RAY WEAVER and DRAZEN PRELEC*

The Bayesian truth serum (BTS) is a survey scoring method that creates truth-telling incentives for respondents answering multiple-choice questions about intrinsically private matters, such as opinions, tastes, and behavior. The authors test BTS in several studies, primarily using recognition questionnaires that present items such as brand names and scientific terms. One-third of the items were nonexistent foils. The BTS mechanism, which mathematically rewards "surprisingly common" answers, both rewarded truth telling, by heavily penalizing foil recognition, and induced truth telling, in that participants who were paid according to their BTS scores claimed to recognize fewer foils than control groups, even when given competing incentives to exaggerate. Survey takers who received BTS-based payments without explanation became less likely to recognize foils as they progressed through the survey, suggesting that they learned to respond to BTS incentives despite the absence of guidance. The mechanism also outperformed the solemn oath, a competing truth-inducement mechanism. Finally, when applied to judgments about contributing to a public good, BTS eliminated the bias common in contingent valuation elicitations.

Keywords: truth-telling incentives, survey design, contingent valuation method, Bayesian inference, false consensus effect

Creating Truth-Telling Incentives with the Bayesian Truth Serum

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—the "correct" answer for a particular person is simply the one that best matches his or her private opinions or preferences, and the survey administrator is in no position to judge whether any given answer faithfully reflects these. To evaluate answers, the administrator would apparently need to know which answer is personally correct for each respondent; however, such an omniscient administrator would not need to conduct a survey in the first place.

Prelec (2004) proposes 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. Thus, the system rewards responses that are "surprisingly common" and penalizes those that are "surprisingly uncommon." The BTS theorem states that under certain conditions, personally truthful answers maximize expected scores. By compensating survey takers according to these scores, we can potentially create financial incentives for truth telling, even when the sur-

*Ray Weaver is Assistant Professor of Business Administration, Harvard Business School, Harvard University (e-mail: rweaver@hbs.edu). Drazen Prelec is Professor of Management, Management Science and Economics, MIT Sloan School of Management, Massachusetts Institute of Technology (e-mail: dprelec@mit.edu). This research was supported by NSF SES-0519141 and the John Simon Guggenheim Memorial Fellowship (to the second author) and MIT's Center for Innovation in Product Development (to the first author). The authors thank their colleagues Shane Frederick and John Hauser for discussion and comments. The hospitality of the Institute for Advanced Study is gratefully acknowledged. Teck Ho served as associate editor for this article.

ISSN: 0022-2437 (print), 1547-7193 (electronic)

Journal of Marketing Research
Vol. L (June 2013), 289–302

vey administrator knows neither the respondent's true responses nor the expected aggregate responses of the surveyed population.

However, the mechanism's effectiveness in actual application remains an open question, notwithstanding some initial encouraging results (Barrage and Lee 2010; John, Loewenstein, and Prelec 2012). The truth-telling theorem requires assumptions that are not likely to be satisfied in any actual data set. Moreover, even if information scores (hereinafter, "iscore") reward truthfulness, they may not necessarily *induce* it, particularly when survey takers face competing incentives to dissemble. Moreover, the extent of any improvement in data quality under BTS has not been weighed against added complexity imposed by the method or against alternative truth-inducing methods.

The objective of this article is to address these questions. In four studies, we applied BTS to recognition questionnaires containing items such as brand names and scientific terms. One-third of these were nonexistent "foils" that were interspersed among the legitimate items. The rate of claimed recognition of these foils is a plausible measure of deception across experimental treatments. We found that, in general, the BTS mechanism robustly penalized claimed recognition of foils (even though the classification of items into foils and "reals" was not an input to the mechanism). Moreover, participants who were paid according to BTS scores responded to these incentives by recognizing fewer foils (but similar numbers of real items) than control groups, even when they were offered competing payments to exaggerate. When we implemented BTS-based payoffs without explanation, survey takers became less likely to recognize foils over the course of the survey, suggesting that they learned to respond to BTS incentives despite the absence of guidance. In addition, BTS outperformed a rival truth-inducement mechanism, proposed by Jacquemet et al. (2009), that employs a form of cheap talk in which participants are asked to sign a "solemn oath" before beginning the survey. Finally, in an application to judgments about contributing to a public good, BTS eliminated the bias common in contingent valuation elicitations. Collectively, these results suggest that BTS is an effective and practical way to induce truth telling in many settings, including cases in which survey respondents have nontrivial implicit or explicit incentives to stray from the truth.

INFORMATION SCORING AND EXPERIMENTAL APPROACH

The Scoring Mechanism

The BTS works at the level of an individual multiple-choice survey question, assigning a numerical score to each possible response. To implement BTS, the administrator elicits from each survey respondent not only his or her personal answer but also his or her estimates in percentage terms of how others will respond. The iscore formula is as follows:

$$\text{(1)}\qquad \text{Answer's iscore} = \log \frac{\text{answer's actual relative frequency}}{\text{(geometric) mean of answer's predicted frequency}}$$

The scoring system transforms a survey into a competitive, zero-sum contest in which truth telling is a strict Bayesian Nash equilibrium. Specifically, personally truthful answers maximize the expected iscore for any respondent who believes that others are giving truthful answers and providing perfect Bayesian predictions of the distribution of answers (Prelec 2004).[1] Like other Bayesian mechanisms, BTS exploits the subjective correlation between a person'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). Unlike previous mechanisms, however, BTS does not incorporate assumptions about this correlation into the scoring function. It can be contrasted with consensus scoring, which assigns a high score to the most popular answer, thereby creating deception incentives among respondents who suspect that their opinion is in the minority. Information scoring creates no such incentives: Untruthful answers have lower expected scores, regardless of whether the respondent believes that his or her opinion is common or rare.

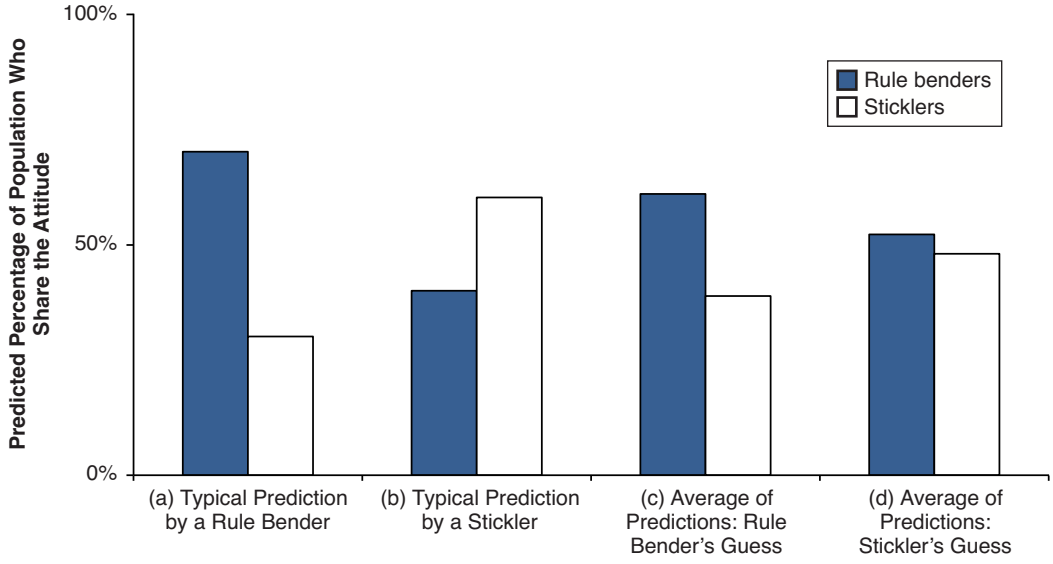
Intuition Behind Information Scoring

Mathematically, iscoring is straightforward: Answers that are more common than the group collectively predicts receive high scores, and answers that are less common than predicted receive low scores. It is less obvious, however, why the most surprisingly common response should also be the personally truthful one. A hypothetical example can provide some intuition.

Suppose a survey group is asked, "Have you ever exaggerated the deductions on your income tax return?" Each respondent is instructed to privately answer yes or no and to guess the percentage of people who will give each response. To understand the relationships between personal answers and predictions, it is useful to consider how predictions will differ between people whose true answer is yes ("rule benders") and those whose true answer is no ("sticklers"). Figure 1 shows predictions that our hypothetical survey takers might make. Although it is not essential to the argument, the figure assumes that both rule benders and sticklers estimate that they are in the majority (first two columns). What is essential, however, is that rule benders believe rule bending to be more common than sticklers believe that it is and that this difference in predictions is itself anticipated by both categories of subjects. In that case, the rule bender's guess of the predictions averaged across all respondents (third column) will fall somewhere between his or her own predictions (first column) and estimate of the stickler's predictions (second column). Because the rule bender's own predictions are, by definition, the best predictions he or she can make, the rule bender will conclude that the average predictions (third column) will be directionally incorrect—that is, the average will underestimate the percentage of rule benders and overestimate the percentage of sticklers (first column), and thus that his honest answer "yes" ("I have exaggerated my income tax deductions") will prove surprisingly common. A stickler, in contrast, should expect "no"—again, the

[1]In addition to scoring answers with iscores, a full implementation of BTS technically requires that predictions be scored as well. It is a standard result that truthful predictions are optimal with a log rule (Cooke 1991), and we do not further discuss prediction scoring here.

Figure 1
HYPOTHETICAL EXAMPLE SHOWING THE RELATIONSHIP BETWEEN PERSONAL ATTITUDES AND ESTIMATES OF OTHERS' ATTITUDES



honest response—to be surprisingly common, because his or her own estimate of sticklers' numbers (second column) is higher than the guess of the average prediction of that number (fourth column). In other words, every person's true answer is subjectively likely to be more common than collectively predicted. Therefore, rewarding surprisingly common answers is tantamount to rewarding truthful answers.[2]

The critical assumption here is that respondents will use their own opinions as evidence about the distribution of opinions in the population. Therefore, the practicality of the method hinges on a testable empirical proposition: that there does indeed exist a positive relationship between individual opinions and population estimates. It must be true empirically that rule benders will give relatively higher estimates of the frequency of rule bending. There is a great deal of experimental evidence in support of this relationship, across many domains. Ross, Greene, and House (1977) performed the seminal experiment by asking students whether they would be willing to wear a "Repent" sign around their campus. They observed that students who were personally willing to wear the sign were also likely to give higher estimates of how many others would also be willing to do so. This finding was initially described as a "false consensus"—evidence of an egocentric assumption that others are necessarily similar to oneself. Notably, although the result was replicated in dozens of studies (Marks and Miller 1987), it took some time before the normative status of the finding was properly addressed. Dawes (1989, 1990), in particular, argues strongly for a Bayesian interpretation of the phenomenon as a rational updating on the basis of a "sample of one."[3] The theoretical support for the present method follows Dawes' Bayesian interpretation: Personal preferences and opinions effectively act as private signals that inform survey takers about responses that are most likely to be surprisingly common and thus most likely to receive high iscores.

It may seem counterintuitive that BTS can create truth-telling incentives if different answers are truthful for different people. However, consider a test of objective fact such as the SAT. If two test-takers arrive at different answers to a question, they cannot both be right, but each is best off reporting the answer he or she *believes* to be correct. Similarly, although iscoring makes some answers winners and others losers, every respondent has reason to believe that the answer matching his or her own private opinion is *ex ante* most likely to receive the highest score. And just as a wrong answer on the SAT does not imply that the student intended to err, a low iscore does not necessarily indicate a dishonest response.

Assumptions of the Bayesian Truth Serum

The BTS theorem makes several assumptions that have not been tested in practice. In particular, it models people as Bayesian statisticians when constructing predictions about the distribution of responses to a question: They begin with a common prior belief and then update this belief in the direction of their own preference according to Bayes' rule.

[2]Although both types in our example think they are in the majority, we emphasize that truthful responses are surprisingly common in expectation even for respondents who are, or think they are, in the minority.

[3]Some debate remains about whether inferences from such samples are efficient. Krueger and Clement (1994) argue that people hold an egocentric bias in favor of personal preferences. Engelmann and Strobel (2000) disagree. They first informed participants of the (factual) preferences of a random subset of peers and then elicited population predictions. 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 compared with signals from random, anonymous others (no "false consensus effect").

Strictly speaking, this assumption implies that all people with the same preference will have the same posterior distribution (i.e., make the same frequency predictions). The theorem also assumes that the sample is large enough that no individual response or prediction has a meaningful impact on the sample frequencies.

There is reason to be skeptical that these assumptions will hold in an actual survey setting. First, the idea of a "prior" belief, uninformed by a person's own preference, probably has little psychological reality. Moreover, the predictions of like-minded respondents vary considerably in practice. Finally, although truth telling is an equilibrium of the game, there are also other equilibria. In light of these potential pitfalls, the experiments that follow are intended to evaluate the effectiveness and practicality of BTS for use in marketing research, opinion polling, and other surveys. Specifically, we explore the following:

- Given the sophistication required for Bayesian reasoning and the noise inherent in survey data, does information scoring successfully reward truthful answers and penalize untruthful ones?
- If so, do survey takers respond to financial incentives based on these scores? How much explanation and coaching about the system do survey takers need?
- Is the method robust to conditions that might undermine the truth-telling equilibrium, such as small sample sizes or competing incentives to deceive?
- Does BTS improve the quality of survey data enough to justify its use?
- How does BTS perform relative to alternative truth-inducement mechanisms?

The Overclaiming Questionnaire

Evaluating sincerity in matters of private opinions or other subjective topics presents the challenge of distinguishing truthful responses from untruthful ones. To address this problem, we borrowed a method first used by Phillips and Clancy (1972), who indexed consumer overstatement of brand awareness with a survey consisting entirely of nonexistent goods. More recently, Paulhus and Bruce (1990) developed the overclaiming questionnaire, a survey of familiarity containing items from various categories in which legitimate items are interspersed with a minority of nonexistent foils. The surveys in our Experiments 1–4 are based on the overclaiming questionnaire, with foils making up one-third of total items. (Participants were not warned that the surveys contained foils.) We replaced the typical seven-point familiarity scale with a binary choice: Respondents indicated that they either do or do not recognize the item. This change simplifies the prediction task necessary for information scoring because respondents must estimate the percentage of answers that will fall into each response category.

The proportion of foils that a survey taker claims to recognize provides a measure of untruthfulness that we use to evaluate the effect of BTS on survey data. Importantly, "untruthfulness" means not only intentional deception but also departures from true answers caused by carelessness, inattentiveness, or lack of introspection. The overclaiming technique also has a built-in incentive to deceive: Paulhus et al. (2003) shows that people with a psychological need for self-enhancement tend to exaggerate their knowledge. In some experimental trials, we supplement this implicit incentive with explicit financial incentives to overclaim.

EXPERIMENT 1

Many people might reasonably be skeptical of a mysterious formula that scores survey responses in such a way as to reward honesty and penalize deception. The main objectives of Experiment 1 were to determine if survey takers regard this claim as plausible and if iscores indeed create the desired incentives. We studied these questions by administering a version of the overclaiming exercise we described as a "general knowledge questionnaire." By varying respondents' incentives across experimental groups, we could assess whether the promise of financial rewards for truth telling was credible.

Design and Procedure

We created a computer-administered survey based on Paulhus and Bruce's (1990) overclaiming questionnaire, selecting 72 items (of which 24 were foils) from six categories: arts, historic names, authors and characters, computers and electronics, life sciences, and philosophy. These items were presented by category, 12 per screen. Participants reported whether they personally did or did not recognize each item and estimated—on an 11-point scale in increments of ten percentage points—the proportion of survey takers who would claim recognition of that item. The survey itself was identical for all participants, but the instructions and end-of-survey procedures varied according to the financial incentives present in a 2 × 2 between-subjects design: truth-telling incentives (BTS or control) and deception incentives (overclaiming or none). We conducted the experiment in laboratories on two university campuses.

Before completing the questionnaire, participants in the BTS treatment received a brief nontechnical introduction to the concept. We explained that "BTS scoring" assigns scores to survey answers in a way that rewards honesty and that although we cannot know true private opinions or beliefs, and questions about such matters have no objectively correct answer, respondents nevertheless score higher on average by telling the truth. To lend credence to these claims, we reported that the method was "recently invented by an MIT professor, and published in the academic journal *Science*." We informed participants that the top one-third of people with the highest total scores would earn $25 each. In the control treatment, we explained that one-third of participants chosen at random would receive $25 and asked this group to "please answer as honestly as you can." To create deception incentives, the overclaiming treatment included an additional instruction: "We will pay you an extra 10 cents for each item that you recognize." However, we admonished the recipients of this incentive not to exaggerate their knowledge: People in the BTS treatment were reminded that untruthfulness would jeopardize their chance at the $25 bonus, and people in the control treatment were again asked to respond honestly.

In the BTS treatment, a summary screen at the conclusion of the survey reported the participant's total iscore across all 72 items. To simplify the real-time computations, we based these scores on data from the previous sessions' participants; iscores for the first session were based on a pilot study.[4] Control participants' summary showed a random

[4]Information scores become more stable as the number of survey takers increases, but truth telling is optimal in expectation even for very small N.

number. At the end of the session, we distributed $25 payments and item recognition payments according to the rules we had specified.

Results and Discussion

A total of 133 people completed the questionnaire. One of the authentic items was mislabeled on the survey and subsequently excluded, leaving 47 reals and 24 foils for analysis. We turn first to item recognition rates. Table 1, Panel A, presents the average percentages of items participants recognized in each experimental condition. Unsurprisingly, overclaiming incentives alone led to higher recognition of both item types than that found in the control group. In contrast, participants in the BTS treatment—regardless of whether they were also given overclaiming incentives—recognized roughly the same number of reals as, and fewer foils than, control participants. Regressing each participant's recognition percentages for each item type on the experimental treatments, with item type as a within-subject effect, confirms that recognition rates were lower under BTS incentives and for foils and higher under overclaiming incentives. There are also significant interaction effects between BTS and overclaiming incentives, indicating that the truth serum had a greater effect in the presence of competing incentives, and between BTS incentives and foils, indicating that the truth serum reduced recognition of bogus items more than it did legitimate ones (Table 1, Panel B). Survey takers in the BTS treatment also spent more time on the questionnaire (7:23 vs. 6:28; t(131) = 2.87, $p < .005$), suggesting that they were more careful than other participants.

Using the collective data from all experimental sessions, we then recomputed the average iscores for reals and foils in each experimental group (Table 2). The most important scores are those for the BTS treatment. In both BTS conditions, recognition scored slightly higher than nonrecognition among real items (significantly so only for the BTS + overclaiming group). However, among foils, nonrecognition outscored recognition by a wide margin (BTS only: t(23) = −7.52, $p < .0001$; BTS + overclaiming: t(23) = −6.62, $p < .0001$). Recognition was penalized for 21 of the 24 foils in the BTS cell and 22 of 24 foils in the BTS + overclaiming cell. This pattern seems appropriate for creating truth-telling incentives: iscores give a small reward for recognizing legitimate items and a large penalty for recognizing bogus ones. The introduction of competing incentives to exaggerate does not seem to significantly distort these incentives.

The scores for the control and overclaiming only conditions are less meaningful: No information about or incentives based on iscores were given, so we do not necessarily expect a truth-telling equilibrium to emerge. However, two points are worth noting. First, the penalties for recognizing fake items are decisive in the BTS conditions, larger than for the control group. Second, the pattern of iscores that emerged for the BTS treatment is by no means guaranteed: in the pure overclaiming condition, recognizing foils is rewarded, not penalized, on the whole. Why foil recognition was surprisingly common in this treatment—a pattern we also observed in our other experiments under similar incentives—is unclear, particularly considering the group's (modest) restraint in overclaiming. They recognized an average of 43.1 items, forgoing $2.89 (28.9 items at 10¢ each) in recognition payments. Mazar, Amir, and Ariely (2008) find similar forbearance and argue that an internal motivation to perceive oneself as honest limits the exploitation of rewards for dishonesty, even as the risk of exposure declines. It is possible, then, that participants in our overclaiming condition convinced themselves that they were not exaggerating and thus failed to appreciate the extent to which others would do so.

Experiment 1 shows that BTS can both encourage and reward more truthful survey responses. Moreover, the presence of competing deception incentives undermined neither the system's truth-inducing property, in that item recognition rates were similar between the BTS and BTS + overclaiming groups, nor its truth-rewarding property, in that iscores—and in particular, the steep penalty for recognizing foils—were similar for the two groups. Note that we purposefully set payment levels such that rational actors who found the truth-telling incentives credible would favor them over competing inducements. Respondents in the BTS + overclaiming group claimed to recognize 30.3 items on average (virtually the same as in the control and BTS conditions, so presumably reflecting their actual knowledge), earning $3.03 for recognition. Had they claimed to recognize all 72 items, they would have earned $7.20 – $3.03 =

Table 1

EFFECTS OF ITEM TYPE AND INCENTIVES ON REPORTED RECOGNITION (EXPERIMENT 1)

A: Mean Recognition Rate by Treatment

	Control	*Overclaiming Only*	*BTS Only*	*BTS + Overclaiming*
Reals	58%	71%	57%	57%
Foils	20%	42%	14%	14%

B: Regression of Participants' Recognition Rates on Item Type and Incentive Treatments

Variable	*Coefficient*	*Robust SE*	*t*	p > |*t*|
BTS incentives	−.120	.029	−4.20	.000
Overclaiming incentives	.085	.029	2.98	.003
Foil	−.382	.017	−23.12	.000
BTS incentives × overclaiming incentives	−.167	.057	−2.92	.004
BTS incentives × foil	−.095	.033	−2.88	.005
Overclaiming incentives × foil	.038	.033	1.16	.246
BTS × overclaiming × foil	−.093	.066	−1.41	.162
Intercept	.417	.014	29.16	.000

Notes: $R^2 = .56$, $F(7, 132) = 127.63$, $p < .0001$.

Table 2

ISCORES UNDER VARYING INCENTIVES (EXPERIMENT 1)

	Control	*Overclaiming Only*	*BTS Only*	*BTS + Overclaiming*
Reals				
Recognize	+.28	+.38	+.16	+.22
Do not recognize	−.11	−.28	+.08	−.02
Foils				
Recognize	−.70	+.23	−.99	−.93
Do not recognize	+.21	+.08	+.34	+.27

$4.17 more. However, the expected reward from truth telling was nearly twice as large: one-third of $25, or $8.33.

EXPERIMENT 2

Experiment 1 suggests that participants found the concept of an honesty-rewarding scoring system plausible and responded accordingly. However, because participants were given no information about actual iscores during the survey, it is unclear whether they were influenced by the *actual* truth-telling incentives or by our *claims* of such incentives, which can be regarded as a form of cheap talk. In Experiment 2, we attempt to distinguish these hypotheses. Specifically, we tied truth-telling incentives directly to each item's iscores and gave immediate feedback about payments for recognizing and not recognizing each item. In addition, to investigate changes in behavior as people progressed through their surveys, we randomized the order in which items were presented. Finally, we calculated iscores in real time to examine trends in scores as the sample size increased.

Design and Procedure

As in Experiment 1, we varied incentives between subjects using a 2 (truth telling: BTS or control) × 2 (deception: overclaiming or none) design. Participants assigned to the BTS treatment received a payment of 1.5 times their iscore; participants in the control group received a flat 25¢ per item. Those in the overclaiming treatment were paid an additional 25¢ for each item they reported recognizing. Appendix A reproduces our instructions.

The knowledge questionnaire for this experiment contained 60 items—40 reals and 20 foils—from five categories. Some come from Paulhus and Bruce's (1990) questionnaire, and others we developed specifically for this study (for the full list of items, see Appendix B; we used the identical set in Experiments 3 and 4). These items were presented in random order for each participant independent of category, though the relevant category was always announced alongside each item. We again elicited from participants both personal judgments and sample group predictions. After a participant submitted responses for an item, we showed a feedback screen reporting his or her earnings. In the BTS conditions, we showed payments for both recognition and nonrecognition and confirmed the respondent's actual earnings according to his or her answer. These payments were based on iscores calculated using all prior responses on the relevant item from participants in the same experimental condition. To have some basis for computing iscores (and payments) for the first survey takers, we seeded each experimental group with the responses of ten people who completed the questionnaire before the study. These seeds were volunteers who were not given financial incentives for participation. The survey was administered to members of an online panel of university undergraduate students (N = 117) who were randomly assigned to one of the four experimental conditions. They took the survey remotely and were paid afterward with Amazon.com gift certificates.

Item Recognition and Payments Results

The pattern of recognition rates is similar to that found in Experiment 1 (see Table 3, Panel A). Overclaiming incentives in isolation resulted in much higher recognition rates compared with the control group (recall that the incentives were stronger here: 25¢ per item recognized vs. 10¢ in Experiment 1). Participants in the BTS condition, however, recognized slightly fewer reals (54% vs. 58%; $\chi^2(1) = 3.45$, $p = .06$) and substantially fewer foils (13% vs. 24%; $\chi^2(1) = 23.15$, $p < .001$) than participants in the control group. Again, the results for the combination of truth-telling and deception incentives were similar to those for truth-telling incentives alone. We ran a three-way full-factorial logit model of recognition judgment (nonrecognition = 0, recognition = 1) on BTS incentives, overclaiming incentives, and item type as a within-person effect, clustering standard errors by participant (Table 3, Panel B). All main effects and two-way interactions were significant except overclaiming × item type, which was marginally significant ($p = .06$). This analysis shows that the truth serum lowers recognition likelihood overall, overclaiming incentives increase recognition, and foils have lower recognition rates. Furthermore, BTS acts more strongly (i.e. lowers recognition likelihood more) on foils than on reals and acts more strongly in the presence of competing incentives to overclaim.[5]

This improvement in truth telling in the BTS group is rational given the payments for recognition and nonrecognition that emerged (Figure 2). Payments for the BTS-only condition (equal to 1.5 times the underlying iscores) are analogous to the iscores in Experiment 1: a small reward for recognizing reals and a large penalty for recognizing foils, appropriate for encouraging truth telling. In the BTS + overclaiming condition, the reward for recognizing reals is

Table 3
BTS INDUCES TRUTH-TELLING, EVEN WITH COMPETING (WEAKER BUT MORE CERTAIN) EXAGGERATION INCENTIVES (EXPERIMENT 2)

A: Mean Recognition Rates Across Participants

	Control	*Overclaiming Only*	*BTS Only*	*BTS + Overclaiming*
Reals	58%	77%	54%	56%
Foils	24%	55%	13%	16%

B: Logistic Regression of Recognition Judgments on Item Type and Incentive Treatments

Variable	*Coefficient*	*Robust SE*	*z*	p > *z*
BTS incentives	−.813	.125	−6.52	.000
Overclaiming incentives	.605	.125	4.86	.000
Foil	−1.610	.086	−18.80	.000
BTS incentives × overclaiming incentives	−.908	.249	−3.65	.000
BTS incentives × foil	−.717	.171	−4.19	.000
Overclaiming incentives × foil	.322	.171	1.88	.060
BTS × overclaiming × foil	−.227	.343	−.66	.508
Intercept	−.064	.062	−1.03	.304

Notes: Pseudo $R^2 = .13$, Wald $\chi^2(7) = 523.36$, $p < .0001$.

[5]We obtained similar results when we add item fixed effects to the model, except that the overclaiming incentives × item type interaction (which is irrelevant to our hypothesis) becomes insignificant ($p = .31$). Our conclusions are also the same under the model we ran in Experiment 1 (i.e., a linear regression with each participant's overall recognition rate by item type, not individual recognition judgments, as the dependent variable).

Figure 2
AVERAGE PAYMENTS PER ITEM UNDER BTS INCENTIVES (EXPERIMENT 2)

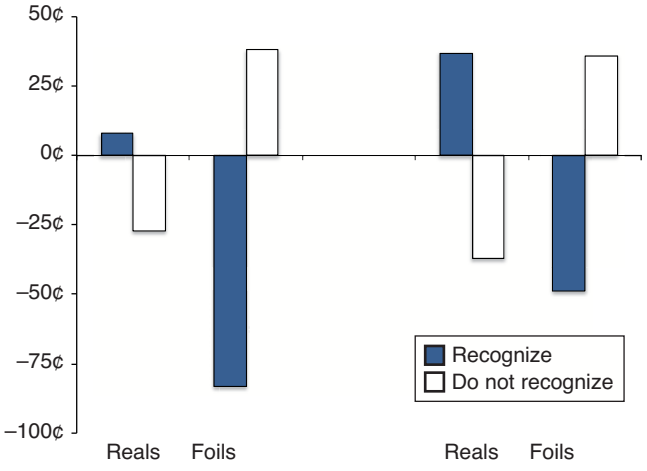
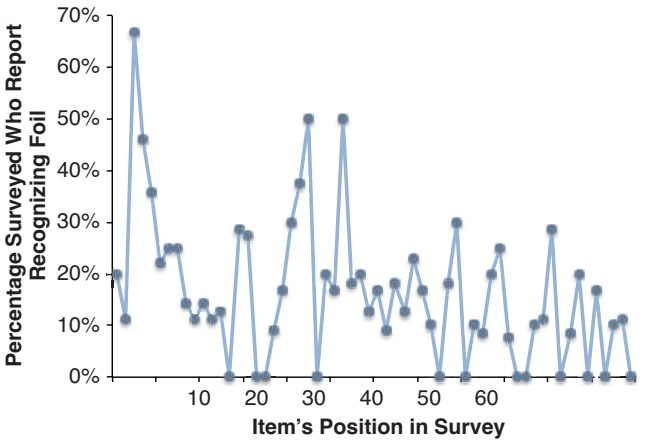


Figure 3
FOIL "RECOMMENDATION" DECLINES OVER THE SURVEY AS PARTICIPANT RECEIVES BTS PAYMENT FEEDBACK (EXPERIMENT 2)



larger, and the penalty for recognizing foils smaller because recognition also carries a 25¢ bonus. Because two-thirds of questionnaire items are reals, it may therefore seem that a participant could gain by exploiting this bonus and claiming to recognize every item. However, this is not the case. From the perspective of a questionnaire respondent, an unfamiliar legitimate item is probably indistinguishable from a fake, and because reals are recognized at much higher rates, an unrecognized item is likely to be a foil. Genuine judgments also tend to score higher on average even when a person is truly unfamiliar with a legitimate item. On average, a participant in the BTS + overclaiming group who claimed to recognize everything would have earned an average of 8¢ per item (–17¢ from iscores + 25¢ for recognition). However, people in this group actually earned 35¢ per item on average.

Learning and Iscore Trends

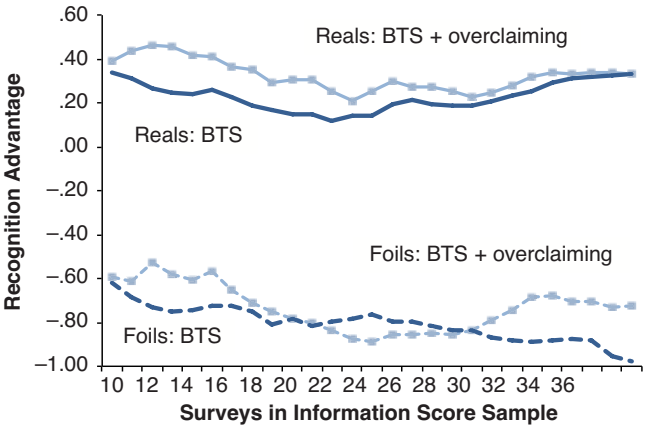
Randomizing the items' presentation order across participants permits us to look for changes in behavior as respondents progressed through their questionnaires. In particular, we are interested in whether exposure to recognition and nonrecognition payment values for particular items engendered trust that truth telling would be rewarded. If so, we would expect lower recognition for the items, especially foils, presented toward the end of the survey. Figure 3 shows that for the BTS + overclaiming group, there indeed appears to be a learning trend. To test the statistical reliability of the pattern, we ran a logit model of individual item recognition judgments on survey position, item type, and their interaction, with item fixed effects. We dropped the main effect of item type because of collinearity with the fixed effects, and we dropped several items because of unanimous recognition or nonrecognition. In the resulting model, there is no main effect of item position (b = .0001, z = .03, $p > .95$), but the interaction of position and foils is negative and significant (b = –.03, z = –2.86, $p < .005$), confirming that foil recognition likelihood declined as the survey progressed. A linear probability model with the same regressors yielded similar results. In that model, the coefficient on the item position × item type interaction term is –.0029 (t = –2.59, $p < .02$), meaning that for each survey item preceding a foil, recognition dropped by .3 percentage points—an aggregate decline of approximately 18 percentage points across all 60 items.

Similar regressions for the other experimental conditions showed no significant trends. A plausible explanation for the absence of a learning trend in the BTS (without overclaiming incentives) condition is that because those participants had no financial incentives to exaggerate, they tentatively accepted the truth-rewarding claim at the survey's outset. If so, experience may have increased their confidence in this claim but would not meaningfully change their behavior. (Recall that this group recognized the fewest foils overall: 13% vs. 16% for the BTS + overclaiming condition.) In contrast, participants facing competing incentives may have been tempted initially to exploit the certain recognition payments but discovered through experience that the penalties for exaggeration outweighed these gains.

We also examined changes in iscores as the number of survey respondents increased. Figure 4 plots the average recognition advantage (the difference between the scores for recognition and nonrecognition) as a function of sample size for both item types in each BTS condition.[6] The figure shows that the advantage of recognizing reals, and disadvantage of recognizing foils, holds for all values of N for which we have data. The scores also appear to be fairly stable as N changes. To test these impressions more precisely, we regressed recognition advantage on sample size, item type, and their interaction. Two trends emerged: in the BTS condition, the foil recognition penalty improved somewhat (t = –6.15, $p < .0001$), and in the BTS + overclaiming group, the real recognition advantage deteriorated slightly (t = –2.40, $p = .02$). Both effects are modest, especially the sec-

[6]The minimum sample size is ten because, as explained previously, we seeded the database in each condition with the survey responses of ten volunteers who were naive to our purpose and hypotheses.

Figure 4
ISCORE DIFFERENTIALS ARE MOSTLY STABLE AS SAMPLE SIZE INCREASES



Notes: "Recognition advantage" is the difference between iscores for recognizing and not recognizing an item. As we expected, these values are positive for reals and negative for foils; the point here is that while absolute iscores may drift somewhat, there is no significant change in their difference: The BTS discrimination between responses is largely stable.

ond, whose regression coefficient implies a change of only –.005 for each additional N. In summary, iscores were generally well behaved over the sample sizes in our experiment.

Analysis of Specific Items

To better understand the effect of truth-telling incentives, we investigate in more detail the results of the BTS-only condition versus the control group. We note that BTS reduced recognition rates for 46 of the 60 items on the survey. The four biggest reductions were for invented items: Miriam Fischer (–27 percentage points), Michail Stoika (–27), capacitance (–27), and granine (–24). Number five was Vicente Fox (–19), the real-life former President of Mexico. The prominence of foils is unsurprising because we have shown that foils are more responsive than reals to truth-telling incentives. A more subtle observation is the *content* of items on which BTS exerts the most pull. As we explained previously, some overclaiming is motivated by a desire for self-enhancement (Paulhus et al. 2003). A plausible hypothesis, then, is that items whose recognition promotes a positive self-image are more susceptible to overclaiming in the control condition and are therefore subject to a larger correction by BTS. In our survey, we propose that the self-enhancement items are those in the three education categories (world leaders, language arts, and life sciences) as opposed to the lifestyle categories (alcohol brands, movie comedies). The top five reductions are all for education items, and a t-test confirms the effect across all items ($M_{education} = –.08$, SD = .09; $M_{lifestyle} = –.03$, SD = .08; $t = 2.21$, $p < .05$). This is not because education items had "farther to fall," because the control group recognized education and lifestyle items at approximately the same rate ($M_{education} = .49$, SD = .32; $M_{lifestyle} = .43$, SD = .28; $t = –.77$, $p > .40$).

Of the remaining 14 survey items, recognition rates were unchanged on 4 under BTS versus the control group. For 10 items, recognition rates modestly increased, of which 9 were reals (recognition of the made-up film *School for Dogs* increased by one percentage point). It is also instructive to examine foils that were resistant to truth-telling incentives. Of the 20 foils, 3 had recognition rates that were both abnormally high under BTS and essentially unchanged compared with the control group. These outliers are the bogus language arts term "interjunction" (recognized by 54% of BTS participants) and the fake movies *Sister Act 3: Soul Remedy* (43% recognition) and *School for Dogs* (36%). (All other foils had recognition rates of less than 20%.) A likely explanation is that because these items are sufficiently similar to familiar real items (interjection, *Sister Act* and *Sister Act 2: Back in the Habit*, *Hotel for Dogs*), many survey takers genuinely believed them to be legitimate.

Can Iscores Identify Individual "Liars"?

Although a low iscore on a particular item does not imply that the recipient has lied, our method does hold out the possibility of identifying untruthful people by aggregating their scores across a battery of items. We find promising evidence of this. Again focusing on the BTS-only condition, we compared each respondent's total iscore to the proportion of foils recognized and to d′, which measures how well a participant distinguishes real items from fakes. (The signal detection measure d′ is defined as z[hit rate] – z[false alarm rate], where in our data a "hit" is a recognized real, and a "false alarm" is a recognized foil.) The correlations are strong: r(iscores, foil recognition) = –.60, and r(iscores, d′) = +.69.

In particular, the results suggest that overall data quality might be improved by eliminating three participants with very low iscores of 4.1, 5.2, and 6.9 (all others scored between 9.4 and 14.6). In addition to recognizing foils at high rates, these three gave some particular responses that seem unlikely to be true. For example, one claimed to recognize the movie *The Deli*, which was never released in U.S. theaters, but not the cult comedy classic *The Big Lebowski*, ranked 135th among IMDb.com users' favorite all-time movies. Another was familiar with Oronoco, an obscure rum brand, but not Jim Beam, one of the top ten selling spirits in the United States according to the Adams Liquor Handbook.

EXPERIMENT 3

The results of our first two experiments suggest that survey takers respond to iscore-based incentives—not merely promises to reward truth telling—to a degree that can meaningfully reduce untruthful responding. In Experiment 3, we implemented BTS incentives without explaining the basis of the payments and without asking people to answer the questionnaire honestly. We created these conditions to minimize demand effects or other social pressures to respond truthfully. This is a strong test, in that the mechanism can induce truthfulness only if people are aware that truthful answers maximize expected earnings.

Design and Procedure

We closely followed the procedures of Experiment 2: We presented the same questionnaire of 60 items in random order with feedback after each item, administered remotely to the same online panel. Payments were 1.5 times the iscores. We departed from the previous protocol, however,

in three ways. First, the presurvey instructions were intentionally uninformative:

> This survey lists products, people, and things from various categories. We want you to tell us which of these items you recognize. For each complete answer, you will earn money based on your response and the responses of people who took the survey before you. You may receive negative payments on some items, but your total earnings will not be negative.

Second, because we left participants "in the dark" as to how payments were calculated, it would be natural to presume a causal relationship between their predictions and those payments, making the correct inference unrealistically difficult in our estimation. We therefore asked only for personal recognition judgments and not predictions of others. Third, as a consequence of this decision, it was not possible to update iscores in real time, so instead we borrowed the ending scores (and their associated payments) from Experiment 2's BTS condition.

Results and Discussion

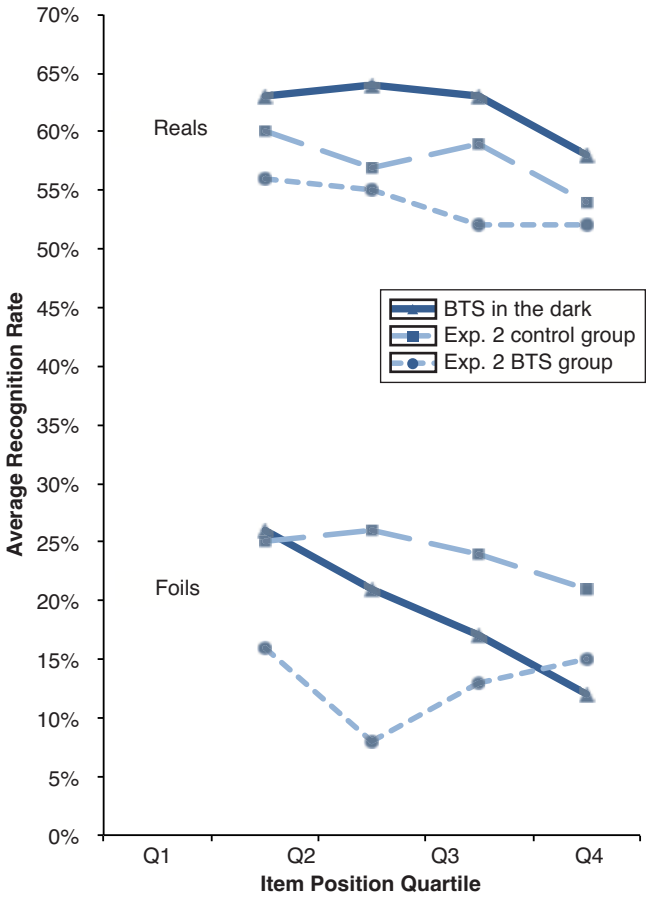
Twenty-seven people completed the questionnaire. We first looked for evidence of learning as participants progressed through the survey. Here, the potential for learning is even stronger because people are given no context for the payments presented on the feedback screens. As in Experiment 2, we regressed item recognition judgments in a logit model on survey position, item type, and survey position × item type, with item fixed effects. We dropped the main effect of item type due to collinearity with the fixed effects. The main effect of item position is slightly negative but not significant ($b = -.005$, $z = -1.15$, $p = .25$), meaning that the recognition of reals was not substantially affected by survey position. However, there is a significant interaction between survey position and foil types ($b = -.02$, $z = -2.08$, $p < .05$), confirming the previous experiment's finding that respondents were less likely to claim recognition of foils when they were presented later in the survey.

Participants recognized 62% of the real items and 19% of the foils overall. To evaluate these results, we compared them with recognition rates for Experiment 2, particularly the BTS condition, which had identical incentives, and the control group. Viewed as a whole, the truth-inducing performance of our "BTS-in-the-dark" implementation does not seem impressive. Participants recognized reals at a higher rate than both Experiment 2's control condition (58%;$\chi^2(1) = 4.31$, $p = .04$) and its BTS condition (54%;$\chi^2(1) = 15.08$, $p = .0001$). On foils, BTS-in-the-dark did better than the control (24%; $\chi^2(1) = 4.16$, $p < .05$) but worse than BTS (13%; $\chi^2(1) = 7.57$, $p = .006$).

However, these aggregate data mask changes in recognition rates as survey takers gained experience with the truth-rewarding payments. Figure 5 plots recognition rates for the three survey groups as a function of item position quartile (i.e., the first 15 items make up Q1, and so on). Consistent with the learning analysis, we observe that the performance of BTS-in-the-dark improves with item position, particularly for foils. Restricting our analysis to items presented in Q4, the recognition likelihood among reals was not significantly different than for the Experiment 2 comparison groups (vs. control: $\chi^2(1) = .68$, $p = .41$ vs. BTS: $\chi^2(1) = 1.90$, $p = .17$).

Figure 5

AVERAGE RECOGNITION RATES BY ITEM POSITION FOR BTS "IN THE DARK" (EXPERIMENT 3; DATA FROM EXPERIMENT 2 SHOWN FOR COMPARISON)



Furthermore, the foil recognition rate was actually lowest for BTS-in-the-dark, though the difference was significant only compared with the control ($\chi^2(1) = 3.91$, $p < .05$).

Experiment 3 shows that given sufficient exposure to payments derived from iscores, the truth-telling incentives were ultimately compelling despite no explanation of the mechanism and without any plea for honesty. Although we obtained this outcome implementing a "constrained" implementation of BTS (no predictions and static iscores), in a dynamic game, there is a truth-telling equilibrium only when every player believes that others are also responding truthfully. In the BTS-in-the-dark implementation, there is no particular reason to expect that this would have been the case.

EXPERIMENT 4

An important question we have not yet addressed is how well the BTS performs relative to other truth-telling mechanisms. Several such mechanisms are well-known. One is the "bogus pipeline" (Jones and Sigall 1971), in which participants are hooked up to an apparatus resembling a polygraph and (falsely) told the machine can detect deception. Because they believe their answers are actively monitored, people

often respond more truthfully than they would otherwise. A second method is to ensure participant anonymity, which can encourage candor about socially stigmatized behaviors. However, straightforward procedures are often insufficient, as Turner et al. (1998) find when surveying adolescents about various risky behaviors. Reports of male–male sex, injection drug use, and sexual contact with intravenous drug users tripled when traditional anonymous self-administered questionnaires were replaced by a more elaborate approach called "audio computer-assisted self-interviewing," which speaks recorded questions to respondents through headphones; respondents answer by pressing numbers on a computer keyboard.

Both methods are somewhat complex to implement and have other drawbacks that limit their practicality: The bogus pipeline's deception of survey takers is ethically objectionable, and anonymity techniques might do little to correct untruthfulness motivated by reasons other than social sensitivity (e.g., boredom, carelessness). Therefore, for our comparison, we chose a benchmark method that not only has been shown to be effective but also is easy to implement: the "solemn oath," in which participants are asked to sign a declaration to tell the truth before participating. Jacquemet et al. (2009) find that asking people to sign an oath elicited true preferences better than either nonbinding hypothetical judgments or binding judgments made incentive-compatible through a second price auction. In an induced values experiment, people who took the oath submitted bids that were closer to their true valuations. Moreover, when eliciting bids for dolphin protection, the solemn oath reduced two biases common in contingent valuation studies: the tendency for hypothetical bidders to declare high bids, seemingly in violation of their budget constraint, and the tendency for incentive-compatible bidders to bid zero, often a sign that they have opted out of the exercise in protest.

We administered our knowledge questionnaire to two groups of participants. One group was asked to sign a presurvey oath, and the other group was given BTS incentives. Both groups were also paid a small amount for each item they recognized, creating a competing incentive to exaggerate their knowledge.

Design and Procedure

We again borrowed the protocol used in Experiment 2: All participants completed our 60-item knowledge questionnaire with randomized presentation order, gave personal recognition judgments as well as predictions of others', and received payment information after each item. So that we could properly administer the oath, this experiment was conducted in person in a computer lab. Each lab session was randomly assigned to one of two groups, BTS or solemn oath.

Participants in the BTS condition were given the instructions shown for BTS + overclaiming incentives from Experiment 2, as shown in Appendix A. However, the financial incentives were somewhat smaller: BTS payments were 1 times the iscores (rather than 1.5 times), and recognition payments were 15¢ per recognized item (not 25¢). We calculated iscores in real time using all prior responses from members of the same experimental group, plus the ten-person seed data we used before.

Participants in the solemn oath condition were given the previously used instructions for the control + overclaiming group (also in Appendix A), again with reduced incentives: a flat payment of 10¢ per item, plus 15¢ per recognized item. Before being given these instructions, participants in the oath condition were asked to sign a solemn oath. In their research, Jacquemet et al. (2009) reviewed the theory of commitment in social psychology (Kiesler and Sakumura 1966) and concluded that to be effective, an oath should be freely undertaken, publicly expressed, and signed. Accordingly, we distributed copies of the oath on paper and emphasized that signing it was voluntary and that declining to sign would not affect earnings. The oath read as follows:

> Topic: Research study number F17806—103.
>
> I, [print name□ swear on my honor that, during the experiment, I will:
>
> [sign and date□

Oaths were collected before the survey began. In both groups, a summary screen at the end of the questionnaire reported the participants' total earnings, which we paid in cash at the end of the session.

Results and Discussion

Two people in the solemn oath condition declined to sign the oath and were excluded from analysis (remaining N = 70). The participants in the BTS condition reported recognizing substantially fewer items, both legitimate and fake, than those who signed the solemn oath and were given flat payments (Table 4, Panel A). A factorial logistic regression confirms that the likelihood of recognition was significantly lower for the BTS group, foils, and their interaction (Table 4, Panel B). We found no meaningful learning trend in either group. These results suggest that BTS was more effective than the solemn oath in inducing truth telling.

Because our study did not have a baseline condition, we cannot conclude that the oath was completely ineffective. However, comparing the percentage of foil recognition under oath in Study 4 with the percentage without oath in Study 1 suggests that the impact of oath on truth telling is marginal at best. A possible explanation for the difference

Table 4
BTS VERSUS SOLEMN OATH

A: Proportion of Items Recognized

	BTS	*Oath*
Reals	54%	72%
Foils	21%	52%

B: Logistic Regression of Recognition on Mechanism and Item Type

Variable	*Coefficient*	*Robust SE*	*z*	p > *z*
BTS incentives	−1.08	.185	−5.84	.000
Foil	−1.17	.083	−14.17	.000
BTS incentives × foil	−.68	.165	−4.12	.000
Intercept	−.05	.093	−.51	.628

Notes: Pseudo R^2 = .08, Wald $\chi^2(3) = 290.51$, $p < .0001$.

between these results and those of Jacquemet et al. (2009) is that their participants did not face a financial incentive to deceive. In our study, recognition payments created such an incentive, though BTS appears to have overcome it. Finally, the BTS group earned more money (M = $12.95, SE = $.45) than the oath group (M = $11.53, SE = $.35; t = 2.54, p = .007), though this does not explain the observed differences in claimed recognition.

EXPERIMENT 5

A domain in which BTS might have particular use is in improving elicited valuations for nonmarket goods. A frequently used technique is the contingent valuation (CV) method, in which respondents are asked to declare the amount of money they would personally pay to acquire, or demand for the loss of, the good of interest—for example, a national park. A common criticism of CV is that because it lacks truth-telling incentives, respondents might ignore income constraints, lodge protest responses, or otherwise deviate from their true preferences.

Various techniques have been proposed to improve the quality of CV data. One is cheap talk—that is, explicitly instructing participants that hypothetical responses are sometimes inflated. Such instructions have shown mixed results (Cummings and Taylor 1999; List 2001). A different approach is calibration: Even if hypothetical valuations are biased, they are informative if the direction and degree of bias are stable. List and Shogren (1998), however, find that calibration functions are both good and context specific. These results suggest that iscoring might further improve the quality of CV responses, a proposition we test in Experiment 5.

Design and Procedure

The nonmarket good we chose for the study was the National Endowment for the Arts, an agency of the U.S. government that distributes public funds to arts projects. We conducted referenda on personal contributions to the NEA. Each respondent was asked to vote either in favor of or in opposition to donating $4, with the understanding that the majority's will would be imposed on all study participants. We conducted four separate referenda, one with binding outcomes and three with nonbinding outcomes, for different experimental groups:

1. *Real*: Participants were instructed that the outcome was binding: "In the referendum about donating $4 to this organization, what is your vote? Remember, the stakes are real: if the majority is in favor, we will donate $4 on your behalf; if the majority is opposed, we will pay you a $4 bonus."
2. *Hypothetical*: Participants were instructed that the referendum was nonbinding but to treat it as real: "If you participated in a referendum about donating $4 to this organization, how would you vote? Remember, although no real money is at stake, we want you to respond as if the stakes were real, and the majority's choice really applied to you."
3. *BTS*: Participants in the BTS treatment received the same instructions as the hypothetical group and were also told that they would receive an extra payment based on their response's "Truth Score." We explained BTS incentives using the same language as in Appendix A. To facilitate information scoring, here we also elicited predictions of the proportion of respondents who would vote in favor of the referendum.
4. *BTS with training:* Instructions were identical to those for the BTS group. In addition, this group received prior exposure to BTS incentives through a preceding task involving a 36-item subset of the recognition questionnaire used in our prior studies. In this task, we explained the BTS mechanism and disclosed the payments for both recognition and nonrecognition after each item.[7]

We ran BTS treatments both with and without prior exposure to evaluate the importance of experience in establishing the method's credibility for CV judgments. Experiment 1 showed that the mechanism was both truth inducing and truth rewarding without prior training for the recognition questionnaire. However, in applying BTS to contingent valuations as we do here, Barrage and Lee (2010) find that it reduces, but does not altogether eliminate, hypothetical bias.

Results and Discussion

The study was administered to an online panel of 121 participants. We eliminated 7 of them due to data integrity concerns stemming from large numbers of unanswered items, implausibly short time spent on the questionnaire, and failures to pass screening questions. (Three screening items were embedded in the general knowledge task in which we asked respondents if they recognized the U.S. states of Colorado, Maryland, and "Alberta.")

Among the remaining participants (N = 114), the proportions in each treatment who voted in favor of donating $4 to the NEA were as follows:

Treatment	*Percentage in Favor of Donating to the NEA*
Real	44%
Hypothetical	76%
BTS	47%
BTS with training	50%

The real treatment serves as a baseline against which any bias in the nonbinding groups may be measured. The hypothetical treatment shows the common pattern in CV elicitations of overstated intentions of contributing to public goods ($\chi^2(1) = 6.09$, $p = .01$). However, this bias is eliminated in both of the BTS treatments: The comparison of real versus BTS has a $\chi^2(1)$ statistic of .04 ($p = .84$); for real versus BTS with training, $\chi^2(1) = .18$ ($p = .67$).[8] The success here of BTS without prior training is consistent with the results of Experiment 1, but less so with Barrage and Lee's (2010) findings. The reason for this inconsistency is unclear, but it is possibly related to differences in the instructions participants received.

SOME PRACTICAL CONSIDERATIONS

In four experimental administrations of the overclaiming questionnaire, the BTS generally rewarded truth telling, in particular by assigning large penalties to demonstrably untrue responses, and induced truth telling, in that the application of BTS-based payments reduced claims of familiarity

[7]As a control, we administered the same preceding task in all other experimental treatments, but those groups were not told about BTS incentives, asked to provide predictions, or given feedback.

[8]Voting in favor of donating was surprisingly common. In the BTS (without training) condition, iscores were .35 for in favor and −.10 for opposed; for BTS with training, they were .26 and −.03, respectively.

with fake items. This improvement in truth telling was robust to (smaller but more certain and straightforward) competing incentives to exaggerate knowledge. Participants also became more truthful as they progressed through their questionnaires, suggesting that the truth inducement was the result of actual BTS incentives rather than placebo, demand, or cheap talk effects. In a side-by-side comparison, BTS outperformed the solemn oath, a rival truth-inducement mechanism, and it also eliminated hypothetical bias from contingent valuations of a public good, both with and without prior exposure to the mechanism.

Different Types of Untruthfulness

To implement BTS successfully, two criteria must be met. First, participants must believe that the method rewards truth telling on average. This is not to say that administrators must provide a full explanation of the scoring formula or the intuition behind it; we did not. Indeed, such details may do more harm than good by creating confusion or encouraging attempts to game the system.

We found a simple explanation to be credible. In our studies, a presurvey demonstration of the method using scores from previously collected data was not necessary, but Barrage and Lee (2010) find that without it, BTS only partially eliminated untruthfulness. Prior exposure is likely to be more important in inducing truth telling for survey questions that are deeply personal, in which case the truth serum's face validity might be especially low, and when respondents have significant motivations, financial or otherwise, to be less than fully truthful.

The second criterion is that BTS-based payments (which can be arbitrarily scaled as a function of iscores) must be large enough to overcome participants' competing incentives to misrepresent themselves. Here, it is useful to consider three types of untruthfulness: intentional deception, carelessness, and inauthenticity (i.e., answers that are biased—possibly without the respondent's conscious awareness—by social norms, cognitive heuristics, or environmental factors). It is clear that some responses to our surveys were intentionally deceptive, in that item recognition rates increased substantially when we created incentives to overclaim. In our studies, BTS appears to have largely eliminated such deception. However, for some content, including questions about socially stigmatized behavior, even substantial BTS incentives might be inadequate, particularly because payments are a function of survey responses, making it difficult to implement the method anonymously. In these cases, preserving anonymity may be a more effective means of motivating truthful disclosure (Turner et al. 1998).

The BTS also improved truth telling in the absence of financial incentives to exaggerate, in which case the "recognition" of foils was probably the result of carelessness or inauthenticity. These effects are difficult to distinguish: When BTS reduces foil recognition compared with a control group, it is unclear if participants have become more careful, suppressed a desire for self-enhancement, or both. Evidence that BTS groups spent more time on their surveys, however, suggests that at least some of the effect results from greater care. Moreover, some sources of inauthenticity may be fully unconscious and stubbornly resistant to even large financial rewards for truth telling. For example, Paulhus (1984) distinguishes two types of socially desirable responding: impression management, the purposeful manipulation of answers to create a positive social image, and self-deceptive positivity, which may be unconscious and is used to help maintain self-esteem and optimism. Variations in demand for social desirability (such as an expectation that survey responses will be made public) influence impression management but have no effect on self-deception.

Is BTS Worth the Overhead?

It is difficult to extrapolate from the effectiveness of the BTS in our experimental setting to the wide variety of market research surveys now or potentially in use. The method clearly asks more of survey takers—in particular, their predictions about the distribution of responses. From a theoretical standpoint, it would be sufficient to elicit predictions from a small number of randomly selected respondents and use their predictions to calculate initial iscores. The remaining respondents would only be required to provide answers, as in a traditional survey. The iscores could be periodically updated as the data on empirical proportions accumulates. An advantage of having provisional iscores in hand is that respondents could be scored and rewarded as soon as they enter a response.

On the issue of 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 seems 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 on the survey, just as they would on a standardized 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. The benefits of the BTS 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, beliefs, or preferences.

APPENDIX A: KNOWLEDGE QUESTIONNAIRE INSTRUCTIONS FOR EXPERIMENT 2

Instructions for All Participants

This survey lists products, people, and things from various categories. We want you to tell us if you recognize these items, and to estimate the percentage of other people who will recognize them.

Additional Instructions for BTS Incentives Group

For each complete answer, you will earn money based on that answer's "Truth Score." Truth Scoring, recently invented by an MIT professor and published in the academic journal

Science, rewards you for answering truthfully. Even though only you know if you really recognize a survey item, people who tell the truth score higher overall.

You are most likely to maximize your earnings if you answer every item truthfully. By "truthfully," we mean: consider each item carefully, answer honestly, and take care to avoid mistakes.

If Also Given Overclaiming Incentives: You will also earn 25¢ for each item that you recognize. But remember that your Truth Score will be lower if you don't respond truthfully, so the best way to earn more money overall is to answer every item honestly.

Additional Instructions for Control Group

For each complete answer, you will earn 25¢. Please consider each item carefully, answer honestly, and take care to avoid mistakes.

If Also Given Overclaiming Incentives: You will also earn an extra 25¢ for each item that you recognize. But we still want you to answer every item honestly.

Appendix B

KNOWLEDGE QUESTIONNAIRE ITEMS USED IN EXPERIMENTS 2, 3 AND 4

Alcohol Brands
- 99 Bananas
- Bacardi Breezer
- Cabo Wabo
- Cardinal Sin[a]
- Crescent Moon[a]
- Gettysburg[a]
- Hypnotiq
- Jim Beam
- Labatt's
- Oronoco
- St. Petersburg[a]
- Yellowtail

Language Arts
- Alliteration
- Aphorism
- Capacitance[a]
- Eponym
- Euphemism
- Grapheme
- Hyperbole
- Interjunction[a]
- Interrogative
- Lexical shunt[a]
- Limiting adjective
- Sentence stigma[a]

Life Sciences
- Cellular gulf[a]
- Cerebral cortex
- Echinoderm
- Granine[a]
- Hayflick limit
- Hemoglobin
- Lupine
- Lysogenic bacterium
- Meiosis
- Retroplex[a]
- Superplacentals[a]
- Telomere

Movie Comedies
- *Be Cool*
- *Couples Retreat*
- *Do You Want Fries With That?*[a]
- *Half Baked*
- *Man of the Year*
- *Money for Something*[a]
- *Rumor Has It...*
- *School for Dogs*[a]
- *Sister Act 3: Soul Remedy*[a]
- *The Big Lebowski*
- *The Deli*
- *Young Frankenstein*

World Leaders
- Angela Merkel
- Atal Behari Vajpayee
- Garrett Andersen[a]
- Golda Meir
- Idi Amin
- Irwin Anders[a]
- Keizo Obuchi
- Kim Jong Il
- Michail Stoika[a]
- Miriam Fischer[a]
- Stephen Harper
- Vicente Fox

[a]Foil (nonexistent item).

REFERENCES

Barrage, Lint and Min Sok Lee (2010), "A Penny for Your Thoughts: Inducing Truth-Telling in Stated Preference Elicitation," *Economics Letters*, 106 (2), 140–42.

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 (1–3), 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 (6), 1247–57.

Cummings, Ronald G. and Laura O. Taylor (1999), "Unbiased Estimates for Environmental Goods: A Cheap Talk Design for the Contingent Valuation Method," *The American Economic Review*, 89 (3), 649–65.

d'Aspremont, Claude and Louis-André Gerard-Varet (1979), "Incentives and Incomplete Information," *Journal of Public Economics*, 11 (1), 25–45.

Dawes, Robyn M. (1989), "Statistical Criteria for Establishing a Truly False Consensus Effect," *Journal of Experimental Social Psychology*, 25 (1), 1–17.

——— (1990), "The Potential Nonfalsity of the False Consensus Effect," in *Insights in Decision Making* , R. Hogarth, ed. Chicago: University of Chicago Press, 179–99.

Engelmann, Dirk and Martin Strobel (2000), "The False Consensus Effect Disappears if Representative Information and Monetary Incentives Are Given," *Experimental Economics* , 3 (3), 241–60.

Jacquemet, Nicolas, Robert-Vincent Joule, Stéphane Luchini, and Jason F. Shogren (2009), "Preference Elicitation Under Oath," CES Working Papers, 43, (June 9), (accessed October 18, 2012), [available at ftp://mse.univ-paris1.fr/pub/mse/CES2009/09043.pdf□

John, Leslie K., George Loewenstein, and Drazen Prelec (2012), "Measuring the Prevalence of Questionable Research Practices with Incentives for Truth-Telling," *Psychological Science* , 23 (5), 524–32.

Johnson, Scott J., John Pratt, and Richard J. Zeckhauser (1990), "Efficiency Despite Mutually Payoff-Relevant Private Information: The Finite Case," *Econometrica*, 58 (4), 873–900.

Jones, Edward E. and Harold Sigall (1971), "The Bogus Pipeline: A New Paradigm for Measuring Affect and Attitude," *Psychological Bulletin*, 76 (5), 349–64.

Kiesler, Charles A. and Joseph Sakumura (1966), "A Test of a Model for Commitment," *Journal of Personality and Social Psychology*, 3 (3), 349–53.

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 (4), 596–610.

List, John A. (2001), "Do Explicit Warnings Eliminate the Hypothetical Bias in Elicitation Procedures? Evidence from Field Auctions for Sportscards," *The American Economic Review*, 91 (5), 1498–1507.

——— and Jason F. Shogren (1998), "Calibration of the Difference Between Actual and Hypothetical Valuations in a Field Experiment," *Journal of Economic Behavior and Organization*, 37 (2), 193–205.

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 (November), 633–44.

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), 2421–54.

Miller, Nolan H., Paul Resnick, and Richard J. Zeckhauser (2005), "Eliciting Informative Feedback: The Peer-Prediction Method," *Management Science*, 51 (9), 1359–73.

Paulhus, Delroy L. (1984), "Two-Component Models of Socially Desirable Responding," *Journal of Personality and Social Psychology*, 46 (3), 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 (June).

———, 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), 921–40.

Prelec, Drazen (2004), "A Bayesian Truth Serum for Subjective Data," *Science*, 306 (5695), 462–66.

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 (3), 279–301.

Turner, C.F., L. Ku, S.M. Rogers, L.D. Lindberg, J.H. Pleck, and F.L. Sonenstein (1998), "Adolescent Sexual Behavior, Drug Use, and Violence: Increased Reporting with Computer Survey Technology," *Science*, 280 (5365), 867–73.