Running head: INCENTIVES, ERROR, AND DATA SHARING

Incentives, Error, and Data Sharing

Alexander L. Davis*

Department of Social and Decision Sciences

Carnegie Mellon University

*E-mail address: alexander.l.davis1@gmail.com, Department of Social and Decision Sciences, Carnegie Mellon University, Pittsburgh PA 15213, USA, 412-216-2040.

Abstract

This research examines when individuals solving Wason's 2-4-6 rule discovery task attribute disconfirming feedback to error, and when they decide to share perceived errors with another person trying to solve the same problem. Participants invoked error for disconfirming feedback more often than for affirming feedback. Data sharing decisions during the task (Experiment Two) and at the end of the task (Experiments Three and Four), found that participants were less likely to share trial results when feedback was disconfirming or when trials were attributed to error, even after controlling for actual error. Experiments Two through Four found that incentives had no effects on rule discovery or attributions to error. Although, the perverse incentive used in Experiments Three and Four gave participants a financial motive to suppress data, they typically did the opposite, deciding to share unconvincing data that could cost them money. The experiments suggest that scientists will fail to share disconfirming data, attributing them to error even when they are not, and that perverse incentives will not lead to deception of oneself or others as long as the audience is kept in mind.

Incentives, Error, and Data Sharing

Hypothesis testing has been found to follow a *positive test strategy*. Researchers collect data that they expect to conform to their prior beliefs and then exaggerate its information value (Klayman & Ha, 1987), while discounting any inconsistent evidence that comes their way (Dunbar, 1995, 2001; Gorman, 2005; Lord, Ross, & Lepper, 1979). This result has been found with both simple experimental tasks and in dynamic artificial environments, such as simulated molecular biology (Dunbar, 1993), programmed robots (Klahr & Dunbar, 1988), and multiple-cue probability learning (O'Connor, Doherty, & Tweney, 1989). Similar patterns have been found in scientific laboratories. For example, in an observational study of a biological sciences laboratory, Dunbar (2001) found that scientists did not immediately reject their hypotheses after they were contradicted by data. Rather, their first reaction was to invoke experimental error (Dunbar, 1995, 2001; Gorman, 2005). In thirty-seven experimental treatments conducted by one biologist, twenty-one had unexpected results, most of which were treated as errors (Dunbar, 2001).

Wason's 2-4-6 rule discovery task (Wason, 1960) has served as an archetype for studying scientific inference. This task has been fruitful because it captures two important features of real scientific research: an unbounded hypothesis space and strong prior biases toward a specific hypothesis (ascending evens). However, it also lacks several key features that scientists must face; one of them is uncertainty. When the possibility of error is added to the feedback provided in the rule-discovery task, participants attribute disconfirming feedback to error more than affirming feedback, and have greater difficulty discovering the rule (Gorman, 1986, 1989; Penner & Klahr, 1996).

Another key feature that is missing from this task is the social context of scientific research. In the present research, we focus on two contextual factors: data sharing and incentives. After inspecting their data for errors, researchers must decide whether to communicate any data that they consider flawed. Indirect evidence suggests that people are reluctant to give diffuse (imprecise) statements about uncertain quantities because they seem 'uninformative' (Yaniv & Foster, 1995), and because communicating data that one believes to be inaccurate (even if one does not know the accuracy for sure) violates conversational norms (Grice, 1975). Studies looking at publication bias consistently find that statistical significance is often (incorrectly) interpreted as the probability of error in data, and is usually a necessary condition for publication (Fanelli, 2012; Sterling, 1959).

If the researchers' attributions to error are accurate (Oaksford & Chater, 2007), then omitting these errors from published reports may avoid distracting readers. If attributions to error are inaccurate (Kahneman & Tversky, 1996; Tversky & Kahneman, 1974), then failing to publish those data will allow false theories to emerge and persist. Accuracy in the rule-discovery task should be difficult because the hypothesis space is unbounded, ensuring that people rarely benefit from unambiguous feedback about the presence and causes of experimental error (Lichtenstein & Fischhoff, 1980; Lichtenstein, Fischhoff, & Phillips, 1981). Data attributed to error are unlikely to be shared, whether or not these attributions are accurate.

Data sharing decisions are not only affected by whether the data are perceived to be faulty, but also by professional rewards for publishing positive (usually statistically significant) results (Angell, 2000; Bodenheimer, 2000; Nathan & Weatherall, 1999; Weatherall, 2000). These rewards can produce a healthy motivation to make a discovery, resulting in more complex evaluation of the data and better error identification (Tetlock & Kim, 1987; Harkness, DeBono, & Borgida, 1985).

However, the effect of incentives on performance in experimental tasks is not uniformly positive (Camerer & Hogarth, 1999). Motivation helps judgment in easy tasks (e.g., resistance to persuasion), but is ineffective or harmful in difficult ones (e.g., judgment under uncertainty) (Pelham & Neter, 1995). Participants who were offered a monetary reward in an insight task had more difficulty achieving insight compared to those offered no reward (Glucksberg, 1962). These harmful effects may be due to incentives undermining accurate data inspection by increasing scrutiny of results that indicate the discovery is false, while simultaneously making affirming results a wanted relief from the pressure to produce (Kunda, 1990). Supporting this account, there is evidence that higher rewards for publishing are associated with publication bias (Ioannidis & Trikalinos, 2005; Fanelli, 2010).

Aside from encouraging self-interested data suppression, pressures to publish may also lead scientists to deceive themselves, as they simultaneously suppress disconfirming data but attempt to maintain the belief that they are honest and objective (Gino, Ayal, & Ariely, 2009; Gino & Ariely, 2012). Attributing disconfirming results to error can serve this purpose by providing particularly convincing justification, both to oneself and others, for not sharing data.

Here we present four experiments involving decisions to share possibly faulty data. We use Wason's 2-4-6 rule-discovery task (Wason, 1960). It asks participants to discover the rule that generated a set of numbers (2, 4, 6), by proposing a new set of three numbers (a proposed triple), then getting feedback as to whether the numbers that they proposed fit the rule. We use Penner and Klahr (1996)'s version, in which participants are told that some percentage of the time, the feedback will be false—a feature that adds something like the uncertainty that is inevitable with scientific inferences.

We add several new features to the task. On each trial (a) Before receiving feedback, participants assess the probability that it will affirm their expectations. (b) After receiving it, they indicate whether they would share the trial, including the feedback, with a second person trying to discover the same rule. In Experiment One, the sharing decision is done at the end of the task. In Experiment Two, the data sharing decision is done immediately after participants make their error judgments. In Experiments Three and Four, it is done both after each trial and at the end of the task. (c) We also use two types of incentives intended to simulate the rewards that may lead to motivated reasoning. Experiment Two provides participants with a large incentive ($100) for correctly guessing the rule, and a small incentive ($1) for concluding that they do not know the answer. Experiments Three and Four provide participants with an incentive to convince a matched participant that they discovered the rule, whether or not they actually did.

Using this task, we first replicate the finding that error is more likely to be invoked with disconfirming than with affirming feedback (Penner & Klahr, 1996; Gorman, 1986, 1989). We then examine whether these attributions to error are justified using two evaluative criteria: (a) *accuracy*, defined as whether the judgments are correct; and (b) *Bayesian consistency*, defined as attributing feedback to error if and only if either: (i) participants strongly expected the triple they proposed to fit the rule but it did not, or (ii) participants strongly expected the triple they proposed to not fit the rule but it did.[1] *Selective reporting* is the degree to which trials attributed to error are not shared with the matched participant, compared to those attributed to other sources. Previous research suggests that incentives to reach a particular conclusion (i.e., that one has made a discovery) or to convince others that one is correct, will decrease sharing of disconfirming evidence through self-interested and self-deceptive mechanisms such as attributing the data to error, as well as undermine objective error identification.

Experiment One

Experiment One looks at whether attributions to error are consistent with prior beliefs, whether they correspond to actual error, and whether trials are less likely to be shared with another person when the feedback is attributed to error. We used the Wason 2-4-6 rule discovery task with feedback error (Penner & Klahr, 1996).

Method

Participants. Eighteen Carnegie Mellon University undergraduates completed the task for course credit.

Procedure. Participants were seated at a computer, asked to sign an informed consent document, and then instructed that they had 30 minutes to complete the task. Participants completed the task online as a Qualtrics questionnaire with embedded Javascript used for feedback. Each page (trial) of the questionnaire had the same format. In order, participants proposed a rule, proposed a new triple, assessed the probability that the triple they proposed fit the Actual Rule, received feedback, judged whether the feedback reflected error, and then decided if they wanted to give their Final Answer. They were reminded to record all responses both on the computer and on the spreadsheet they were given. After participants decided to stop new trials and give their Final Answer, or thirty minutes had passed, they were asked to review their spreadsheet and mark the trials that they thought should be shared, in order to help a new participant solve the problem.

Materials. The materials were a modification of Penner and Klahr (1996)'s version of the Wason 2-4-6 rule discovery task (see `http://hdl.handle.net/1902.1/18699` for full materials).

Introduction. Participants were shown the following introduction on the computer along with a separate paper copy as a reminder:

> "You will be given three numbers that are related somehow. For example: 3, 5, and 15. This is called a triple. There are many possible rules that could relate these three numbers. We have selected only one of them. The rule that we selected is called the Actual Rule. You will not be given the Actual Rule. Your task is to discover it. The initial triple on the next page is an example drawn from the Actual Rule."
>
> "Our study is using several versions of this task. Yours is a particularly difficult one. Sometimes, even if your Proposed Triple FITs the Actual Rule, the computer may output that it DOES NOT FIT. Conversely, sometimes, when your Proposed Triple DOES NOT FIT the rule, the

> computer may output that it FITs. On any trial there is a 20% chance that you will get false feedback. For each trial if you think false feedback occurred mark “F” in the “Feedback” column on your spreadsheet. If you think true feedback occurred, mark “T” in the “Feedback” column.”
>
> “At any time you may try to guess the Actual Rule that we selected. This is called the Final Answer. You only get one Final Answer and it may be wrong. Once you make your Final Answer you can no longer get feedback from the computer and the experiment will end.”

Initial Triple. At the top of each page, the initial triple (2,4,6) was shown. Participants were told:

> “The initial triple above is an example drawn from the Actual Rule.”

Proposed Triple. After writing their best explanation of the initial triple, they were instructed to propose a new triple:

> “You may propose additional triples to help you discover the Actual Rule. The computer will tell you whether the triple you proposed fits the Actual Rule. Record all information on the spreadsheet you were given. Write one number of your triple in each box below.”

Prior Probability. On each trial, before they received feedback, participants assigned a probability that the triple they proposed fit the actual rule, by answering the following question:

> “What is the probability that the triple you proposed fits the Actual Rule? (must be a number between 0 and 100)”

We denote this $P(TFTR)$ for '[P]robability that the [T]riple [F]its [T]he actual [R]ule'.

Attribution to Error. Immediately after receiving feedback that the triple fit (FIT) or did not fit (DNF) the rule, participants judged whether they thought the feedback was due to error:

> "Do you think this feedback was true or false? (True/False)"

Final Answer. After participants felt they had completed enough trials, or the 30–minute window expired, they were asked to make their Final Answer:

> "Write your Final Answer for the Actual Rule in the box below (it can be mathematical or in words)."

Data sharing. Participants then decided which trials they wanted to share with a new participant:

> "In this experiment, a trial is a page where you proposed a rule, a triple, a probability estimate, received feedback, and judged whether you thought the feedback was false or true."
>
> "In a future experiment we will have a new participant try to discover the same rule you tried to discover.
>
> You can choose trials that you think will help him or her solve the rule. For each trial you indicate, all of the information would be shared, including:
>
> 1. your rule

2. the proposed triple
3. your probability estimate
4. the feedback
5. whether you thought the feedback was false or true

In the space below, please indicate the trials you conducted that you think would help this person."

Results

Unless otherwise noted, all estimation was done using hierarchical logistic models with subject-level varying intercepts (Gelman & Hill, 2007; Gelman et al., 2010). The model assumes that multiple observations from the same person are conditionally independent given the subject-specific intercept. Tests, standard errors, and p-values based on these models were calculated using non-parametric bootstrap with 200 simulations per statistic (Efron & Tibshirani, 1993).

Performance. Participants completed a median of eight trials. Each participant's task performance score was determined by their final answer, scored on a 5–point scale awarding one point for each element of the rule that they had discovered. The five elements were: 1) even numbers, 2) consecutive numbers, 3) ascending numbers, 4) the lower bound is 2, and 5) the upper bound is 100.

Attributions to Error. Replicating Penner and Klahr (1996), participants judged disconfirming feedback to be error more often (38%; $SE = 5.9\%$) than affirming feedback (8.6%; $SE = 4.4\%$), $t(162) = 3.77$ $p < 0.05$, $d = 0.30$. In multiple regression, there was only a main effect of feedback type on attributions of error

($t(159) = 3.1$, $p = 0.0075$), with no significant main effect of actual error ($t(159) = 1$, $p = 0.46$) or interaction between the two factors ($t(159) = 0.72$, $p = 0.62$).

Bayesian Consistency. These error attributions were consistent with prior beliefs. When participants received disconfirming feedback, they correctly attributed 11 of 12 trials to error when they strongly expected the triple to fit the rule beforehand ($P(TFTR) > 0.8$), and incorrectly attributed 15 of 65 trials to error when the strength of their prior beliefs did not justify attributing the feedback to error, ($P(TFTR) < 0.8$), $\chi^2(1) = 23$, $p < 0.05$, $\phi = 0.5$. When receiving affirming feedback, they correctly attributed 1 of 4 trials to error when they strongly expected the triple to not fit the rule beforehand ($P(TFTR) < 0.2$), and 4 of 63 incorrectly when the strength of their prior beliefs did not justify attributing the feedback to error ($P(TFTR) > 0.2$), $\chi^2(1) = 1$, $p = 0.31$, $\phi = 0.12$. However, in multiple regression, there was a main effect of feedback type on attributions of error ($t(159) = 3$, $p = 0.0095$), no significant main effect of Bayes' Rule requiring error attribution ($t(159) = 1.2$, $p = 0.37$), and no interaction between the two factors ($t(159) = 1.4$, $p = 0.3$). The overall correlation between their judgments and the consistency criterion was $\phi = 0.38$, $\chi^2(1) = 23$, $p < 0.05$.

Accuracy. Although error attributions were consistent with prior beliefs, they did not match actual error. When participants believed that feedback was false, it was as likely to be accurate as inaccurate (23% vs. 29%), $\chi^2(1) = 2.6$, $p = 0.35$, $\phi = 0.11$.

Data Sharing. Participants were as likely to share data when feedback affirmed their hypothesis as when it did not, (40%; $SE = 11\%$ vs. 34%; $SE =$

12%), $t(122) = 0.48$, $p > 0.05$. They were also equally likely to share feedback when they saw it as accurate or inaccurate (40%, $SE = 10\%$ vs. 32%, $SE = 10\%$), $t(122) = 0.53$, $p > 0.05$. When including both main effects and the interaction between actual error and attribution of error to predict whether each trial would be shared, there was neither a significant main effect of error attribution ($t(119) = 0.55$, $p = 0.68$), actual error ($t(119) = 0.045$, $p = 0.8$), or an interaction between the two factors ($t(119) = 0.19$, $p = 0.78$).

Discussion

The results replicate the findings of Gorman (1989) and Penner and Klahr (1996), who found that people are more likely to question feedback when it disconfirms their hypothesis. For attributions to error, most of these judgments were normatively justified, matching the Bayesian consistency criterion on 92 of 144 trials (64%). In spite of this consistency, participants were unable to identify actual errors. Finally, on a task new to this study, participants shared information at equal rates regardless of whether the feedback was affirming or disconfirming and regardless of whether it was attributed to error.

The positive test strategy (Klayman & Ha, 1987) entails seeking affirming evidence and discounting disconfirming evidence. Experiment One found that this strategy is both internally consistent and inaccurate. Participants were, however, no less likely to share disconfirming or seemingly flawed data. Although this pattern of data sharing contradicts the positive test strategy, we observed that some participants had difficulty interpreting the open-ended data-sharing question. Namely, when asked which trials they wanted to share, some responded with a triple (e.g., “2, 4, 6”), rather than a trial (e.g., “trial 3”). Experiment Two addresses this problem by using a fixed response format after each trial rather than an open-ended one at the end.

Experiment Two

Experiment One replicated the positive test strategy found in previous studies, with participants invoking error more often for disconfirming feedback. These attributions to error were justified in terms of the consistency criterion, but not the accuracy criterion. The second experiment examines the effects of incentives on these judgments, using a monetary payoff that encourages participants to convince themselves that they know the rule. Specifically, participants were offered $100 for guessing the Actual Rule correctly, and $1 for concluding that they do not know it. This incentive scheme sought to encourage motivated reasoning, so that hopeful participants believe that they've reached a correct conclusion rather than assess their knowledge candidly.

Experiment Two also improves the data-sharing decision. Immediately after receiving feedback, participants make a binary (Yes–No) decision about whether each trial should be shared. Having data-sharing decisions at the end of each trial rather than at the end of the experiment sought to make it clearer that the sharing decision applies to the current trial, resolving any ambiguity about whether triples or trials should be shared. It also elicits sharing judgments earlier in the task, before participants might become tired or frustrated.

Method

Participants. Fifty-eight Carnegie Mellon University undergraduates participated in the experiment for course credit. There were thirty-four women, with average age of 20 years (range: 18 – 24).

Design. Participants were randomly assigned to either the control or the incentive condition. This was a one-way between-subjects design with two levels.

Procedure. The entire experiment lasted 30 minutes. Participants were given informed consent, instructions, and the response sheet. At the end of the experiment, they were asked to leave their email address, with the promise that they would be contacted later if they had solved the rule to receive their bonus payment. This delay of bonus payment was done to prevent participants from telling their friends the correct answer.

Materials. All materials were the same as those in the first study except for the following three changes. First, participants were given the spreadsheet, but were not required to use it.

Second, in the financial incentive condition, participants were told:

> "At the end of the experiment you will be given a chance to win money by guessing the rule. If you decide to guess the rule you will receive 100 dollars if the guess is exactly correct, but 0 dollars if the guess is incorrect. On the other hand, you can decide that you do not know and receive 1 dollar for sure."

Third, decisions to share a trial were made immediately after participants made their attributions to error:

> "We are also interested in how people share information. In a future experiment, a new participant will try to discover the same Actual Rule

> that you are trying to discover. You can share information with this new participant to help him or her solve the Actual Rule. If the new participant solves the rule, you will receive an additional 50 dollars."

The trials were described in the same way as Experiment One, but the sharing judgment was now binary:

> "Do you think this trial should be shared with a new participant? (Yes/No)"

Results

Incentives and Performance. Incentives doubled the median number of trials from 4.5 to 9.[2] Using the same scoring method as Experiment One, those in the incentive condition scored about the same on average ($M = 1.58$, $SD = 1.21$) as those in the control condition ($M = 1.66$, $SD = 1.21$), $t(56) = 0.80$, $p > 0.05$. One participant solved the rule exactly, and was compensated with a $99 Amazon gift card.

Attributions to Error. As in Experiment One, those in the control condition were significantly more likely to see feedback as error when it was disconfirming (29%, $SE = 5.6\%$), than when it was affirming (10%, $SE = 4.1\%$), $t(171) = 2.89$, $p < 0.05$, $d = 0.22$. In contrast, participants in the incentive condition were equally likely to attribute error to disconfirming feedback (16%, $SE = 3.8\%$) and to affirming feedback (20%, $SE = 6.6\%$), $t(309) = 0.62$, $p > 0.05$, $d = 0.04$. Thus, although we expected the incentives to increase motivated reasoning, they appeared to reduce the tendency for participants to attribute disconfirming results to error.

In multiple regression, there was a significant main effect of feedback type ($t(476) = 2.5$, $p = 0.038$), incentive ($t(476) = 2.4$, $p = 0.05$), and a significant interaction between the two factors, where disconfirming feedback only increased error attributions for those in the control condition ($t(476) = 3.1$, $p = 0.0069$). There were no other main effects, two-way, or three-way interactions between feedback type, incentive condition, and actual error.[3]

Bayesian Consistency. Attributions to error for participants in the control condition were consistent with their prior beliefs. For affirming feedback, they correctly attributed 3 of 10 trials to error and incorrectly attributed 0 of 47 trials to error, $\chi^2(1) = 3.9$, $p = 0.057$, $\phi = 0.24$.[4] For disconfirming feedback they correctly attributed 9 of 13 trials to error and incorrectly attributed 16 of 82 trials to error, $\chi^2(1) = 10$, $p < 0.05$, $\phi = 0.32$. The overall correlation between their attributions to error and the consistency criterion was $\phi = 0.24$, $\chi^2(1) = 10$, $p = 0.0015$.

Participants in the incentive condition exhibited similar consistency. For affirming feedback, they correctly attributed 12 of 38 trials to error and incorrectly attributed 15 of 76 trials to error, $\chi^2(1) = 2.9$, $p = 0.088$, $\phi = 0.16$. For disconfirming feedback they attributed 11 of 31 trials to error correctly and incorrectly attributed 22 of 155 trials to error, $\chi^2(1) = 16$, $p < 0.05$, $\phi = 0.29$. The overall correlation between their error attributions and the consistency criterion was $\phi = 0.19$, $\chi^2(1) = 12$, $p = 0.001$.

Accuracy. As in Experiment One, participants in the control condition were unable to identify when actual errors occurred. They correctly identified 26% of actual errors and incorrectly identified 20% of non-errors as error, $\chi^2(1) = 2.3$, $p = 0.36$, $\phi = 0.094$. For the incentive condition, participants were also unable to identify actual errors. They correctly identified 21% of actual errors and incorrectly identified 21% of non-errors as error, $\chi^2(1) = 1.3$, $p = 0.46$, $\phi = 0.054$.

Data Sharing. In contrast to Experiment One, participants in the control condition shared a smaller proportion of trials when the feedback was disconfirming (84%, $SE = 8.3\%$) than when it was affirming (93%, $SE = 5.6\%$), $t(171) = 1.96$, $p = 0.05$, $d = 0.15$. Similarly, they shared a smaller proportion of trials when they judged the feedback to be an error (79%, $SE = 12\%$) than when they judged it to be accurate (91%, $SE = 6\%$), $t(171) = 1.98$, $p < 0.05$, $d = 0.15$. Participants in the incentive condition also shared a smaller proportion of trials when the feedback was disconfirming (84%, $SE = 6.3\%$) than when it was affirming (94%, $SE = 3.6\%$), $t(309) = 2.95$, $p < 0.05$, $d = 0.17$. They also shared a smaller proportion of trials when they judged the feedback to be an error (71%, $SE = 12\%$) than when they judged the feedback to be accurate (91%, $SE = 3.8\%$), $t(309) = 3.94$, $p < 0.05$, $d = 0.22$.

In multiple regression, there was only a significant main effect of error attribution ($t(476) = 2$, $p = 0.053$), and a marginally significant interaction between actual error and incentive condition, such that those in the incentive condition were more likely to share actual errors than those in the control condition ($t(476) = 1.7$, $p = 0.086$). There were no other main effects, two-way, or three-way interactions between feedback type, actual error, and incentive condition.

Discussion

Experiment Two again found that participants more often attribute error to disconfirming feedback when given no incentive beyond their intrinsic motivation to solve the problem. However, participants who were offered a large incentive for getting the rule attributed error to affirming and disconfirming feedback at equal rates. Although we had expected the incentive for getting the rule to increase motivated reasoning, it actually reduced the tendency for participants to attribute disconfirming feedback to error. It did not, however, lead to error attributions that were either more accurate or more consistent with prior expectations. Participants in the control condition met the consistency criterion on 125 of 152 trials (82%), which was a higher rate than those in the incentive condition (217 of 300 trials, 72%). One possible explanation is that the incentive helped participants maintain a more balanced perspective on the likelihood of error after receiving feedback; however, in spite of their motivation, they lacked the understanding (e.g., of Bayes' Rule) needed to respond consistently. An alternative explanation is that participants in the incentive condition rushed through the prior probability and error attribution questions in order to complete more trials, thereby creating more chances to propose triples and get feedback. This strategy would reduce consistency and make attributions of error more equal across feedback types, and is consistent with the finding that participants in the incentive condition completed twice as many trials in the same time period as those in the control condition.

For both the control and the incentive groups, participants shared disconfirming feedback less frequently than affirming feedback. They also shared feedback that they attributed to error less frequently than feedback that they saw as accurate. Those error attributions were loosely justified by internal consistency, but not by accuracy. Extrapolating to scientific contexts, researchers may have defensible reasons to omit data from publication based on their expectations, but this consistency may not prevent harm to those who must use the data. Before reaching that conclusion, we address one possible artifact in Experiment Two's procedure: placing the sharing decision immediately after the error attribution task, perhaps suggesting that the two should be related. Experiment Three remedies this possible confound by eliciting data sharing decisions and error attributions both during each trial and at the end of the task, also allowing participants to reflect on all the data before making their final error attributions and data-sharing decisions.

Finally, Experiment Two's incentive scheme sought to motivate participants to believe they knew the rule. However, the value of data is usually determined not by the person who collects the data themselves, but by others, such as reviewers (for journals) or regulatory bodies (for drug approval). These people, who are external to the data collection process, determine the reward to the researcher based on their prior beliefs and their evaluation of the data shared with them. To simulate this incentive system more closely, Experiment Three uses the natural expectations that participants have about how to convince another person. We expect that an incentive to convince another person should increase the preference for discounting disconfirming feedback.

Experiment Three

Experiment Three replicates Experiment Two with several modifications. Most importantly, a new condition provides an incentive for participants to convince another person that their proposed Final Answer is correct, with data-sharing as the sole mode of communication between them. To do this, we embed the Wason task in a teacher–learner game, a type of principal–agent game (Fudenberg &

Tirole, 1991; Shafto, Eaves, Navarro, & Perfors, n.d.). In this task, the participant collecting the data (the teacher) shares data with another person (the learner) who has to guess the rule based on the data that the teacher decides to share.

The teacher is in one of two incentive conditions. The *compatible* incentive condition rewards both the teacher and learner if the learner guesses the rule. In the *perverse* incentive condition, the learner's rewards remain the same, but the teacher receives money if the learner accepts the teacher's Final Answer. Thus, the perverse incentive allows the teacher to distort the data supplied to the learner, potentially increasing her own payoff while reducing the learner's reward. In this scenario, the teacher knows the entire game structure, but the learner does not. Specifically, the learner is not told that the teacher does not have to share all the trials that were conducted, and the teacher is told that the learner only knows about the shared trials.

Experiment Three also deals with two methodological issues that arose in Experiment Two. One is that participants in the incentive condition attributed affirmation and disconfirmation to error equally, but were slightly less consistent in their attributions to error than participants in the control condition. This may have reflected their rushing through the task to complete more trials. To reduce this threat, we use a penalty for making incorrect prior probability and error attributions. Any payoff to the participant is reduced in proportion to their inaccuracy on these two measures. This penalty prevents them from performing one element of the task well (collecting many trials) at the cost of the other elements (rushing through attributions to error). The second was the possibility that participants assumed that the data sharing and error attribution judgments should be related because they occurred sequentially on each trial. This could create a false correlation between the two measures based on the participant's belief that the experimenter put the questions close to each other for a reason. To deal with this, we also elicit data sharing decisions and attributions to error at the end of the task, using a fixed-response format rather than the open-ended format used in Experiment One.

Method

Participants. One hundred Amazon Mturk volunteers completed the task for $5. There were 46 women, with average age of 32 years (range: 18–65).

Design. The design was a 2 level (perverse or compatible incentive) between-subjects design.

Materials. The procedure and materials were the same as in Experiment Two except for the following modifications. First, participants completed three 'practice trials' to help them understand the task. They were then told the following:

> "We are also interested in how people share information. The information comes in trials. A trial is a page where you proposed a triple and received feedback. The practice trials you conducted are shown below. For each trial you share, another person will get the triple you proposed and the feedback you received. The person will also receive the Final Answer you propose at the end of the task, regardless of the trials you share."

Participants were then told about possible bonus money:

> “Both you and the person you share trials with can earn up to a $5 bonus in addition to the $5 you receive for participating in the experiment.”

The perverse incentive condition was followed with this text:

> “How you earn bonus money:
>
> - If the other person thinks your Final Answer matches the Actual Rule exactly, then you get $5.
> - If the other person thinks your Final Answer does not match the Actual Rule at all, then you get $0.
> - If the other person thinks your Final Answer somewhat matches the Actual Rule, then you get somewhere between $0 and $5.”
>
> “How the person you are sharing trials with earns bonus money:
> The person you are sharing trials with can also earn money.
>
> - This person gets the most money ($5) by correctly judging how well your Final Answer matches the Actual Rule.
> - If this person thinks your Final Answer matches the Actual Rule, but it does not, the other person gets less money.
> - If this person thinks your Final Answer does not match the Actual Rule, but it is does, the other person gets less money.”

Those in the compatible incentive condition were told:

> - “If the other person’s guess matches the Actual Rule exactly, then you both get $5.

> - If the other person's guess does not match the Actual Rule at all, then you both get $0.
> - If the other person's guess somewhat matches the Actual Rule, then you both get somewhere between $0 and $5."

Finally, participants were told the penalty for making incorrect attributions:

> "Penalty for wrong answers
>
> Any bonus you get will be reduced if your false feedback and probability judgments are wrong. Thus, to earn the most money you should make your false feedback and probability judgments as accurate as possible."

Results

Incentives and Performance. As in Experiment Two, participants in the compatible and perverse incentive conditions completed a median of about 8 trials (9 and 7, respectively), $t(97)$ = -0.12, $p = 0.91$, d = -0.012.

Attributions to Error. Those in the compatible incentive condition were more likely to see feedback as in error when it was disconfirming than when it was affirming both during (38% vs. 4.8%) and at the end of the task (41% vs. 12%), ($t(510) = 7.7$, $p < 0.001$, $d = 0.34$; $t(525) = 6.5$, $p < 0.001$, $d = 0.29$, respectively). Similarly, both during (39% vs. 9.3%) and at the end of the task (47% vs. 7.9%), those in the perverse incentive condition were significantly more likely to see feedback as in error when it was disconfirming than when it was affirming ($t(537) = 7.3$, $p < 0.001$, $d = 0.32$; $t(504) = 8.5$, $p < 0.001$, $d = 0.38$, respectively).

Bayesian Consistency. For both incentive groups, adding the penalty for incorrect error attributions and probability judgments greatly improved accuracy and consistency, as compared to Experiments One and Two. For the compatible condition, the overall correlation between their attributions to error and the consistency criterion was $\phi = 0.37$, $\chi^2(1) = 76$, $p < 0.001$. Participants in the perverse incentive condition exhibited even greater consistency, $\phi = 0.55$, $\chi^2(1) = 179$, $p < 0.001$.

Accuracy. Participants both in the compatible and perverse incentive conditions were able to accurately identify error during the task ($\chi^2(1) = 79$, $p < 0.001$, $\phi = 0.37$; $\chi^2(1) = 139$, $p < 0.001$, $\phi = 0.5$, respectively). Participants in the compatible incentive group correctly identified 44 of 99 actual errors and incorrectly identified 60 of 458 non-errors as error. For the perverse incentive condition, participants correctly identified 61 of 121 actual errors and incorrectly identified 66 of 474 non-errors as error. This accuracy also slightly improved in judgments made at the end of the task for both the compatible and perverse incentive conditions ($\chi^2(1) = 121$, $p < 0.001$, $\phi = 0.45$; $\chi^2(1) = 168$, $p < 0.001$, $\phi = 0.57$, respectively).

Data Sharing. Participants in the compatible incentive condition shared 147 of 211 trials when the feedback was disconfirming ($76\%, SE = 6.6\%$) and 191 of 203 when it was affirming ($97\%, SE = 1.3\%$), $t(499) = 5.8$, $p < 0.001$, $d = 0.26$. Similarly, they shared 37 of 82 trials when they attributed feedback to error ($53\%, SE = 12\%$) and 369 of 400 when they judged it to be accurate ($97\%, SE = 1.4\%$), $t(570) = 7.6$, $p < 0.001$, $d = 0.32$.

At the end of the task, participants shared 112 of 180 trials that they judged to be an error ($71\%, SE = 17\%$) and 391 of 414 when they judged it to be accurate

$(99\%, SE = 1.4\%)$, $t(515) = 2.5$, $p = 0.036$, $d = 0.11$. However, there was also significant variation across participants in how much data they shared when they perceived the feedback to be an error, $\chi^2(1) = 52$, $p < 0.001$. As can be seen in Figure 1, most participants in the compatible incentive condition shared all of the trials they attributed to error at the end of the task, while a significant proportion shared none of those trials. However, there was no such variation for data sharing in response to disconfirming feedback $\chi^2(2) = 2$, $p = 0.5$, or error attributions during the task, $\chi^2(2) = 1.5$, $p = 0.64$.

Unexpectedly, participants in the perverse incentive condition did not share trials at lower rates than those in the compatible incentive condition. They shared 171 of 218 trials when the feedback was disconfirming $(88\%, SE = 6.5\%)$ and 183 of 218 trials when it was affirming $(97\%, SE = 2.2\%)$, $t(520) = 1.8$, $p < 0.16$, $d = 0.079$. They also shared 68 of 102 trials when they judged the feedback to be an error during the task $(80\%, SE = 10\%)$ and 286 of 334 trials when they judged the feedback to be accurate $(96\%, SE = 2.8\%)$, $t(520) = 2.4$, $p < 0.043$, $d = 0.11$. At the end of the task, they shared 74 of 130 trials that they judged to be an error $(81\%, SE = 20\%)$ and 333 of 354 trials that they judged to be accurate $(100\%, SE = 0.33\%)$, $t(481) = 2.1$, $p = 0.08$, $d = 0.098$.

As seen in Figure 1, there was significant variation across participants in their decisions to share data after receiving disconfirming feedback, $\chi^2(1) = 7$, $p = 0.083$, whether they shared data that they perceived to be error during the task, $\chi^2(1) = 19$, $p = 0.029$, or whether they shared data that they perceived to be error at the end of the task, $\chi^2(1) = 27$, $p < 0.001$. For all three judgments, most participants in the perverse incentive condition shared all of their trials.

Our prediction was that some participants would be seduced by the perverse incentive, thus deciding only to share trials that were consistent with their final answer. However, there was no difference between conditions in the probability of omitting data that were inconsistent with their final answer, $t(999) = 0.13$, $p = 0.79$. A second way that participants could produce these results while exploiting the perverse incentive would be to seek out only affirming data, knowing that the data would make a simple and convincing story. One way to implement this weak testing strategy is to propose the (2,4,6) triple, knowing that they would receive affirming feedback unless the feedback is in error. However, participants in the two incentive conditions were equally likely to propose (2,4,6) triples, $t(1154) = 0.59$, $p = 0.67$.

As participants were both accurate and consistent in their error attributions, they may have been able to remove actual errors from the data they shared. Overall, at the end of the task participants shared 63 of 118 (53%) trials that were both actual errors and perceived as errors, 62 of 65 (95%) trials that were actual errors but not perceived as errors, 106 of 163 (65%) trials that were perceived as errors but not actual errors, and 615 of 650 (95%) trials that were neither perceived as error nor actual error. When including both main effects and the interaction between actual error and attribution of error to predict whether each trial would be shared at the end of the task, there was only a significant main effect of error attribution, and not actual error, for both compatible and perverse conditions ($t(509) = 4.7$, $p < 0.001$ vs. $t(479) = 5.2$, $p < 0.001$, respectively). This means that error attributions, but not actual errors, mattered in determining whether data is shared.

The reason perceived and actual errors diverged was that disconfirmation had a systematic and additive effect on perceived error, even after controlling for actual error. Main effects of both actual error ($t(1025) = 7$, $p < 0.001$) and disconfirming feedback ($t(1025) = 7.3$, $p < 0.001$) increased the chance of attributing a trial to error at the end of the task, with no significant interaction between the two ($t(1025) = 1.6$, $p = 0.23$). Thus, affirming trials were shared more often, as they were less likely to be perceived as errors than disconfirming trials even when they were actually errors, whereas disconfirming trials were shared less frequently because they were inappropriately seen as errors when they were not.

Discussion

Participants with a compatible or perverse incentive to share data were equally likely to attribute disconfirming feedback to error. The financial penalty for making incorrect probability judgments and attributions to error produced greater consistency and accuracy, compared to Experiments One and Two. Participants in both incentive conditions also shared fewer trials whose feedback was disconfirming or attributed to error, either during or at the end of the task. Although participants were successful in identifying actual errors, it was attributions to error that determined whether they shared trials, indicating that being able identify error does not preclude failing to share trials with accurate disconfirmations, while sharing ones with inaccurate affirmations.

We expected the perverse incentive to reduce the consistency and accuracy of attributions to error, as well as to reduce the sharing of data attributed to error. However, such motivated reasoning was not observed. Rather, data sharing behavior in the two conditions differed in an unexpected way. Both for decisions made after each trial and at the end of the task, participants in the perverse incentive condition shared *more* data than those in the compatible incentive condition–thereby demonstrating a more ethical data sharing stance. While it is possible that higher stakes, such as those involved in pharmaceutical or academic research, would lead to motivated reasoning and data sharing policies, participants responded to the moderate stakes used in this research with reasoned and ethical behavior.

In decisions made at the end of all trials, however, some participants in the perverse incentive condition decided to share none of the data they attributed to error. Contrary to our prediction, these participants did not omit more trials that were inconsistent with their final answer than those in the compatible incentive condition. Additionally, those in the perverse incentive condition did not try to produce a convincing story in as few trials as possible, in order to reduce the risk of collecting inconvenient data that would make their Final Answer less convincing or requiring selective reporting.

There are several possible explanations for why participants in the perverse incentive condition shared trials at a higher rate than those in the compatible incentive condition. First, they may have thought that the learner knows they can hide data, even though the instructions indicated that the other participant would only know about the trials they decided to share. Second, they may have believed that sharing more trials increases the learner's confidence, regardless of whether the trials are consistent with their Final Answer. Third, they may have been more strongly motivated to do the right thing and give the learner all the data available, even if that came at the cost of their own compensation. Open-ended responses at the end of the task show four examples of such motivation:

1. "Yes. / It was an exercise in thinking about probabilities and cooperating with another. "

2. “I think there was some deception involved. This experiment may be about how willing the participant is to share money.”
3. “seems more like a trust then a math problem. ”
4. “I shared everything because, not knowing if the FIT/DNF response by the computer was correct, I didn’t want to deliberately bias the info I passed on by being selective.”

Thus, Experiment Three extends the positive test strategy to communication of results, seen in selective reporting, such that disconfirming data are seen as both caused by error and not worthy of sharing with others. Contrary to our prediction of motivated reasoning (Kunda, 1990), the perverse incentive condition not only did not increase error attributions, but increased the sharing of data that were disconfirming or attributed to error.

Experiment Four

Experiment Three found that participants given a perverse incentive shared more trials that were disconfirming or attributed to error than did participants given a compatible incentive. On the surface it appears that they were genuinely willing benefit others once the incentives increased the importance of doing so, at potential financial cost to themselves. To test this explanation, at the end of Experiment Four all participants were told the Actual Rule and then were allowed to revise their decisions regarding which trials to share (although not to adjust their Final Answer).[5]

If participants care about the success of the person receiving the data, then they should change the data they share to match the correct answer after they learn it. This entails sharing trials that they attributed to error but were not, in particular trials that disconfirm their Final Answer. Likewise, any trials that they thought were not errors but actually were should not be shared. This behavior would be altruistic, as sharing data that are inconsistent with their Final Answer will make it less convinging, likely reducing their own payoff.

Alternatively, if participants care about maximizing their own payoff, then knowing the Actual Rule should make no difference in their data sharing judgments, as they would have already selected the data that is most convincing for their Final Answer before knowing the Actual Rule. Participants who share data that are only consistent with their Final Answer will have no reason to change the data they share. Similarly, participants who shared all of their data because they believed more data is more convincing should also not change the data they share upon learning the Actual Rule. Thus, knowing the Actual Rule should change the data shared by those who care about the payoff of the receiver, but not affect those who only care about their own payoff.

Experiment Four also seeks to rule out two alternative explanations for the data sharing results of Experiment Three. Participants may have thought that the person receiving the data knew that they could hide trials, hence might become suspicious if the data were too orderly. To clarify that this was not possible, we modified the wording to be clear that the other person cannot know that they collected more trials than they shared, if they decide to do so. Second, explicitly mentioning the other person may evoke concerns about the welfare of the person receiving the data. Similarly, data sharing judgments were worded as “information sharing,” possibly sending the message to participants in the perverse incentive condition that they should share, rather than hide, data. To control this, we used more neutral language, where all mentions of ‘sharing’ were changed to ‘communicate’, and any mention of the other participant was removed, wherever possible. This wording sought to allow participants to decide whether the data communication was about sharing or deception themselves, rather than try to infer what the experimenter wants (Orne, 1962; Weber & Cook, 1972; Nichols & Maner, 2008).

Method

Participants. One hundred twenty three Amazon Mturk volunteers completed the task for $5. There were 61 women, with an average age of 32 years (range: 18–69).

Design. The design was a 2 level (perverse or compatible incentive) between-subjects design.

Materials. The procedure and materials were the same as in Experiment Three except for the following modifications. Most importantly, participants were given the Actual Rule after they gave their final answer, and were allowed to change their data sharing judgments.

> “In words, the actual rule is ascending consecutive even numbers from 2-100.
>
> In math, the actual rule is:
>
> - There are three numbers, call them num1, num2, num3, in order.
> - $num2 - num1 = 2$

- $num3 - num2 = 2$
- $num1$ is *even*
- $num1 > 2$
- $num3 < 100$

If you want, you can now decide to change the trials that you communicate below." [6]

Next, all mentions of 'sharing' were changed to 'communication'. The data sharing task was described as follows:

> "For each trial you communicate, another person will get only the triple you proposed and the feedback you received, nothing else.
>
> The person will also receive the Final Answer you propose at the end of the task, regardless of the trials you communicate. However, this person will only know about the trials you communicate, and does not know how many trials you completed or that you did not have to communicate all of your trials."

All data-sharing judgments were changed to the following wording:

Do you think this trial should be communicated?

Results

Attributions to Error. There were main effects of both actual error ($t(1365) = 7.7$, $p < 0.001$), and feedback ($t(1365) = 6.1$, $p < 0.001$), on attributions to error, with an interaction between actual error and incentive, such that actual error had less of an influence on attributions to error in the perverse incentive condition than compatible incentive condition ($t(1365) = 2.7$, $p = 0.011$). For attributions at the end of the task, those in the perverse incentive condition attributed affirming feedback to error less than those in the compatible incentive condition (3.8% vs. 9.9%), but the reverse was true for disconfirming feedback (43% vs. 38%), ($t(1369) = 3.7$, $p < 0.001$, $d = 0.099$).

Bayesian Consistency. For the compatible condition, the overall correlation between attributions to error and the consistency criterion was $\phi = 0.56$, $\chi^2(1) = 246$, $p < 0.001$. Participants in the perverse incentive condition exhibited similar consistency, $\phi = 0.56$, $\chi^2(1) = 205$, $p < 0.001$.

Accuracy. Participants both in the compatible and perverse incentive conditions were able to accurately identify error during the task ($\chi^2(1) = 109$, $p < 0.001$, $\phi = 0.39$; $\chi^2(1) = 111$, $p < 0.001$, $\phi = 0.43$, respectively).

Data Sharing. Participants in the compatible and perverse incentive conditions shared trials at similar rates. They shared fewer trials (69% vs. 73%) when the feedback was disconfirming than affirming (93% vs. 95%), a main effect of feedback only ($t(1072) = 7.2$, $p < 0.001$, $d = 0.22$), and no other main effects or interactions. During the task they shared fewer trials that they attributed to error (37% vs. 38%) than they saw as accurate (94% vs. 97%) a main effect of attribution only ($t(1073) = 10$, $p < 0.001$, $d = 0.31$), and no other main effects or interactions. At the end of the task they shared fewer trials that they attributed to error (56% vs. 55%) than trials they saw as accurate (94% vs. 97%) a main effect of attribution only ($t(1478) = 11$, $p < 0.001$, $d = 0.27$), and no other main effects or interactions.

Both conditions also exhibited the bimodal sharing pattern as in Experiment Three, with some participants sharing all data attributed to error at the end of the task, and others sharing none ($\chi^2(1) = 283$, $p < 0.001$), and no interaction with incentive condition.

Differing from Experiment Three, both attribution to error and actual error determined data sharing, with no interaction with the incentive. There was a significant main effect of attribution to error ($t(1217) = 6.9$, $p < 0.001$) and actual error ($t(1217) = 4.5$, $p < 0.001$). Overall, at the end of the task participants shared 61 of 154 (40%) trials that were both actual errors and perceived as errors, 71 of 85 (84%) trials that were actual errors but not perceived as errors, 101 of 184 (55%) trials that were perceived as errors but not actual errors, and 825 of 905 (91%) trials that were neither perceived as error nor actual error.

Data Sharing with Full Information. After learning the Actual Rule, participants in the perverse incentive condition did not share trials that they knew to be errors at a higher rate than did those in the compatible incentive condition, $t(1185) = 0.67$, $p = 0.32$. Thus, they did not behave in an actively deceptive manner by communicating results that they knew to be false. Instead, participants in the perverse incentive condition were more likely to share trials overall ($t(1187) = 1.9$, $p = 0.071$), but less likely to share trials that fit the Actual Rule, a marginally significant interaction ($t(1187) = 2$, $p = 0.054$).

After being told the Actual Rule, those in the perverse incentive condition shared *more* trials that fit the Actual Rule but were inconsistent with their Final Answer than those in the compatible condition, a significant three-way interaction, $t(1131) = 2.3$, $p = 0.027$.

Discussion

As in Experiment Three, participants attributed disconfirming feedback to error more than affirming feedback. However, those in the perverse incentive condition attributed affirming feedback to error at a lower rate, and disconfirming feedback at a higher rate, than those in the compatible condition. Although this interaction was small, likely due to chance, it may reflect the perverse incentive blinding participants to finding fault in affirming feedback while making them more sensitive to finding fault in disconfirming feedback. Supporting this interpretation, those in the perverse incentive condition were less sensitive to actual error when making their attributions. Although these two findings can be interpreted as evidence of motivated reasoning, neither effects were large enough to undermine either Bayesian consistency or accuracy.

As in Experiment Three, participants in both conditions shared fewer disconfirming trials, fewer trials attributed to error during the task, and fewer trials attributed to error at the end of the task. Likewise, participants in both treatment groups exhibited a bimodal data sharing pattern of trials attributed to error at the end of the task, either sharing all or none of these data. These data sharing judgments depended on attributions to error even after controlling for actual error.

In response to a task that was new to Experiment Four, when participants were given full information about the Actual Rule, those in the perverse incentive condition shared fewer trials that fit the Actual Rule than those in the compatible incentive condition, but did not share trials that they knew were errors at a higher rate. Consistent with the results of Experiment Three, participants in the perverse incentive condition who knew a triple fit the Actual Rule but did not fit their Final

Answer shared those trials at a higher rate than those in the compatible incentive condition. Whether explained by reactivity, ethics, or something else, there was no evidence that participants took advantage of the perverse incentive to be deceptive by sharing less data. This indicates that it was not uncertainty or partial knowledge that led to greater data sharing. The perverse incentive did not lead to data sharing policies that a simple game-theoretic analysis would suggest.

In sum, participants given a perverse incentive either implicitly or explicitly considered the welfare of the person receiving the data, and decided to share data to help the other participant. This stands in contrast to the perception of biased researchers who dispose of data that would harm their company's (or their own) profit (Angell, 2000; Bodenheimer, 2000; Nathan & Weatherall, 1999; Weatherall, 2000). Instead, participants with and without an incentive to deceive another person shared trials that they attributed to error less than those that they saw as accurate, even after controlling for actual errors. Instead of motivated reasoning, in both Experiments Three and Four those in the perverse incentive condition performed slightly, but not significantly, better in terms of both Bayesian consistency and accuracy. Thus, we find strong evidence for a 'cognitive' file-drawer problem, and weak or no evidence for a motivational one.

General Discussion

Over 50 years of psychological research has found that hypothesis testing follows a positive test strategy (Klayman & Ha, 1987), whereby people collect data that they expect to affirm their expectations and discount disconfirming data, should it nonetheless reach them. The present study asks how the positive test strategy affects data sharing. We use the Wason 2-4-6 rule discovery task (Wason, 1960), adding the possibility of error to simulate the uncertainty of actual research (Penner & Klahr, 1996). In this task, participants seek to discover a rule by conducting ‘experiments’ to test their hypotheses about its answer, then receive affirming or disconfirming feedback, known to have a 20% error rate. We extended the task by adding several incentive schemes, then examining their effects on participants’ decisions about sharing the feedback they received with another person. We also evaluated participants’ performance in terms of the accuracy and consistency of their judgments of whether the feedback is error.

Experiment One replicated the pattern of results from previous studies, finding that disconfirming feedback is attributed to error more often than is affirming feedback (Penner & Klahr, 1996). A new result is that participants’ error attributions were generally consistent with their prior beliefs, in the sense of their being more likely to attribute affirmative feedback to error when they had strongly expected that the triple would not fit the rule, and being more likely to attribute disconfirming feedback to error when they had strongly expected the triple to fit the rule. However, their judgments of whether the feedback was in error were unrelated to its accuracy. Whether they shared trial results was unrelated to whether the feedback was disconfirming or attributed to error.

Experiment Two replicated Experiment One along with a new condition that provided participants with a large financial incentive for discovering the rule. As in Experiment One, participants attributed disconfirming feedback to error at a greater rate than affirming feedback in the control condition, but not the incentive condition. Those in the control condition again made error attributions that were somewhat consistent with their expectations but were quite inaccurate. In contrast, participants in the incentive condition were neither consistent nor accurate. Experiment Two elicited data sharing decisions after each trial, using a fixed-response format, unlike Experiment One which asked a single open-ended question at the end. Participants in both conditions were more likely to share feedback if it was affirming and perceived to be accurate.

Experiment Three introduced two incentive schemes for sharing data: (a) *compatible* incentives rewarded the sharer and receiver based on the receiver's success; (b) *perverse* incentives rewarded the sharer based on whether the receiver believed that the problem had been solved, and did not disclose when data were not shared. Both conditions penalized participants for making inaccurate probability and error judgments. As before, participants in both conditions were more likely to attribute feedback to error when it was disconfirming. The penalty increased both the accuracy and consistency of error attributions for participants in both conditions, compared to Experiments One and Two. Contrary to prediction, participants with the perverse incentive shared more trials that were disconfirming or attributed to error than did participants with the compatible incentive. In both conditions, despite these participants' ability to identify error feedback, their perception of error was more important than actual error in determining their data sharing.

Experiment Four replicated Experiment Three and used three additional controls. Most importantly, uncertainty about the validity and usefulness of each trial was resolved by giving participants the Actual Rule at the end of the task, allowing them to change the data they shared but without letting them change their

Final Answer. Additionally, demand characteristics were controlled by changing the 'data sharing' wording to 'communication', and participants were explicitly told that the person receiving the data could not know that they chose not to share trials. Experiment Four again found that participants in both incentive groups attributed disconfirming feedback to error and shared perceived errors at a lower rate. Again, those in the perverse incentive condition did not exhibit motivated reasoning in terms of reduced Bayesian Consistency or accuracy, and shared more trials than those in the compatible incentive condition that were consistent with the Actual Rule but inconsistent with their Final Answer.

Four experiments provide support for the following interpretation of human error identification and communication with and without pressures to publish. In terms of error identification, people naturally attribute disconfirming feedback to error, and this is above and beyond what is justified by actual error. However, when penalized for inaccurate judgments, their judgments of error are strongly correlated to actual errors, and they do behave in a consistent (Bayesian) manner. Thus, people were biased but Bayesian. In terms of communication, participants consistently saw trials that they perceived to be errors as less worthy of sharing with another person, even after controlling for whether these trials were actual errors. There was no evidence of motivated reasoning or self-deception, either in terms of misattributing disconfirming results to error, undermining Bayesian consistency, or not sharing unwanted results that could reduce one's financial profit. In fact, participants frequently exhibited 'reactive' (Dillard & Shen, 2005) or 'ethical' behavior when confronted with a perverse incentive that may have been seen as encouraging deception.

Prior to this research there was no evidence on whether laypeoples' (Penner & Klahr, 1996; Gorman, 1986, 1989) attributions of disconfirmations to error are biased. Experiments One and Two found that people have poor error identification ability when no financial penalties are present, but Experiments Three and Four demonstrate both substantial accuracy in identifying error as well as Bayesian consistency when participants were threatened with financial penalties. The success of financial penalties is consistent with research showing that making people accountable for their judgments can substantially improve the quality of their judgments (Tetlock & Kim, 1987).

Incentives for rule discovery or convincingness neither improved nor harmed performance on the task, did not distort error attributions and probability judgments, and did not lead to deceptive data sharing. These results are surprising, as participants had both of the necessary conditions to deceive themselves and others: a motivation to reach a particular conclusion (Kunda, 1990) and justifications for doing so in the form of the possibility of error (Gino, Schweitzer, Mead, & Ariely, 2011; Gino et al., 2009). One difference between this and previous research is that participants may have cared about the performance of the person receiving the data. In our research, there was always someone who would receive the data who could potentially benefit. In deception and cheating research (Gino et al., 2011, 2009), the only person that is harmed by the deception is the experimenter, in the sense that the expermenter pays participants for work they didn't do or skills they don't have. When sharing data with the scientific community, the people who are helped or harmed are somewhat diffuse and nameless. Thus, a crucial determinant of ethical data sharing policies may be whether one considers who is receiving and using the data: if the receiver is made clear, then people will share data ethically out of concern; if the receiver is opaque, then people will fail to share data as they do not think about how others can be harmed.

These experiments also found two unexpected results. First, although participants shared trials they believed to be error at a lower rate than those they thought were accurate, they still shared a substantial proportion of trials they thought were errors. Thus, although we provide the first data on an important determinant of data sharing—whether the data are perceived to be error—there are still unanswered questions on why people would share data they believe to be error at all. Second, Experiments Three and Four found a very strong ability for participants to objectively identify error and act in a Bayesian manner. The Wason rule-discovery task, as with all scientific tasks, has a logically infinite space of potential solutions. Even in the face of this impossible task, in the sense of being unsolvable by a Bayesian agent (although some have tried (Austerweil & Griffiths, 2008)), participants were able to demonstrate substantial accuracy and Bayesian consistency. How they do this merits further research, as it can lend insight into how scientists identify error when they also face logically infinite possibilities.

The circumstances of the experiments differ from those of working scientists in several ways. First, scientists never know the exact error rates in their experiments, but have, instead, just a range of plausible values based on their experience and intuition. Those ambiguous error rates may be more readily modified to fit results than the fixed ones used in the experiments. Second, although the patterns observed here generally parallel those observed in real labs (Dunbar, 1995), the participants were either undergraduates or MTurk respondents, not scientists. One critical difference between experts and novices is that experts are better able to ignore irrelevant information (Shanteau, 1992). Thus, the training of working scientists may allow them to identify and report only accurate data, especially when their knowledge is built on firm theoretical ground (Dawes, Faust, & Meehl, 1989), appropriately omitting errors that would confuse readers.

However, empirical evidence suggests that the results from these four experiments can generalize to scientists, and importantly, our own behavior when conducting research. The most significant benefit of expertise on judgment comes from having theoretical knowledge (Dawes et al., 1989), but this knowledge does not help when unexpected results occur, as this explicitly means that theory, experiment, or both are wrong.

Since the ability to use theory-mediated judgment and the provision of feedback are limited or impossible, scientists and laypeople must rely on their intuition. In this case, evidence suggests that experts and laypeople perform similarly on tasks that involve intuitive judgment, for example in identifying omissions in a fault tree (Fischhoff, Slovic, & Lichtenstein, 1978). Although experts are more consistent in applying judgment policies than laypeople, they do not exhibit non-linear or configural patterns, do not match the environment better, and do not perform better overall than non-experts (Karelaia & Hogarth, 2008).

Most importantly, psychologists who have observed scientists (Dunbar, 1995) find the same tendency to attribute disconfirming results to error as those who have examined laypeople (Gorman, 1989; Penner & Klahr, 1996), including our experiments. The convergent finding is that both expert scientists and laypeople attribute unexpected results to error. The present research extends this, showing that laypeople see unexpected results as both due to error and not worth sharing with others. Generalizing from this previous convergence, we can expect that scientists identify error and share data similarly to laypeople.

The difficulty participants had when trying to avoid sharing faulty feedback shows that trying to be helpful by selectively reporting only results that one believes are accurate is not easy. One strategy participants could have used to achieve accurate selective reporting would be to use exact replications. Participants in all three experiments did not have the perfect accuracy in error attributions that would be required to selectively exclude errors from shared data. At the end of the task, exact replications would allow participants to clearly identify which trials were error and which were accurate, and, in turn, selectively report only accurate data.

Similar policies can help real scientists share data. Experiment Three found that penalties for incorrect probability judgments and error attributions greatly increased consistency and accuracy. One way to implement such a penalty would be to require that statistical analyses and experimental methods presented in published reports provide enough detail, in the paper or ancillary material, to be reproducible—with appropriate professional penalties for those who fail. As a protection, researchers can adopt the protocols of impartial organizations dedicated to independent replication of experiments and analyses (e.g., `https://www.scienceexchange.com/`). Another way of improving error identification is to encourage researchers to complete exact replications. These replications allow researchers to identify errors with high accuracy and make selective reporting of perceived errors highly accurate.

Without corrective measures that go beyond financial penalties, we can expect our file-drawers to contain accurate disconfirmations that would help us ignore false hypotheses, and the published literature to contain inaccurate affirmations that lead us to believe these false hypotheses. Ignoring what is in the file-drawer and attending only to what is not can only harm scientific progess, in the best case by wasting the time used to produce the ignored research, and in the worse case leading entire disciplines to pursue false hypotheses and miss true ones. In the former case, science will not progress, and we will lose the public's trust and willingness to pay for our work. In the latter case, we will actively pursue misleading research, a situation that is not new to science as in the cases of cold fusion and wormrunning (Collins & Pinch, 1998), as well as the discovery of n-rays (Wood, 1904; Klotz, 1980). Awareness of the cognitive processes that lead us to see disconfirming results as error can help us consider ways that our judgment may be wrong, allowing value, to ourselves and others, to be seen in sharing disconfirmations (Collins, 2003).

References

Angell, M. (2000). Is academic medicine for sale? *New England Journal of Medicine*, *342*(20), 1516–1518.

Austerweil, J., & Griffiths, T. (2008). A rational analysis of confirmation with deterministic hypotheses. In *Proceedings of the 30th annual conference of the cognitive science society* (pp. 1041–1046).

Bodenheimer, T. (2000). Uneasy alliance–clinical investigators and the pharmaceutical industry. *New England Journal of Medicine*, *342*(20), 1539–1544.

Camerer, C., & Hogarth, R. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty*, *19*(1), 7–42.

Collins, H. (2003). Lead into gold: the science of finding nothing. *Studies In History and Philosophy of Science Part A*, *34*(4), 661–691.

Collins, H., & Pinch, T. (1998). *The golem: What you should know about science.* Cambridge Univ Pr.

Dawes, R., Faust, D., & Meehl, P. (1989). Clinical versus actuarial judgment. *Science*, *243*(4899), 1668–1674.

Dillard, J., & Shen, L. (2005). On the nature of reactance and its role in persuasive health communication. *Communication Monographs*, *72*(2), 144–168.

Dunbar, K. (1993). Concept discovery in a scientific domain. *Cognitive Science*, *17*(3), 397–434.

Dunbar, K. (1995). How scientists really reason: Scientific reasoning in real-world laboratories. In R. J. Sternberg & J. E. Davidson (Eds.), *The nature of insight*

(Vol. XVIII, pp. 365–395). Cambridge, MA, US: The MIT Press.

Dunbar, K. (2001). What scientific thinking reveals about the nature of cognition. In K. Crowley, C. Schunn, & O. T. (Eds.), *Designing for science: Implications from everyday, classroom, and professional settings* (pp. 115–140). Lawrence Erlbaum Hillsdale: NJ.

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap* (Vol. 57). Chapman & Hall/CRC.

Fanelli, D. (2010). Do pressures to publish increase scientists' bias? an empirical support from u.s. states data. *PLoS One*, *5*(4), e10271.

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, 1–14.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1978). Fault trees: Sensitivity of estimated failure probabilities to problem representation. *Journal of Experimental Psychology: Human Perception and Performance*, *4*(2), 330-344.

Fudenberg, D., & Tirole, J. (1991). *Game theory.* MIT Press.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models* (Vol. 1). Cambridge University Press New York.

Gelman, A., Su, Y., Yajima, M., Hill, J., Pittau, M., Kerman, J., et al. (2010). arm: Data analysis using regression and multilevel/hierarchical models. *R package version*, 1–3.

Gino, F., & Ariely, D. (2012). The dark side of creativity: Original thinkers can be more dishonest. *Journal of Personality and Social Psychology*, *102*(3), 445–459.

Gino, F., Ayal, S., & Ariely, D. (2009). Contagion and differentiation in unethical behavior the effect of one bad apple on the barrel. *Psychological Science*, *20*(3), 393–398.

Gino, F., Schweitzer, M., Mead, N., & Ariely, D. (2011). Unable to resist temptation: How self-control depletion promotes unethical behavior. *Organizational Behavior and Human Decision Processes*, *115*(2), 191–203.

Glucksberg, S. (1962). The influence of strength of drive on functional fixedness and perceptual recognition. *Journal of Experimental Psychology*, *63*(1), 36–41.

Gorman, M. (1986). How the possibility of error affects falsification on a task that models scientific problem solving. *British Journal of Psychology*, *77*(1), 85–96.

Gorman, M. (1989). Error, falsification and scientific inference: An experimental investigation. *The Quarterly Journal of Experimental Psychology*, *41*(2), 385–412.

Gorman, M. (2005). *Scientific and technological thinking.* Lawrence Erlbaum.

Grice, H. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics. vol. 3, speech acts* (pp. 41–58). Academic Press.

Harkness, A., DeBono, K., & Borgida, E. (1985). Personal involvement and strategies for making contingency judgments: A stake in the dating game makes a difference. *Journal of Personality and Social Psychology*, *49*(1), 22–32.

Ioannidis, J., & Trikalinos, T. (2005). Early extreme contradictory estimates may appear in published research: The proteus phenomenon in molecular genetics research and randomized trials. *Journal of Clinical Epidemiology*, *58*(6), 543–549.

Kahneman, D., & Tversky, A. (1996). On the reality of cognitive illusions. *Psychological Review*, *103*(3), 582–591.

Karelaia, N., & Hogarth, R. (2008). Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological Bulletin*, *134*(3), 404–426.

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science*, *12*(1), 1–48.

Klayman, J., & Ha, Y. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, *94*(2), 211-228.

Klotz, I. (1980). The n-ray affair. *Scientific American*, *242*(5), 122–131.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498.

Lichtenstein, S., & Fischhoff, B. (1980). Training for calibration. *Organizational Behavior and Human Performance*, *26*(2), 149–171.

Lichtenstein, S., Fischhoff, B., & Phillips, L. (1981). *Calibration of probabilities: The state of the art to 1980* (Tech. Rep.). DTIC Document.

Lord, C., Ross, L., & Lepper, M. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, *37*(11), 2098–2109.

Nathan, D., & Weatherall, D. (1999). Academia and industry: lessons from the unfortunate events in toronto. *The Lancet*, *353*(9155), 771–772.

Nichols, A., & Maner, J. (2008). The good-subject effect: investigating participant demand characteristics. *The Journal of General Psychology*, *135*(2), 151–166.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning.* Oxford University Press, USA.

O'Connor, R., Doherty, M., & Tweney, R. (1989). The effects of system failure error on predictions. *Organizational Behavior and Human Decision Processes*, *44*(1), 1–11.

Orne, M. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, *17*(11), 776-783.

Pelham, B., & Neter, E. (1995). The effect of motivation of judgment depends on the difficulty of the judgment. *Journal of Personality and Social Psychology*, *68*(4), 581–594.

Penner, D., & Klahr, D. (1996). When to trust the data: Further investigations of system error in a scientific reasoning task. *Memory & Cognition*, *24*(5), 655–668.

Shafto, P., Eaves, B., Navarro, D., & Perfors, A. (n.d.). Epistemic trust: Modeling children's reasoning about others' knowledge and intent. *Developmental Science*.

Shanteau, J. (1992). Competence in experts: The role of task characteristics. *Organizational Behavior and Human Decision processes*, *53*(2), 252–266.

Sterling, T. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance–or vice versa. *Journal of the American Statistical Association*, 30–34.

Tetlock, P., & Kim, J. (1987). Accountability and judgment processes in a personality prediction task. *Journal of Personality and Social Psychology*, *52*(4), 700–709.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

Wason, P. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, *12*(3), 129–140.

Weatherall, D. (2000). Academia and industry: increasingly uneasy bedfellows. *The Lancet*, *355*(9215), 1574.

Weber, S., & Cook, T. (1972). Subject effects in laboratory research: An examination of subject roles, demand characteristics, and valid inference. *Psychological Bulletin*, *77*(4), 273–295.

Wood, R. (1904). The n-rays. *Nature*, *70*, 530–531.

Yaniv, I., & Foster, D. (1995). Graininess of judgment under uncertainty: An accuracy-informativeness trade-off. *Journal of Experimental Psychology: General*, *124*(4), 424–432.

Author Note

All materials and data, including completely reproducible statistical analyses in Sweave, can be obtained from the first author's Dataverse at `http://hdl.handle.net/1902.1/18699`. Open lab notebook can be obtained at `http://openwetware.org/wiki/User:Alexander_L._Davis`. We would like to thank NSF for the dissertation enhancement grant. Thank you to John Sperger and Terence Einhorn for help collecting the data

Footnotes

[1] More precisely, using Bayes' Rule it can be shown that feedback should be attributed to error whenever one believes that there is greater than an 80% prior probability that the triple fit the rule, but the feedback indicates it does not fit, or conversely one believes there is less than a 20% prior probability that the triple fit the rule, but the feedback indicates that it does fit.

[2] Although the median number of trials increased, a non-parametric Kolmogorov-Smirnov (KS) test for differences in empirical cumulative distributions indicates no differences in distribution. Between Experiment One and the control condition of Experiment Two, the KS test was $D = 0.31$, $p = 0.23$. Between Experiment Two control and incentive conditions, the KS test was $D = 0.32$, $p = 0.11$. Thus, although the medians were different, the distributions of trials between the studies and conditions were similar.

[3] This reversal of error attributions may be partially explained by participants inferring that the rule was difficult to solve because the incentive was so large. Participants in the incentive condition expected their triples to fit the rule less often ($M = 0.41$, $SD = 0.36$) than those in the control condition, ($M = 0.49$, $SD = 0.3$), $t(482) = \text{-}2.4$, $p = 0.021$.

[4] A hierarchical model could not be used for the control condition. Only one participant both made an attribution to error and should have not made an attribution to error. Thus, only one subject-level intercept could be fit, as all other participants had zero probability of judging error. To deal with this we pool all of the data to get an approximate answer.

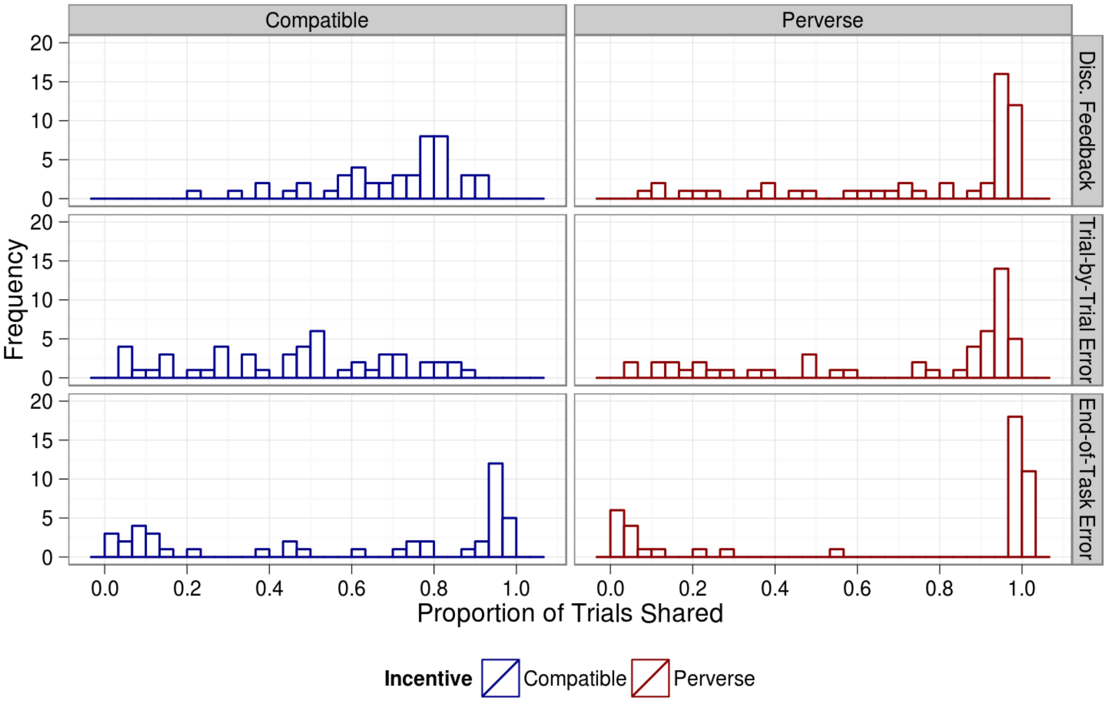
[5] To promote greater exploration of the hypothesis space and give more

disconfirming feedback, a pre-test for Experiment Four added an additional condition to the rule: that only *odd multiples of two* fit the rule. Unfortunately, this did not increase the median number of trials completed (7). All other results replicated: They were more likely to see feedback as in error when it was disconfirming than when it was affirming, ($t(322) = 5.2$, $p < 0.001$, $d = 0.29$; Bayesian Consistency $\phi = 0.45$, $\chi^2(1) = 110$, $p < 0.001$; accuracy ($\chi^2(1) = 48$, $p < 0.001$, $\phi = 0.38$); and data sharing after disconfirmation $t(263) = 3.9$, $p < 0.001$, $d = 0.24$; data sharing after error $t(259) = 4.4$, $p < 0.001$, $d = 0.28$; end of task $t(373) = 4.7$, $p = 0.001$, $d = 0.24$ with significnat heterogeneity $\chi^2(1) = 21$, $p = 0.011$.

[6] The last two constraints should also include an equal to. This may have confused participants. However, alternative analyses taking these differences into accounts did not change the results substantially.

Figure Captions

Figure 1. Proportion of trials shared by whether the trial was disconfirming (top row), whether participants attributed that trial to error during the task (middle row), and whether participants attributed the trial to error at the end of the task (bottom row).


Compatible
Perverse
Disc. Feedback
Trial-by-Trial Error
End-of-Task Error
Frequency
0
5
10
15
20
0.0
0.2
0.4
0.6
0.8
1.0
Proportion of Trials Shared
Incentive
Compatible
Perverse