Supplementary Materials

Measuring the Prevalence of Questionable Research Practices with Incentives for Truth-telling

Leslie K. John, George Loewenstein, Drazen Prelec

Participation Incentive Survey

We conducted a survey to test the effect of the size of the participation incentive on response rates. We emailed an electronic survey to all authors in the journal Organizational Behavior and Human Decision Processes from Nov 2005 (volume 98, issue 2) to May 2010 (volume 112, issue 1) (N = 499).

The survey was the same as the Control condition from the main study. We manipulated the size of the participation incentive: In the "low incentive" condition, participants were simply told that a donation would be made on behalf of each respondent, and that the total donation would be a maximum of $2,000. In the "high incentive" condition, participants were told that a $50 donation would be made on behalf of each respondent.

There were 176 respondents, for a response rate of 35.3% (176/499; age: 5.8% 20-29; 34.5% 30-39 years, 27.3% 40-49 years, 17.3% 50-59 years; gender: 63.3% male). The response rate was not significantly different between solicitation methods (low = 34.2%, high = 37.8%; ns). Because the overall admission rates to the individual items were generally similar to those obtained in the main study, we do not report them here.

Procedure

The survey was emailed to faculty in all psychology departments in the United States that had a doctoral program. Recipients of this solicitation email (hereafter referred to as "recipients") were asked to participate in a "short survey of researchers' personal perceptions of, and experiences with, different research practices."

There were three links in the solicitation email. One was the link to the online survey, which respondents visited to take the survey. The other two were "tracking links" that recorded which recipients had considered participating, while ensuring complete anonymity of respondents' survey responses. Recipients were instructed to click on one of these two links; the first was labeled "I have participated" and the second was labeled "I choose not to participate." Recipients were told that clicking either of the links confirmed that they had considered participating and ensured that they would not be contacted again, while ensuring anonymity of their survey responses should they choose to participate.

Each recipient had a unique identifying code embedded in the tracking links, enabling the solicitation email to be re-sent to recipients who had not yet considered participating (i.e. those who had not clicked on one of the two tracking links). The identifying code was not contained within the link to the survey; this procedure therefore tracked which individuals had taken the survey, while ensuring complete anonymity of their survey responses.

Recipients who did not click on either of the two tracking links were emailed a maximum of two (this number was determined a priori) follow-up emails at approximately biweekly intervals. Participants therefore responded in one of three solicitation waves. We stopped collecting data approximately ten days after the final follow-up email was sent. By this point, the rate of incoming responses had dropped off substantially (from over 100 per day in the days

immediately following the first solicitation email, to on average fewer than one respondent per day).

Upon clicking the link to the survey, respondents were randomly assigned to one of only two conditions which differed in the incentive to give truthful responses (control vs. BTS, described below). Participants were randomly assigned to the BTS vs. control condition at a rate of 2:1. Respondents were told that: "in the following pages, you will be presented with descriptions of different research practices. For each practice, we will ask you:"

1. "To assess the prevalence of the practice by estimating the percent of research psychologists **that have engaged in the practice on at least one occasion."** (*provided on a scale from 0-100% in increments of 1%; hereafter referred to as "prevalence estimate"*)
2. "To estimate, **among the research psychologists who have engaged in this practice** on at least one occasion, the percent that would indicate that they have engaged in the practice." (*provided on a scale from 0-100% in increments of 1%; hereafter referred to as "admission estimate"*)
3. "Whether **you have ever engaged in the practice."** (*provided on a dichotomous response scale labeled yes/no; hereafter referred to as "admission"*)

Participants were told that these three pieces of information would help to develop more accurate estimates of the prevalence of each practice.

In the BTS condition, participants were further told that the three pieces of information would "let us apply a formula called the Bayesian Truth Serum," which would be used "to determine the size of the donation made to the charity that you selected." They were also told that "the important property of the formula is that it rewards truthful answers. This means that truthful answers about your practices will increase the donation made on your behalf (and will also tend to increase the donations made on behalf of other respondents). For the purpose of this survey it is not necessary for you to understand how the formula works, although the theoretical paper from *Science*, which includes a short abstract, is available here (*link to paper*)." To ensure that participants had read and understood this information, they were asked to complete the following instruction quiz: "Giving truthful responses on this survey _____ the amount of money donated to charity on my behalf" by choosing from the response options: *has no impact on, decreases, increases*. Participants who gave incorrect responses were automatically directed back to the instruction page until they answered the question correctly.

Next, participants were presented with the ten items (Table 1) and for each, asked the three questions listed above. Respondents who indicated that they had engaged in the practice were also asked whether they thought it was defensible to have done so, and, if they wished, to elaborate upon why they thought it was (or was not) defensible. The order in which the ten practices were presented was randomized between-subjects.

After providing the three judgments for each of the items, participants were asked whether they had ever had doubts about the integrity of the research done by a) researchers at other institutions, b) other researchers at their institution, c) graduate students, d) their collaborators, and finally, e) themselves. The response scale was labeled: *never, once or twice,*

occasionally, often.. The survey concluded with demographic questions. All demographic variables are listed in Table S1. There were no other dependent measures.

Item selection. Practices identified as questionable by the NIH and used in a previous survey (Martinson, Anderson, & Devries, 2005) were modified to be specifically relevant to behavioral researchers. Initial drafts of the items were informally pre-tested for clarity and relevance with a small group of behavioral researchers. The items were revised based on input from this pre-test.

Data Analysis

Our primary analysis tested the effect of the experimental manipulation on admission rates to the individual items using the likelihood ratio test of the difference in admission rates between conditions. Applying the Bonferroni correction for multiple comparisons, the critical alpha level was adjusted downward to 0.005 (i.e. α/n comparisons = 0.05/10). We also tested the effect of the experimental manipulation on prevalence estimates and admission estimates to the individual items using t-tests (again applying the Bonferroni correction).

In secondary analyses, we tested whether admission rates differed by sub-group using a random effects logistic regression with a fixed effect for inquiry method and a random intercept for subject, with the repeated measures factor (question number) clustered within-subject. These analyses are restricted to the participants who completed the survey, and hence, were asked the demographic questions. For participants who indicated multiple sub-groups, we weighted the dummy variables representing the chosen sub-groups by the number of sub-groups the given respondent had chosen. The baseline group consisted of respondents who did not specify a sub-group.

Finally, we tested whether perceived defensibility differed by sub-discipline using the same approach but using a random effects ordered probit regression. We also used random effects ordered probit regressions for the follow-up survey analyses reported in the paper.

SPSS, Stata, and R (2011) were used to analyze the data.

Supplementary Results

Response rate

Four percent (257/6,221) of email addresses were invalid. There were 2,155 respondents, for a response rate of 36% (2,155/5,964). The response rate declined with each successive email solicitation (wave 1 response rate: 1,258/5,964 = 21%; wave 2 response rate: 572/4,706 = 12%; wave 3 response rate: 325/4,134 = 8%). Although the response rate was fairly low in absolute terms, it was relatively high, given that a) email survey response rates tend to be low and b) the sample consisted of highly time-constrained academic researchers. Moreover, some have argued that response rates may not play as important a role in determining the validity of prevalence estimates as participants' perceptions of anonymity and confidentiality (Bates & Cox, 2008; Fanelli, 2009). And as noted above, we used a sophisticated procedure to ensure that participants' responses were anonymous.

Completion rate

Of respondents who began the survey, 719 (33.4%) did not complete it. There were therefore 1,436 complete response sets, for a completed response rate of 24% (1,436/5,964). Since the questions were presented in random order, data from all respondents – even those who did not finish the survey – were included in the final data set. However, the demographic data are

restricted to the 1,436 individuals who completed the survey (Table S1) since this information was obtained at the end of the survey.

The study's inclusion criteria may have contributed to the relatively high attrition rate. Because we emailed *all* psychology faculty of research-oriented universities, our distribution list likely included many psychologists who do not typically do research (e.g. adjunct faculty). And indeed, some individuals responded to the solicitation email to indicate that they had withdrawn from the survey upon discovering that it did not pertain to them. But, by the end of the survey, when respondents were asked whether they were researchers, 97.7% indicated that they were. Respondents who were not researchers are likely to have dropped out before this point.

The attrition rate was slightly higher in the BTS condition relative to the control condition (BTS: 518/1,488 or 34.8% vs. Control: 201/667 or 30.1%, $\chi^2(1) = 4.53$, $p = .033$). There was neither a main effect of solicitation wave, nor an interaction between inquiry method and solicitation wave, on attrition rates (Wave 1: 413/1,258 = 32.8%; Wave 2: 204/572 = 35.7%; Wave 3: 102/325 = 31.4%; $\chi^2(2) = 2.10$, $p = .351$; $\beta_{wave*bts} = -.174$, SE = .137, $p = .206$).

Although the between-condition difference in attrition rates is very small in magnitude, the asymmetry nonetheless raises the question of whether observed between-condition differences in the dependent measures can truly be attributed to the BTS procedure, as opposed to an idiosyncratic characteristic of respondents who complete the survey. For example, the greater incentive to tell the truth in the BTS condition may have caused respondents who had engaged in the behaviors to be particularly likely to drop out, posing a threat to internal validity. However, an analysis of the point at which respondents dropped out of the survey suggests that the different attrition rates may be due to the instruction quiz (which did not exist in the control condition), and not the BTS truth incentives per se (Figure S1).

First, among respondents who dropped out, the proportion that did so upon viewing the instruction screen – the point at which the truth incentive was described in the BTS condition – was roughly equal between conditions (Figure S1, 1st group of columns; Control = 10.0% vs. BTS = 14.3%, $\chi^2(1) = 2.40$, $p = .122$). And, among BTS drop-outs, a relatively large proportion (24.9%) dropped out during the instruction quiz, suggesting that it was not the BTS procedure per se that intimidated participants (otherwise one would expect drop-out spikes at the instructions and first item), but rather the increased effort required to answer the instruction quiz. Consistent with this explanation, some respondents who failed the quiz and were re-routed back to the instructions emailed us to say that they had dropped out because they thought the survey had a glitch; clearly, these respondents did not understand that the re-routing was intentional.

Second, if respondents dropped out of the BTS condition because they were more intimidated to tell the truth relative to control participants, one might expect an increased proportion of BTS drop-outs upon viewing the first item – the point at which it may have became apparent that the survey was about QRPs. This was not the case; in fact, control participants were *more* likely to drop out at this point (Figure S1, 3rd group of columns): of the 201 control participants who dropped out, 82 (82/201 = 40.8%) did so upon seeing the first item, compared to only 145 of the 518 BTS participants who dropped out (145/518 = 28.0%) ($\chi^2(1) = 11.0$, $p = .001$).

Taken together, these facts suggest that the observed effects of the BTS incentive are unlikely to have been driven by attrition-based individual differences between conditions. And, these facts notwithstanding, such an effect would likely have only made our tests *more conservative* because if anything, subjects intimidated to tell the truth are (arguably) more likely to have engaged in the practice(s). Therefore, to the extent that these subjects dropped out of the

survey, they are likely to have biased our results in the opposite direction of our hypothesis – driving admission rates *downward* in the BTS condition. Finally, it is worth emphasizing that because the order in which the items were presented was randomized between-subjects, partial response sets were included in the analysis of the primary measures (i.e. admission rates, prevalence estimates, admission estimates).

Demographics

Since the demographic questions were posed at the end of the survey, the demographic data are restricted to the 66.6% (1,436/2,155) of respondents who completed the survey. The demographic makeup of the two inquiry conditions was very similar (Table S1), which further helps to rule out mortality bias as an alternative explanation for the results.

Charities

The distribution of the charities to which participants chose to direct their donation was not significantly different between conditions (Save the Children: 32.2%; World Wildlife Fund: 21.2%; American Cancer Society: 25.3%; Multiple Sclerosis Society: 8.9%; Africa Community Health Alliance: 12.3%; $\chi^2(1) = .94$, $p = .92$). Charity choice was not associated with the primary dependent measures (i.e. admission rates, prevalence estimates, admission estimates). A total of $4,219.27 were donated to charity; receipts are posted online at: http://www.cmuresearchsurvey.com/

Supplementary Material References:

Bates, S., & Cox, J. (2008). The impact of computer versus paper-pencil survey, and individual versus group administration, on self-reports of sensitive behaviors. *Computer in Human Behavior*, *24*, 903-916.

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS One, 4*(5), 1-11.

Martinson, B. C., Anderson, M. S., & Devries, R. (2005). Scientists behaving badly. *Nature, 435*, 737-738.

Team, R. D. C. (2011). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Figure Legends for Supplementary Figures

Figure S1. Analysis of the point at which respondents dropped out of the survey, by inquiry method (Main study). Note: The denominator used to calculate the proportion on the y-axis was the total number of respondents within the given condition that dropped out of the survey.

Figure S1

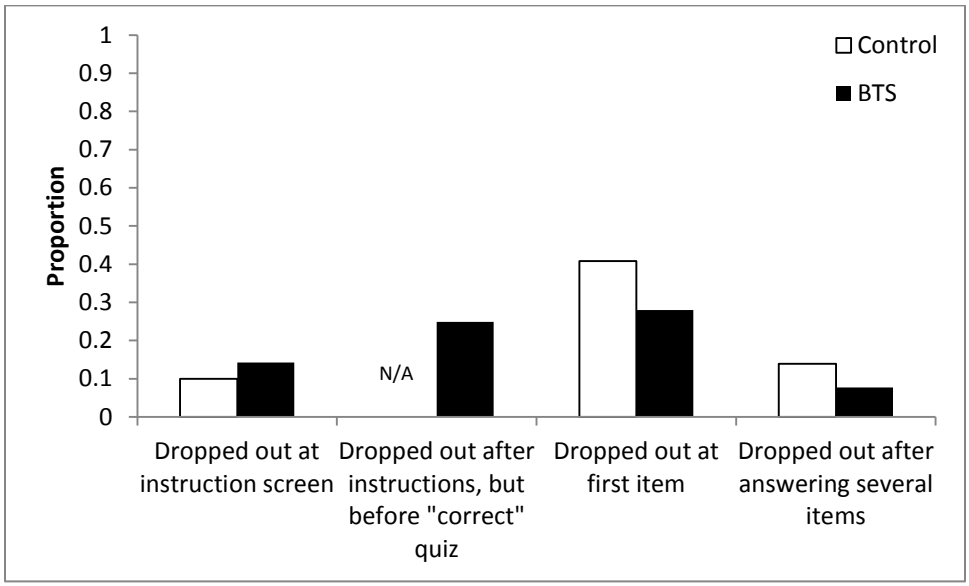
Control
BTS
Proportion
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
N/A
Dropped out at instruction screen
Dropped out after instructions, but before "correct" quiz
Dropped out at first item
Dropped out after answering several items

Table S1. Demographic variables by experimental condition (variables are presented in the order in which they appeared in the survey).

Participant characteristics		Entire sample (N=1436)	Control (n=466)	BTS (n=970)
Psychologist				
	Yes	1356 (94.4)	440 (94.4)	916 (94.4)
Psychologist type				
	Clinical**	315 (21.9)	86 (18.5)	229 (23.6)
	Cognitive**	297 (20.7)	111 (23.8)	186 (19.2)
	Developmental	234 (16.3)	82 (17.6)	152 (15.7)
	Forensic	16 (1.1)	5 (1.1)	11 (1.1)
	Health	71 (4.9)	24 (5.2)	47 (4.8)
	Industrial organizational	58 (4.0)	14 (3.0)	44 (4.5)
	Neuroscience	167 (11.6)	51 (10.9)	116 (12.0)
	Personality	69 (4.8)	22 (4.7)	47 (4.8)
	Social	224 (15.6)	72 (15.5)	152 (15.7)
	Other	192 (13.4)	63 (13.5)	129 (13.3)
Researcher				
	Yes	1377 (97.7)	442 (97.4)	935 (97.9)
Type of Research				
	Behavioral**	755 (52.6)	263 (56.4)	492 (50.7)
	Laboratory*	622 (43.3)	217 (46.6)	405 (41.8)
	Field	468 (32.6)	148 (31.8)	320 (33.0)
	Experiments	845 (58.8)	284 (60.9)	561 (57.8)
	Clinical***	308 (21.4)	82 (17.6)	226 (23.3)
	Modeling	39 (2.7)	11 (2.4)	28 (2.9)
University type				
	Teaching-oriented	49 (3.4)	11 (2.4)	38 (3.9)
	Research-oriented	840 (58.5)	274 (58.8)	566 (58.4)
	Equal emphasis	536 (37.3)	179 (38.4)	357 (36.8)
Professional status				
	Graduate student	7 (0.5)	3 (0.6)	4 (0.4)
	Postdoctoral student	7 (0.5)	1 (0.2)	6 (0.6)
	Non-tenure track instructor	47 (3.3)	16 (3.4)	31 (3.2)
	Non-tenure track researcher	53 (3.7)	17 (3.6)	36 (3.7)
	Untenured professor	284 (19.8)	98 (21.0)	186 (19.2)
	Tenured professor	912 (63.9)	228 (50.0)	624 (64.3)
	Chaired professor	102 (7.1)	30 (6.4)	72 (7.4)
Male				
		782 (55.7)	255 (56.5)	527 (55.3)
Race/ethnicity				
	African American*	30 (2.1%)	14 (3.0)	16 (1.6)
	Caucasian	1258(87.6)	401 (86.1)	857 (88.4)
	Asian	40 (2.8)	13 (2.8)	27 (2.8)
	Indian	4 (.3)	1 (.2)	3 (.3)
	Hispanic	40 (2.8)	13 (2.8)	27 (2.8)
Age				

20-29**	21 (1.5)	12 (2.7)	9 (1.0)
30-39	320 (23.0)	99 (22.1)	221 (23.4)
40-49	307 (22.0)	95 (21.3)	212 (22.4)
50-59	341 (24.5)	103 (23.0)	238 (25.2)
60-69	300 (21.5)	95 (21.3)	205 (21.7)
70+**	104 (7.5)	43 (9.6)	61 (6.4)
Country grew up in			
USA	1186 (82.6)	363 (77.9)	823 (84.8)
Canada	39 (2.7)	14 (3.0)	25 (2.6)

* significantly different between conditions at p<0.1
** significantly different between conditions at p<.05
*** significantly different between conditions at p<.01

Note: Denominators in a given cell are occasionally smaller than the sample sizes listed in the column headings because some participants did not answer all of the demographic questions.

Table S2. Prevalence estimates

Item	Control	BTS	two-tailed p (t-test)
In a paper, failing to report all of a study's dependent measures.	59.03%	60.98%	0.25
Deciding whether to collect more data after looking to see whether the results were significant.	61.01%	62.70%	0.27
In a paper, failing to report all of a study's conditions.	35.64%	37.32%	0.29
Stopping collecting data earlier than planned because one found the result that one had been looking for.	38.98%	40.74%	0.24
In a paper, 'Rounding off' a p value (e.g. reporting that a p value of .054 is less than .05)	40.55%	40.81%	0.88
In a paper, selectively reporting studies that 'worked.'	59.90%	60.80%	0.59
Deciding whether to exclude data after looking at the impact of doing so on the results.	45.24%	45.00%	0.89
In a paper, reporting an unexpected finding as having been predicted from the start.	47.73%	49.94%	0.15
In a paper, claiming that results are unaffected by demographic variables (e.g. gender) when one is actually unsure (or knows that they do).	18.72%	21.37%	0.02
Falsifying data.	9.33%	9.86%	0.38

*Difference between experimental conditions significant at alpha ≤ 0.005

Table S3. Admission estimates

Item	Control	BTS	two-tailed p (t-test)
In a paper, failing to report all of a study's dependent measures.	57.33%	57.84%	0.78
Deciding whether to collect more data after looking to see whether the results were significant.	56.30%	56.45%	0.93
In a paper, failing to report all of a study's conditions.	35.69%	36.98%	0.48
Stopping collecting data earlier than planned because one found the result that one had been looking for.	43.11%	44.05%	0.61
In a paper, 'Rounding off' a p value (e.g. reporting that a p value of .054 is less than .05)	37.67%	37.23%	0.81
In a paper, selectively reporting studies that 'worked.'	50.71%	49.49%	0.51
Deciding whether to exclude data after looking at the impact of doing so on the results.	35.83%	35.07%	0.66
In a paper, reporting an unexpected finding as having been predicted from the start.	38.14%	39.14%	0.54
In a paper, claiming that results are unaffected by demographic variables (e.g. gender) when one is actually unsure (or knows that they do).	17.20%	18.95%	0.20
Falsifying data.	3.34%	4.34%	0.17

*Difference between experimental conditions significant at alpha ≤ 0.005

Table S4. Results of a random effects ordered probit regression testing for differences in perceived defensibility, by sub-groups. The defensibility rating scale was: *no, possibly, yes.* Note: For participants who indicated multiple sub-groups, we weighted the dummy variables by the number of sub-groups they had chosen.

Propensity to judge behaviors defensible, by research type:

Discipline	**Beta**
Cognitive	-0.25
Developmental	-0.30*
Forensic	-0.20
Health	-0.19
Industrial Organizational	-0.20
Neuroscience	0.04
Personality	-0.78**
Social	-0.21
Clinical	-0.57***

Propensity to judge behaviors defensible, by research type:

Research type	**Beta**
Experiments	-0.39
Behavioral	-0.49
Field	-0.39
Modeling	-0.39
Clinical Research	-0.73**
Laboratory	-0.24

* $p < .05$
** $p < .01$
*** $p < .0005$