by about one-third (to 14%). But we do not know if this improvement was due to greater care, a suppression of the (unconscious?) desire for self-enhancement, or some combination of the two. The fact that people in the BTS condition spent more time on the survey, however, suggests that at least some of the improvement results from more careful responding.

### ***Practical Considerations***

To successfully implement BTS, two conditions must be met. First, skeptical respondents must be convinced that the method really can reward the truth. This does not mean that researchers should necessarily explain the specific scoring formula or the intuition behind it (we did not). Indeed, providing these details may be counterproductive, because respondents might be confused by them or falsely conclude they can devise a strategy to beat the system. Our explanation emphasized the reputation of the source (an MIT professor, whose research was published in the journal *Science*) and drew an analogy to scoring for tests of objective

knowledge. Debriefing interviews suggested that respondents found our claim credible, but other instructions may be equally or more effective. Future research can help determine this. Success also requires that BTS-based incentives be large enough to overcome any competing incentives to misrepresent oneself. If the survey questions are sensitive — about drug use or sexual behavior, for example — respondents may prefer to forgo a small financial reward rather than reveal socially stigmatized conduct. It is also possible that at least some sources of inauthenticity are fully unconscious and stubbornly resistant to even large financial rewards for telling the truth.

We have not dealt here with the broader issue of when it is appropriate to introduce performance-based incentives. We take it as uncontroversial that incentives may be useful in some circumstances and counterproductive in others (Camerer and Hogarth 1999). Because scoring transforms a survey of opinions into something that feels like a test of knowledge, it fundamentally changes the relationship between the respondent and survey sponsor. The sponsor, for example, can ask respondents to prepare in advance, such as by trying out a product or service relevant to the questionnaire. In that case, respondents who do their homework have a better chance of doing well in the survey, just as they would on an SAT test. There are also other potential advantages of scoring. Competition creates reputational stakes that can spice up an otherwise dull survey experience; scores can be used to filter more careful or thoughtful respondents, who can then be retained for future studies; scores can also function as performance feedback, teaching respondents how to provide better information. Collectively, the benefits of the Bayesian Truth Serum are substantial and varied enough to make the method a useful tool in many circumstances when a researcher wants to learn about a target group's opinions or beliefs.

## REFERENCES

- Camerer, Colin F., and Robin Hogarth (1999), "The Effects of Financial Incentives in Experiments: A Review and Capital-Labor Production Framework," *Journal of Risk and Uncertainty*, 19, 7-42.
- Cooke, Roger M. (1991), *Experts in Uncertainty: Opinion and Subjective Probability in Science*. New York: Oxford University Press.
- Cremer, Jacques and Richard P. McLean (1988), "Full Extraction of the Surplus in Bayesian and Dominant Strategy Auctions," *Econometrica*, 56, 1247-1257.
- d'Aspremont, Claude and Louis-André Gerard-Varet (1979), "Incentives and Incomplete Information," *Journal of Public Economics*, 11, 25-45.
- Dawes, Robyn M. (1989), "Statistical Criteria for Establishing a Truly False Consensus Effect," *Journal of Experimental Social Psychology*, 25, 1-17.
- (1990), "The Potential Nonfalsity of the False Consensus Effect," in *Insights in Decision Making*, R. Hogarth ed. Chicago: University of Chicago Press, 179-199.
- Engelmann, Dirk and Martin Strobel (2000), "The False Consensus Effect Disappears if Representative Information and Monetary Incentives are Given," *Experimental Economics*, 3 (3), 241-260.
- Johnson, Scott J., John Pratt, and Richard J. Zeckhauser (1990), "Efficiency Despite Mutually Payoff-Relevant Private Information: The Finite Case," *Econometrica*, 58, 873-900.
- Krueger, Joachim and Russell W. Clement (1994), "The Truly False Consensus Effect: An Ineradicable and Egocentric Bias in Social Perception," *Journal of Personality and Social Psychology*, 67, 596-610.
- Marks, Gary and Norman Miller (1987), "Ten Years of Research on the False-Consensus Effect: An Empirical and Theoretical Review," *Psychological Bulletin*, 102 (1), 72-90.
- Mazar, Nina, On Amir, and Dan Ariely (2008), "The Dishonesty of Honest People: A Theory of Self-Concept Maintenance," *Journal of Marketing Research*, 45 (6), 633-644.
- McAfee, R. Preston and Philip Reny (1992), "Correlated Information and Mechanism Design," *Econometrica*, 60 (2), 395-421.
- McLean, Richard and Andrew Postlewaite (2002), "Informational Size and Incentive Compatibility," *Econometrica*, 70 (6), 2421-2454.
- Miller, Nolan H., Paul Resnick, and Richard J. Zeckhauser (2005), "Eliciting Informative Feedback: The Peer-Prediction Method," *Management Science*, 51 (9), 1359-1373.

Paulhus, Delroy L. (1984), "Two-Component Models of Socially Desirable Responding," *Journal of Personality and Social Psychology*, 46, 598-609.

--- and M. Nadine Bruce (1990), "Validation of the OCQ: A Preliminary Study," paper presented at the annual convention of the Canadian Psychological Association, Ottawa, Ontario, Canada.

---, P.D. Harms, M. Nadine Bruce, and Daria C. Lysy (2003), "The Over-Claiming Technique: Measuring Self-Enhancement Independent of Ability," *Journal of Personality and Social Psychology*, 84 (4), 890-904.

Phillips, Derek L and Kevin J. Clancy (1972), "Some Effects of 'Social Desirability' in Survey Studies," *American Journal of Sociology*, 77 (5), 921-940.

Prelec, Drazen (2004), "A Bayesian Truth Serum for Subjective Data," *Science*, 306, 462-466.

Ross, Lee, David Greene, and Pamela House (1977), "The 'False Consensus Effect': An Egocentric Bias in Social Perception and Attributional Processes," *Journal of Experimental Social Psychology*, 13, 279-301.

Swets, John A. (1964). *Signal Detection and Recognition by Human Observers*. New York: Wiley.

## FOOTNOTES

<sup>1</sup> Engelman and Strobel (2000) explored whether the use of own opinions to infer population preferences is egocentric — i.e., whether people weight their own opinions more heavily than others'. They first informed participants of the (factual) preferences of a random subset of peers, then elicited predictions for the population. Although estimates were biased in favor of the sample data (a “consensus effect”), participants’ private signals had no special status and indeed were often underweighted relative to signals from random, anonymous others (no “false consensus effect”).

<sup>2</sup> To simplify the real-time computations, we based each respondent’s score on information scores computed using the responses of the previous sessions’ participants. We used scores computed from a pilot study for the first session. Information scores become more stable as N increases, but in expectation truth-telling is optimal even for very small N.

<sup>3</sup> Interestingly, even in the Control+10¢ group people claimed to recognize only 43 items on average, thus forgoing \$2.89 (= \$7.20 – \$4.31) per person despite the absence of any financial incentive to tell the truth. Similar restraint is explored by Mazar, Amir and Ariely (2008), who argue that an internal motivation to perceive oneself as honest limits the exploitation of rewards for dishonesty.

<sup>4</sup> Consistent with a consensus bias, the mean subjective typicality was .66, i.e. respondents predicted that their peers would agree with two-thirds of their judgments. One person predicted that consensus with his judgments would be 94%!

<sup>5</sup> Incidentally, deception fares even worse among the “bottom ten,” i.e. the most subjectively atypical respondents. For this group, actual answers outscore deception in 52 of 58 cases.

## TABLES

	<b>Control</b>	<b>Control plus 10¢ per item recognized</b>	<b>BTS</b>	<b>BTS plus 10¢ per item recognized</b>
Hits	.58	.71	.57	.57
False alarms	.20	.42	.14	.14
Accuracy	.38	.29	.43	.43
Bias	.79	1.12	.71	.72

Table 1: Signal detection analysis results for the over-claiming questionnaire.

	<b>Control</b>	<b>Control plus 10¢ per item recognized</b>	<b>BTS</b>	<b>BTS plus 10¢ per item recognized</b>
<b>Reals</b>				
I do recognize	+.28	+.38	+.16	+.22
I don't recognize	-.11	-.28	+.08	-.02
<b>Foils</b>				
I do recognize	-.70	+.23	-.99	-.93
I don't recognize	+.21	+.08	+.34	+.27

Table 2: Average information score per survey item for the over-claiming questionnaire, for reals and foils, by experimental condition.

<b>Survey Content</b>	<b>N</b>	<b># Items</b>	<b>Demographic Split</b>	<b>Answer Choices</b>	<b># Deception Strategies Tested</b>
Personality	104	13	n/a	agree disagree	10
Faces	41	24	gender	attractive not attractive	18
Humor	46	13	gender	funny not funny	18
Purchase Intent	106	6	gender	definitely will buy probably will buy probably will not buy definitely will not buy	12

Table 3: Summary of the design of the four surveys conducted for Study 2. Sample sizes reflect the exclusion of participants who did not report their gender or failed to complete large portions of the survey: 16 for Faces, 11 for Humor, and 3 for Purchase Intent.

Strategy	Personality	Faces	Humor	Purchase Intent
Actual answers	.17	.19	.31	.42
Always affirm (agree, attractive, funny, will buy)	.04 ****	-.05 ****	<b>.60</b> ****	-.70 ****
Always reject (disagree, not attr, not funny, won't buy)	.11 ****	.03 ****	-.19 ****	.38 **
Reverse answers	-.02 ****	-.21 ****	.10 ***	-.79 ****
Claim to be a member of the other sex		.09 ****	.25 **	.36 ***
Reverse answers and claim to be other sex		-.20 ****	.13 **	-.75 ****
Consensus: try to be in the majority	.13 ****	.13 ****	.17 ****	.31 ****
Contrarian: try to be in the minority	.03 ****	-.12 ****	<b>.31</b>	-.89 ****
Mild consensus: try to ensure mild agreement	.16 *	.17 ***	.25 ***	
Mild contrarian: try to ensure mild disagreement	.10 ****	-.05 ****	<b>.39</b>	
Mostly consensus, but retain very atypical answers	.12 ****			
Mostly contrarian, but retain very typical answers	.11 ****			
Consensus for other sex (and claim to be that sex)		.10 ****	.08 ****	.22 ****
Contrarian for other sex		-.18 ****	<b>.36</b>	-.89 ****
Mild consensus for other sex		.09 ****	.19 ****	
Mild contrarian for other sex		-.14 ****	<b>.39</b> *	
Consensus for own sex but claim to be other sex		.10 ****	.14 ****	.31 ****
Consensus for other sex but truthfully report own sex		.14 ***	.11 ****	.34 ***
Impersonate a well-known other		.13 ****	.20 **	
Counter-impersonate: reverse impersonated answers		-.17 ****	.20 *	
Answer randomly	.07 ****	.01 ****	.20 **	-.15 ****

**Table 4:** Average information score per survey item across all respondents in Study 2, for respondents' actual answers and the data sets resulting from various deception strategies.

Asterisks indicate results of two-tailed paired t-tests comparing scores under truth (actual answers) to deception: \*  $p < .05$ , \*\*  $p < .01$ , \*\*\*  $p < .001$ , and \*\*\*\*  $p < .0001$ .

## FIGURES

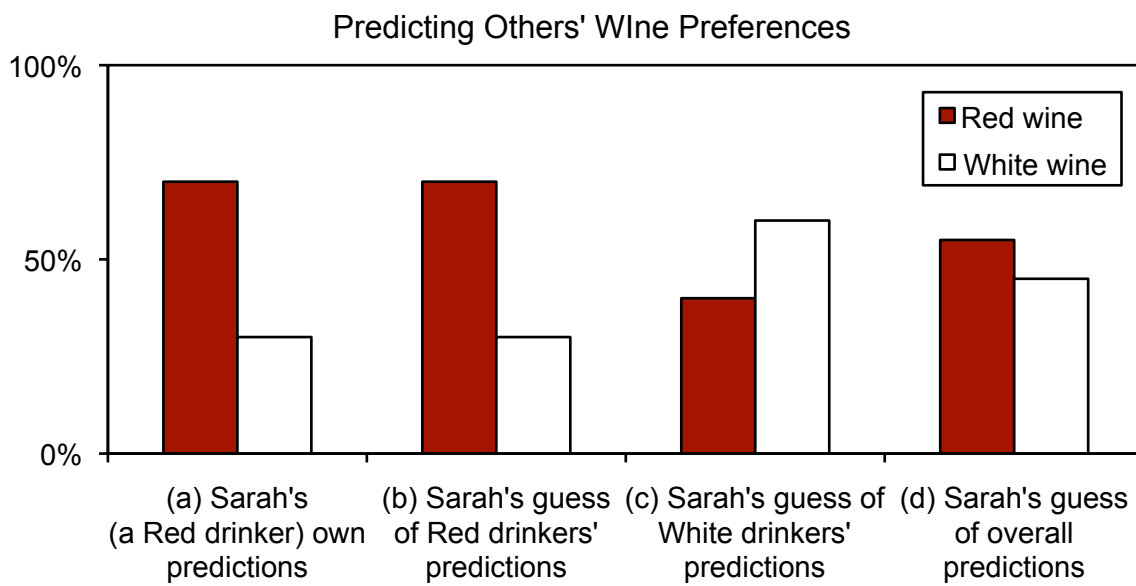
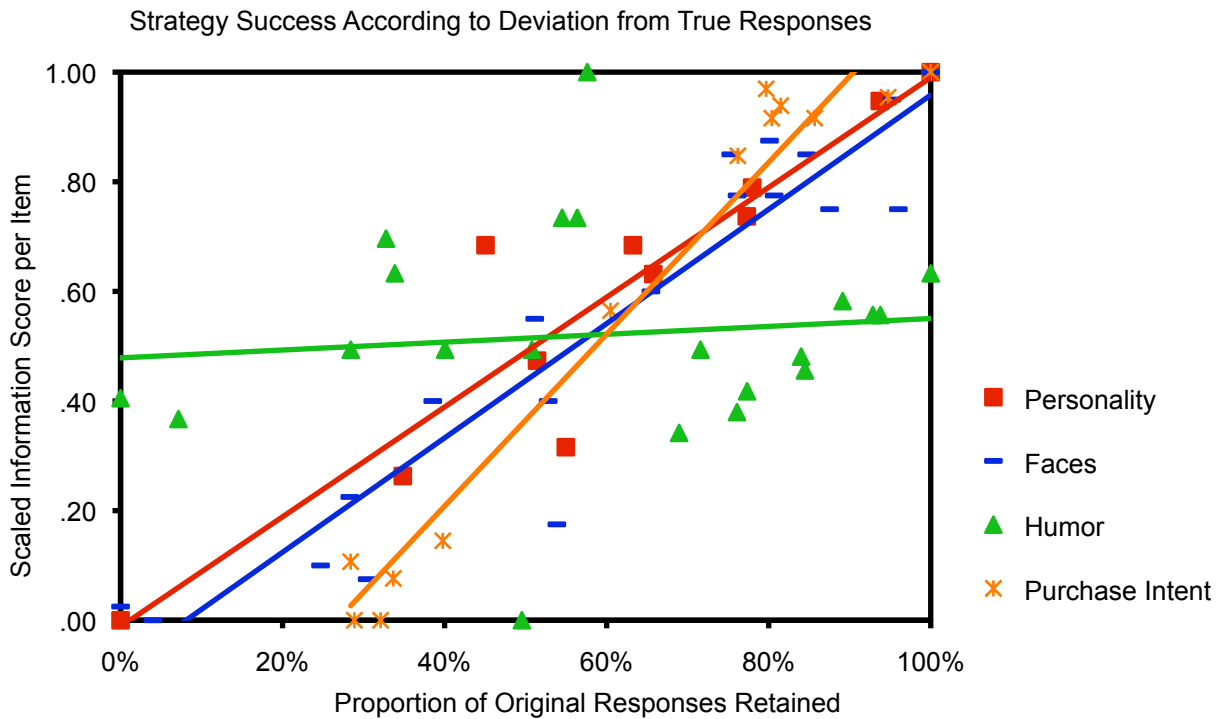


Figure 1: Truthful answers are more likely to be surprisingly common because each person's own preference is a (rational) signal about preferences in general.



**Figure 2:** For each of the four surveys in Study 2, this figure plots information scores under actual responses and deception strategies as a function of the proportion of original responses retained under each strategy. Within each survey, information scores are scaled such that the best-performing strategy (usually actual responses) is 1, and the worst is 0.

## APPENDIX

The deception strategies used to test the ability of information scoring to reward truth-telling in the four surveys of Study 2 are described below. Note that in all cases, we only transformed the survey takers' judgments about the statements, not their estimates of the fraction of people endorsing each answer — with our sample sizes, changing a single respondent's predictions would have a negligible impact on overall scores.

Always affirm: Affirmative response for all items: *agree* (Personality), *attractive* (Faces), *funny* (Humor), and *will buy* (Purchase Intent). Because Purchase Intent has two assenting responses — *probably will buy* and *definitely will buy* — we took the average of the information scores for these two answers.

Always reject: Negative response for all items: *disagree*, *not attractive*, *not funny*, and *will not buy*. Again, we averaged Purchase Intent's two dissenting responses.

Reverse answers: The opposite of each actual answer. For Purchase Intent, which has four possible responses, we changed actual answers to their “mirror image,” e.g. *probably will buy* became *probably will not buy*.

Claim to be a member of the other sex: Retain actual judgments, but apply information scores for the subject's opposite sex.

Reverse answers and claim to be other sex: Apply both of the previous two strategies.

Consensus: Change actual answers to the response the survey taker expects to be in the majority (or, for Purchase Intent, the expected mode) based on his predictions. For the three binary response choice surveys, ties — predictions that exactly half the group would give each response — were handled by retaining the original response. For Purchase Intent ties (multimodal predictions), we averaged the information scores of the modes.

Contrarian: Changed actual answers to the expected minority response. Ties were dealt with in the same way as for consensus.

Mild consensus: Similar to *consensus*, but only change actual answers when the survey taker predicts that less than 30% of others will agree with him. For example, if a person's actual response is *funny* and prediction is that 20% of others will agree, his response becomes *not funny*.

Mild contrarian: Similar to *contrarian*, but only change actual answers when the survey taker predicts that more than 70% of others will agree with him.

Mostly consensus, but retain very atypical answers: If the survey taker expects 30-50% of others to agree with him, switch to the expected majority response. Otherwise, keep original response.

Mostly contrarian, but retain very typical answers: If the survey taker expects 50-70% of others to agree with him, switch to the expected minority response. Otherwise, keep original response.

Consensus for other sex: Switch the survey taker's stated sex, and change to the responses he expects to be in the majority for that sex.

Contrarian for other sex: Switch the survey taker's stated sex, and change to the responses he expects to be in the minority for that sex.

Mild consensus for other sex: Switch the survey taker's stated sex, and change to the expected majority response for that sex when predicting less than 30% agreement.

Mild contrarian for other sex: Switch the survey taker's stated sex, and change to the expected minority response for that sex when predicting more than 70% agreement.

Consensus for own sex but claim to be other sex: Apply both *consensus* and *claim to be a member of the other sex*.

Consensus for other sex but claim own sex: Switch answers to those expected to be in the majority for the other sex, but retain the survey taker's true sex.

Impersonate another: Use the reported sex and responses that the survey taker gave when impersonating the preferences of some other specific person he knows well. Unlike the other strategies, which represent deception by rule-based transformations of truthful responses, impersonation is actual directed deception elicited from the survey takers.

Counter-impersonate: Reverse all responses, including the reported sex, that the survey taker gave when impersonating a well-known other.

Random: Retain actual sex, but generate random responses from a uniform distribution.