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

Leslie K. John

George Loewenstein

Drazen Prelec

Correspondence:
Leslie John
Harvard Business School - Marketing
Morgan Hall 169 Soldiers Field
Boston, MA 02163
T: 6174956394 F: 6174956394
ljohn@hbs.edu

Accepted for publication in *Psychological Science*: 10/20/11

Note: This is an uncorrected version of an author's manuscript accepted for publication in *Psychological Science*. Copyediting, typesetting, and review of the resulting proofs will be undertaken on this manuscript before final publication of the version of record at http://pss.sagepub.com/. During production and pre-press, errors may be discovered that could affect the content, and all legal disclaimers that apply to the journal relate to this version also.

Abstract

Cases of clear scientific misconduct have received significant media attention recently, but less flagrant transgressions of research norms may be more prevalent and in the long run more damaging to the academic enterprise. We surveyed over 2,000 psychologists about their involvement in questionable research practices, using an anonymous elicitation format supplemented by incentives for honest reporting. The impact of incentives on admission rates was positive, and greater for practices that respondents judge to be less defensible. Using three different estimation methods, we find that the proportion of respondents that have engaged in these practices is surprisingly high relative to respondents' own estimates of these proportions. Some questionable practices may constitute the prevailing research norm.

Although cases of scientific misconduct have received significant media attention recently (Altman, 2006; Deer, 2011; Steneck, 2006; 2000), exploitation of the grey-zone of acceptable practice is certainly much more prevalent, and may be more damaging to the academic enterprise in the long run. Questionable research practices (QRPs), such as excluding data points based on post hoc criteria, can spuriously increase the likelihood of finding evidence in support of a hypothesis. Just how dramatic these effects can be is demonstrated by Simmons, Nelson, and Simonsohn (in press), in a series of experiments and simulations which show how greatly QRPs increase the likelihood of finding support for a hypothesis that is false. QRPs are the steroids of scientific competition, artificially enhancing performance while providing considerable latitude for rationalization and self-deception. Concerns over QRPs have been mounting (Marshall, 2000; Sovacool, 2008), and several surveys, largely restricted to medical researchers, have assessed their prevalence (Gardner, Lidz, & Hartwig, 2005; Geggie, 2001; Henry et al., 2005; List, Bailey, Euzent, & Martin, 2007; 2005). In this survey, we measure the proportion of psychologists that have engaged in QRPs.

As with any unethical or socially stigmatized behavior, self-reported survey data may grossly underestimate true prevalence. Respondents have little incentive to provide honest answers, apart from good will (Fanelli, 2009). The goal of the present study was to obtain realistic estimates of QRPs with a new survey methodology that incorporates explicit response-contingent incentives for truth-telling and supplements self-reports with impersonal judgments about the prevalence of practices and about respondents' honesty. These impersonal judgments give rise to alternative estimates, which can be used to infer upper and lower bounds on actual prevalence. Across QRPs, even raw self-admission rates are surprisingly high, and for certain

practices the inferred actual estimates approach 100%, suggesting that these practices may constitute the de facto scientific norm.

Method

The survey was conducted on research psychologists at major U.S. universities. In the test condition, incentives for truth-telling were linked to the Bayesian truth serum scoring algorithm (Prelec, 2004), which uses respondents' personal answers, and their estimates of the sample distribution of answers, as inputs into a truth-rewarding scoring formula. Because of the anonymity requirement, compensation could not be directly linked to individual scores. Instead, respondents were told that we would make a donation to a charity of their choice, selected from five options, and that the size of this donation would depend on the truthfulness of their responses, as determined by the scoring system. By inducing a (correct) belief that dishonesty would reduce donations, we hoped to amplify the moral stakes riding on each answer. Respondents were not given the details of the scoring, but were told that it was based on an algorithm published in *Science* and were given a link to the paper. There was no deception: respondents' BTS scores determined our contributions to the five charities. Respondents in the Control condition were simply told that a charitable donation would be made on behalf of each respondent.

The study was a 2 condition between-subjects design in which we e-mailed an electronic survey to 5,964 academic psychologists (details in supplementary materials section). There were 2,155 respondents, for a response rate of 36%. Participants anonymously indicated whether they had personally engaged in each of ten QRPs (Table 1), and, if so, whether they

thought their actions had been defensible. The order in which the QRPs were presented was randomized between-subjects.

Respondents also provided two impersonal judgments: a) the proportion of other psychologists who had engaged in the behaviors (*prevalence estimate*), and b) among those psychologists who had engaged in the behavior, the percentage that would admit to having done so "in a survey like this" (*admission estimate*). Therefore, each respondent was asked to provide three pieces of information for each QRP.

Observe that each of the three answers: personal admission, prevalence estimate, admission estimate, creates a different route to a final estimate of actual prevalence. The credibility of each estimate hinges on the credibility of one of the three answers in the survey: (1) If respondents answer the personal question honestly, then personal admission rates reveal actual (sample) prevalence. (2) If average prevalence estimates are accurate, then they will also directly estimate actual prevalence. (3) If average admission estimates are accurate, then actual prevalence is estimated by the ratios of admission rates to admission estimates. This would correspond to the case where respondents don't know the actual prevalence of a practice, but do have a good sense of how likely it is that a colleague would admit to it in a survey context. The three estimates should converge if: admission rate = prevalence estimate x admission estimate. To the extent that this equality is violated, we will see gaps between different estimates.

Results

Truth-telling incentives. We present the three sets of estimates in Figure 1, but first describe the impact of BTS. A priori, truth-telling incentives should affect items in proportion to

the baseline (i.e., Control condition) level of false denials. These baseline levels are unknown, but one can hypothesize that they will be minimal for impersonal estimates of prevalence and admission, and greatest for personal admission to unethical practices with high odium, representing 'red card' violations.

As hypothesized, BTS affected neither prevalence estimates (Table S2) nor admission estimates (Table S3), but did increase personal admission rates for some items (Table 1), especially those that are "more questionable." The impact of BTS is captured by the Odds Ratio of BTS to Control condition admission rates. It is large for one practice (falsifying data), moderate for three practices (premature stopping of data collection, falsely reporting a finding as expected, falsely claiming that results are unaffected by certain variables), and negligible for the rest.

The acceptability of a practice can be inferred from the baseline (Control) admission rate or assessed directly by judgments of defensibility (scoring: 0 for Indefensible, +1 for Possibly defensible, +2 for Defensible). The nonparametric correlation of BTS impact, as measured by Odds Ratio, with Control admission rate is –.62 ($p < .06$; parametric correlation = -.65, $p < .05$), and with defensibility it is –0.68 ($p < .03$; parametric correlation = -.94, $p < .001$). These correlations are more modest when item 10 (Falsifying data) is excluded (Odds Ratio with Control admission rate: nonparametric correlation = -.48, $p < .20$; parametric correlation = -.59, $p<.10$; Odds Ratio with defensibility: nonparametric correlation = -.57, $p < .12$; parametric correlation = -.59, $p < .10$).

Prevalence estimates. Figure 1 displays prevalence estimates obtained given the assumption that one of the three types of estimates is accurate, and using the BTS condition only. The numbers in the figure above each set of bars are geometric means of the three estimates,

which in effect give equal credence to the three types of answers (the admission rate estimates are capped at 100%; they exceed 100% by a small margin for a few items). The first of the three estimates, based on raw admission rates, is almost certainly too low, given the likelihood that respondents did not admit to all QRPs that they actually engaged in. Therefore, the geometric means are probably conservative judgments of true prevalence.

One would infer from the means embedded in Figure 1 that one in ten research psychologists has introduced false data into the scientific record (Items 9 and 10), while the majority of research psychologists have engaged in practices such as (1) selective reporting of studies, (3) not reporting all dependent measures, (4) collecting more data, (6) reporting unexpected findings as expected, and (8) excluding data post-hoc.

These estimates are somewhat higher than earlier estimates. For example, a meta-analysis of surveys designed to assess the prevalence of QRPs – none of which provided incentives for truthful responding – found that, across studies, on average, 9.5% of respondents admitted to having engaged in QRPs other than data falsification; the upper-bound estimate was 33.7% (Fanelli, 2009). In the present study, the mean admission rate in BTS (excluding the data falsification item for comparability to (Fanelli, 2009)) was 36.6% – higher than both of the meta-analysis estimates. Moreover, among participants in the BTS condition who completed the survey, 94.0% admitted to having engaged in at least one QRP. Interestingly, the admission rate in our control condition (33.0%) mirrored the upper-bound estimate obtained in the meta-analysis (33.7%).

Response to a given item was predictive of responses to the other items: the survey items approximated a Guttman scale, meaning that an admission to a relatively rare behavior (e.g. falsifying data) usually implied that the given respondent had also admitted to having engaged in

the more common behaviors. Among completed response sets, the coefficient of reproducibility – the average proportion of a person's responses that can be reproduced by knowing the number of items to which he responded affirmatively – was 0.80 (high values indicate close agreement; reproducibility of ≥ 0.90 is considered to be a Guttman scale (Guttman, 1974)). This suggests that the specific QRPs that researchers engage in or avoid are not completely idiosyncratic; researchers differ little in judgments of the relative unethicality of the behaviors, but greatly on where they draw the line when it comes to their own behavior.

Perceived defensibility. Respondents had an opportunity to state whether they thought their actions were defensible. Consistent with the notion that latitude for rationalization is positively associated with engagement in QRPs, respondents who admitted to a QRP tended to think that their actions were defensible. The overall mean defensibility rating was 1.44 (*SD* = 0.76) – between "possibly defensible" and "defensible." Mean judged defensibility for each item is shown in the right-hand column of Table 1. Defensibility ratings did not generally differ by sub-group (Table S4).

Doubts about research integrity. A relatively large proportion of respondents indicated that they had had doubts about research integrity on at least one occasion (Figure 2). The degree of doubt differed by target; for example, respondents were more wary of research generated by those at other institutions than that conducted by their collaborators. Although heterogeneous referent group sizes make these differences difficult to interpret (the number of researchers at other institutions is presumably larger than one's set of collaborators), it is noteworthy that approximately 35% of respondents indicated that they had had doubts about the integrity of their *own* research on at least on occasion.

Frequency of engagement. Although the prevalence estimates obtained in the BTS condition are somewhat higher than previous estimates, they do not distinguish between the researcher who routinely engages in a given behavior, from the researcher who has "only" engaged in it once. To the extent that admission rates are driven by the former type, our results are more worrisome. We conducted a smaller-scale survey in which we tested for differences in admission rates as a function of the response scale.

We asked attendees of an annual conference of behavioral researchers (N=132) whether they had engaged in each of 25 different QRPs (many of which we also used in the BTS study). The study was a 2x2 between-subjects design in which we manipulated the wording of the questions and the response scale. The questions were either worded in infinitive ("Falsifying data.") or first person ("I have falsified data."), and participants indicated whether they had engaged in the behaviors using either a dichotomous (yes/no, as in the BTS study) or frequency (never / once or twice / occasionally / frequently) response scale.

Because the overall admission rates to the individual items were generally similar to those obtained in the BTS study, we do not report them here. The dichotomous response scale yielded fewer affirmative admissions compared to the frequency response scale ($M_{affirm_responses_dichotomous}$ = 3.81 out of 25, *SD* = 2.26; $M_{affirm_responses_frequency}$ = 6.01, *SD* = 3.70; $F(1, 128)=15.6$, $p<.0005$). This result suggests that in the dichotomous scale condition, some non-trivial fraction of respondents who only engaged in a QRP a small number of times reported that they had *never* engaged in the practice, suggesting that the prevalence rates obtained in the BTS study are conservative. There was no effect of the wording manipulation.

We explored the response scale effect further by comparing the distribution of responses between the two response scale conditions across all 25 items and collapsing across the wording

manipulation (Figure 3). Looking among the affirmative responses in the frequency response scale condition (i.e. responses of "once or twice", "occasionally" or "frequently"), although 63% (i.e. .15/(.15+.06+.02)) of the affirmative responses fall into the "once or twice" category, a non-trivial proportion fall into "occasionally" (27%) and "frequently" (10%). This result suggests that the prevalence estimates from the BTS study represent a combination of "one-off" as well as more habitual engagement in the behaviors.

Sub-group differences. Table 2 presents admission rates as a function of subdisciplines within psychology (top panel) and primary methodology used in research (bottom panel). The second column of each panel presents admission rates across all 10 QRPs. Relatively high rates of QRPs were self-reported among cognitive, neuroscience and social subdisciplines, and among those using behavioral, experimental, and laboratory methodologies. Clinical psychologists reported relatively low rates of QRPs.

These differences could reflect the particular relevance of our QRPs to these subdisciplines and methodologies, or they could reflect differences in perceived defensibility of the behaviors. To explore these possible explanations, we sent a brief follow-up survey to 1,440 of the BTS survey participants which asked them to rate the same ten QRPs from the initial study on: 1. The extent to which each practice applies to their research methodology – i.e., how frequently, if at all, they encounter the opportunity to engage in the practice (response scale: Never applicable / Sometimes applicable / Often applicable / Always applicable); and 2. Whether it is generally defensible to engage in each practice (response scale: Indefensible / Possibly defensible / Defensible). We counterbalanced the order in which respondents rated the two dimensions. There were 504 respondents, for a response rate of 35%.

The third and fourth columns of both panels of Table 2 present the results from the follow-up survey. The subgroup differences in applicability and defensibility are partially consistent with the differences in self-reported prevalence: most notably, mean applicability and defensibility were elevated among social psychologists – a sub-group with relatively high admission rates. Similarly, the items were particularly applicable to (but not judged more defensibly by) those conducting behavioral, experimental, and laboratory research.

To test for the relative importance of applicability and defensibility in explaining subfield differences, we conducted an ANOVA of mean admission rates across QRPs and subfields. Both type of QRP (p <.001; Partial η^2 = .87) and subfield (p <.05, partial η^2 =.21) were highly significant predictors of admission rates, and their significance and effect size were largely unchanged after controlling for applicability and defensibility, even though both of the latter variables were highly significant independent predictors of mean admission rates. Similarly, methodology was also a highly significant predictor of admission rates (p <.05, η^2 = .27) and its significance and effect size were largely unchanged after controlling for applicability and defensibility (even though the latter were highly significant predictors of admission rates).

The defensibility ratings obtained in the main study (Table 1, sixth column) stand in contrast to those obtained in follow-up survey (Table 2, fourth column): the behaviors are deemed defensible by people who have engaged in them (main study), but indefensible overall (follow-up study).

Discussion

Concerns over scientific misconduct have led previous researchers to attempt to estimate the prevalence of QRPs that are broadly applicable to scientists (Martinson, et al., 2005). In light of recent concerns over scientific integrity within psychology, this study was designed to provide accurate estimates of the prevalence of QRPs that are specifically applicable to research psychologists. In addition to being one of the first surveys to specifically target research psychologists, it is also the first to test the effectiveness of an incentive compatible elicitation format that measures prevalence rates in three different ways.

All three prevalence measures point to the same conclusion: a surprisingly high proportion of psychologists admit to having engaged in QRPs. The impact of the BTS condition on admission rates was positive, and greater for practices that respondents judge to be less defensible. Beyond revealing the prevalence of QRPs, this study is also, to our knowledge, the first to illustrate that an incentive compatible information elicitation method can lead to higher, and likely more valid, prevalence estimates of sensitive behaviors. The method could easily be used to estimate the prevalence of other sensitive behaviors, such as illegal or sexual activities. For potentially even greater benefit, BTS-based truth-telling incentives could be combined with audio computer-assisted self-interviewing – a technology that has been found to increase self-reporting of sensitive behaviors (Turner, et al., 1998).

There are a number of components to the BTS procedure – both a request and incentive to tell the truth – and we are unable to isolate their independent effects on disclosure. Importantly however, both components rewarded respondents for telling the truth, not for simply responding "yes" regardless of whether they had engaged in the behaviors. Therefore, both components were designed to increase the validity of responses. Future research could test the relative contribution of the various BTS components in eliciting truthful responses.

This research is premised on the assumption that higher prevalence estimates are more valid – an assumption that pervades a large body of research designed to assess the prevalence of sensitive behaviors (Bradburn & Sudman, 1979; de Jong, Pieters, & Fox, 2010; Goodstadt & Gruson, 1975; Lensvelt-Mulders, Hox, van der Heijden, & Maas, 2005; Locander, Sudman, & Bradburn, 1976; Steneck, 2006; Tourangeau & Yan, 2007). This assumption is generally accepted, provided that the behaviors in question are sensitive and/or socially undesirable. The rationale is that respondents are unlikely to be tempted to admit to shameful behaviors in which they have not engaged; instead, they are prone to denying involvement in behaviors in which they actually have engaged (Fanelli, 2009). We think this assumption is also defensible in the present study given its subject matter.

As noted in the introduction, there is a large grey-zone of acceptable practice. While falsifying data (item 10) is never justified, the same cannot be said for all of the items; for example, failing to report all of a study's dependent measures (item 1) could be appropriate if two measures of the same construct show the same significant pattern of results but cannot be easily combined into one measure. Therefore, not all admissions represent scientific felonies, or even misdemeanors; some respondents provided perfectly defensible reasons for engaging in the behaviors. Yet, other respondents provided justifications which, although self-categorized as "defensible," were contentious (e.g. dropping dependent measures inconsistent with the hypothesis because doing so enabled a more coherent story to be told, increasing the likelihood of publication). It is worth noting however, that in the follow-up survey – in which participants rated the behaviors *regardless* of engagement – the defensibility ratings were low, suggesting that the general sentiment is that these behaviors are unjustifiable.

We assume that the vast majority of researchers are sincerely motivated to conduct sound scientific research. Indeed, respondents seem to engage in QRPs unknowingly – those indicating that they have engaged in the practices generally believe their actions to be defensible (Table 1). This belief may be in part a byproduct of publication pressures: the inherent ambiguity in the defensibility of research practices may lead researchers to, however inadvertently, use this ambiguity to delude themselves that their own dubious research practices are "defensible" (Kunda, 1990). This line of thinking could in part explain why the most egregious practices in our survey (e.g. "falsifying data") appear to be less common than the relatively "less questionable" ones (e.g. "failing to report all of a study's conditions") – it is easier to dream up a post-hoc explanation to "justify" removing "nuisance" data points than it is justify outright data falsification, even though both practices produce similar consequences.

Given the findings of our survey, it comes as no surprise that many researchers have expressed concerns over failures to replicate published results (Bower & Mayer, 1985; Crabbe, Wahlsten, & Dudek, 1999; Enserink, 1999; Galak & Nelson, 2010; Ioannidis, 2005, 2005; Palmer, 2000; Steele, Bass, & Crook, 1999). In a New Yorker piece (2010) on the problem of nonreplicability, Jonah Lehrer discusses possible explanations for the "decline effect" – the tendency for effect sizes to decrease with subsequent attempts at replication. He concludes that conventional accounts of this effect (regression to the mean; publication bias) may be incomplete. In a subsequent and insightful commentary, Jonathan Schooler suggests that unpublished data may help to account for the decline effect (Schooler, 2011). By documenting the surprisingly large proportion of researchers that have engaged in QRPs – including selective omission of observations, experimental conditions, and studies from the scientific record – the present research provides empirical support for Schooler's claim. Recent work by Simmons,

Nelson, and Simonsohn (in press) goes further, by showing how easily QRPs can yield invalid findings, and by proposing reforms.

QRPs can waste researchers' time and stall scientific progress, as researchers fruitlessly pursue extensions of effects that are not real and hence do not replicate. But more disheartening is the fact that they threaten research integrity and produce unrealistically elegant results that may be difficult to match without engaging in such practices oneself. This can lead to a 'race to the bottom,' with questionable research begetting even more questionable research. If reforms were effective in reducing the prevalence of QRPs, this would not only bolster scientific integrity but could also reduce the pressure on researchers to produce unrealistically elegant results.

References

Altman, L. (2006, May 2). For Science Gatekeepers, a Credibility Gap. *The New York Times*.

Bower, G. H., & Mayer, J. D. (1985). Failure to Replicate Mood-dependent Retrieval. *Bulletin of the Psychonomic Society, 23*(1), 39-42.

Bradburn, N., & Sudman, S. (1979). *Improving Interview Method and Questionnaire Design - Response Effects to Threatening Questions in Survey Research*. San Francisco, CA: Jossey-Bass Publishers.

Crabbe, J. C., Wahlsten, D., & Dudek, B. C. (1999). Genetics of mouse behavior: interactions with laboratory environment. *Science, 284*(5420), 1670-1672.

de Jong, M. G., Pieters, R., & Fox, J.-P. (2010). Reducing Social Desirability Bias Through Item Randomized Response: An Application to Measure Underreported Desires. Journal of Marketing Research, 47(1), 14-27.

Deer, B. (2011). How the case against the MMR vaccine was fixed. *British Medical Journal, 342*.

Enserink, M. (1999). Fickle mice highlight test problems. *Science, 284*(1599-600).

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.

Galak, J., & Nelson, L. D. (2010). A Replication of the Procedures from Bem (2010, Study 8) and a Failure to Replicate the Same Results.

Gardner, W., Lidz, C. W., & Hartwig, K. C. (2005). Authors' reports about research integrity problems in clinical trials. *Contemporary Clinical Trials, 26*(2), 244-251.

Geggie, D. (2001). A survey of newly appointed consultants' attitudes towards research fraud. *Journal of Medical Ethics, 27*, 344-346.

Goodstadt, M. S., & Gruson, V. (1975). The randomized response technique: A test on drug use. Journal of the American Statistical Association, 70(352), 814-818.

Guttman, L. L. (1974). The Basis for Scalogram Analysis. In G. M. Maranell (Ed.), *Scaling: A Sourcebook for Behavioral Scientists*. New Brunswick, New Jersey: Transaction Publishers.

Hegarty, W. H., & Sims, H. P. (1978). Some determinants of unethical decision behavior: An experiment. *Journal of Applied Psychology, 63*(4), 451-457.

Henry, D. A., Kerridge, I. H., Hill, S. R., McNeill, P. M., Doran, E., Newby, D. A., et al. (2005). Medical specialists and pharmaceutical industry-sponsored research: a survey of the Australian experience. *Medical Journal of Australia, 182*(11), 557-560.

Ioannidis, J. P. A. (2005). Contradicted and Initially Stronger Effects in Highly Cited Clinical Research. *Journal of the American Medical Association, 294*(2).

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), 696-701.

Kahneman, D., Ritov, I., Jacowitz, K. E., & Grant, P. (1993). Stated willingness to pay for public goods. *Psychological Science, 4*(5), 310-315.

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

Lehrer, J. (2010, December 13, 2010). The Truth Wears Off. *The New Yorker*.

Lensvelt-Mulders, G. J. L. M., Hox, J. J., van der Heijden, P. G. M., & Maas, C. J. M. (2005). Meta-analysis of randomized response research: 35 years of validation. Sociological Methods and Research, 33(319), 319-348.

List, J., Bailey, C., Euzent, P., & Martin, T. (2007). Academic economists behaving badly? A survey on three areas of unethical behavior. *Economic Inquiry, 39*(1), 162-170.

Locander, W., Sudman, S., & Bradburn, N. (1976). An investigation of interview method, threat, and response distortion. Journal of American Statistics, 71, 269-275.

Marshall, E. (2000). How prevalent is fraud? That's a million dollar question. *Science, 290*, 1662-1663.

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

Palmer, A. R. (2000). Quasireplication and the Contract of Error: Lessons from Sex Ratios, Heritabilities, and Fluctuating Asymmetry. *Annual Review of Ecology and Systematics, 31*, 441-480.

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

Schooler, J. (2011) Unpublished results hide the decline effect. Nature 470: 437.

Simmons, J., Nelson, L., & Simonsohn, U. (in press). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*.

Sovacool, B. (2008). Exploring scientific misconduct: isolated individuals, impure institutions, or an inevitable idiom of modern science? *Journal of Bioethical Inquiry, 5*(271-282).

Steele, K. M., Bass, K. E., & Crook, M. D. (1999). The Mystery of the Mozart Effect: Failure to Replicate. *Psychological Science, 10*(4), 366-369.

Steneck, N. H. (2006). Fostering integrity in research: Definitions, Current knowledge, and future directions. *Science and Engineering Ethics, 12*(53-74).

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. Psychological Bulletin, 133(5), 859-883.

Turner, C. F., Ku, L., Rogers, S. M., Lindberg, L. D., Pleck, J. H., & Sonenstein, F. L. (1998). Adolescent Sexual Behavior, Drug Use, and Violence: Increased Reporting with Computer Survey Technology. *Science*, *280*, 867-873.

Weiss, R., Rifkin, R., Stewart, F., Theriault, R., Williams, L., Herman, A., et al. (2000). High-dose chemotherapy for high-risk primary breast cancer: an on-site review of the Bezwoda study. *The Lancet*, *355*(9208), 999-1003.

Author Note

Leslie K. John, Harvard Business School; George Loewenstein, Carnegie Mellon University; Drazen Prelec, Massachusetts Institute of Technology.

We thank Evan Robinson for implementing the email procedure that tracked participation while ensuring respondents' anonymity. We also thank Anne-Sophie Charest and Bill Simpson for statistical consulting and members of the Center for Behavioral Decision Research for their input on initial drafts of the survey items.

Address correspondence about this paper to ljohn@hbs.edu.

Table 1. **Admission rates and defensibility ratings, by item.** Items are listed in decreasing order of judged defensibility. Note: Defensibility ratings were provided by respondents who admitted to having engaged in the given behavior.

Item	Control (%)	BTS (%)	Odds Ratio	Two-tailed p (likelihood ratio)	Mean defensibility (*SD*) 0 = Indefensible 1 = Possibly defensible 2 = Defensible
1. In a paper, failing to report all of a study's dependent measures.	63.4	66.5	1.14	0.23	1.84 (.39)
2. Deciding whether to collect more data after looking to see whether the results were significant.	55.9	58.0	1.08	0.46	1.79 (.44)
3. In a paper, failing to report all of a study's conditions.	27.7	27.4	0.98	0.90	1.77 (.49)
4. Stopping collecting data earlier than planned because one found the result that one had been looking for.*	15.6	22.5	1.57	0.00	1.76 (.48)
5. In a paper, 'Rounding off' a p value (e.g. reporting that a p value of .054 is less than .05)	22.0	23.3	1.07	0.58	1.68 (.57)
6. In a paper, selectively reporting studies that 'worked.'	45.8	50.0	1.18	0.13	1.66 (.53)
7. Deciding whether to exclude data after looking at the impact of doing so on the results.	38.2	43.4	1.23	0.06	1.61 (.59)
8. In a paper, reporting an unexpected finding as having been predicted from the start.*	27.0	35.0	1.45	0.00	1.5 (.60)
9. In a paper, claiming that results are unaffected by demographic variables (e.g. gender) when one is actually unsure (or knows that they do).	3.0	4.5	1.52	0.16	1.32 (.60)
10. Falsifying data.	0.6	1.7	2.75	0.07	0.16 (.37)
Aggregate indicator of self-proclaimed 'perfection'					
Respondents with no admissions for any of the ten items	8.6	5.3	0.6	0.03	
Respondents with no admissions, including those who did not finish	8.5	5.4	0.62	0.04	

*Difference between experimental conditions significant at alpha ≤ 0.005

Table 2. Prevalence rates, applicability ratings and defensibility ratings by discipline.
Note: admission rate data are from the BTS study; applicability and defensibility data are from the follow-up study.

Discipline	**Admission rate**	**Applicability** 1 = never applicable 2 = sometimes applicable 3 = often applicable 4 = always applicable	**Defensibility** 0 = Indefensible 1 = Possibly defensible 2 = Defensible
Clinical	0.27*	2.59	0.56
Cognitive	0.37***	2.75*	0.64
Developmental	0.31	2.77**	0.66
Forensic	0.28	3.02*	0.52
Health	0.30	2.56	0.69
Industrial Organizational	0.31	2.80	0.73
Neuro	0.35**	2.71	0.61
Personality	0.32	2.65*	0.66
Social	0.40***	2.89***	0.73**
Research type	**Admission rate**	**Applicability** 1 = never applicable 2 = sometimes applicable 3 = often applicable 4 = always applicable	**Defensibility** 0 = Indefensible 1 = Possibly defensible 2 = Defensible
Clinical	0.30	2.61	0.56
Behavioral	0.34*	2.77**	0.63
Laboratory	0.37***	2.86***	0.66
Field	0.31	2.76**	0.64
Experiments	0.36***	2.83*	0.66*
Modelling	0.34	2.74	0.62

Significance codes:
*p<.05, **p<.01, ***p<.0005
For “Admission rate,” significance codes are based on random effects logistic regression; for “Applicability” and “Defensibility”, significance codes are based on random effects ordered probit regressions.

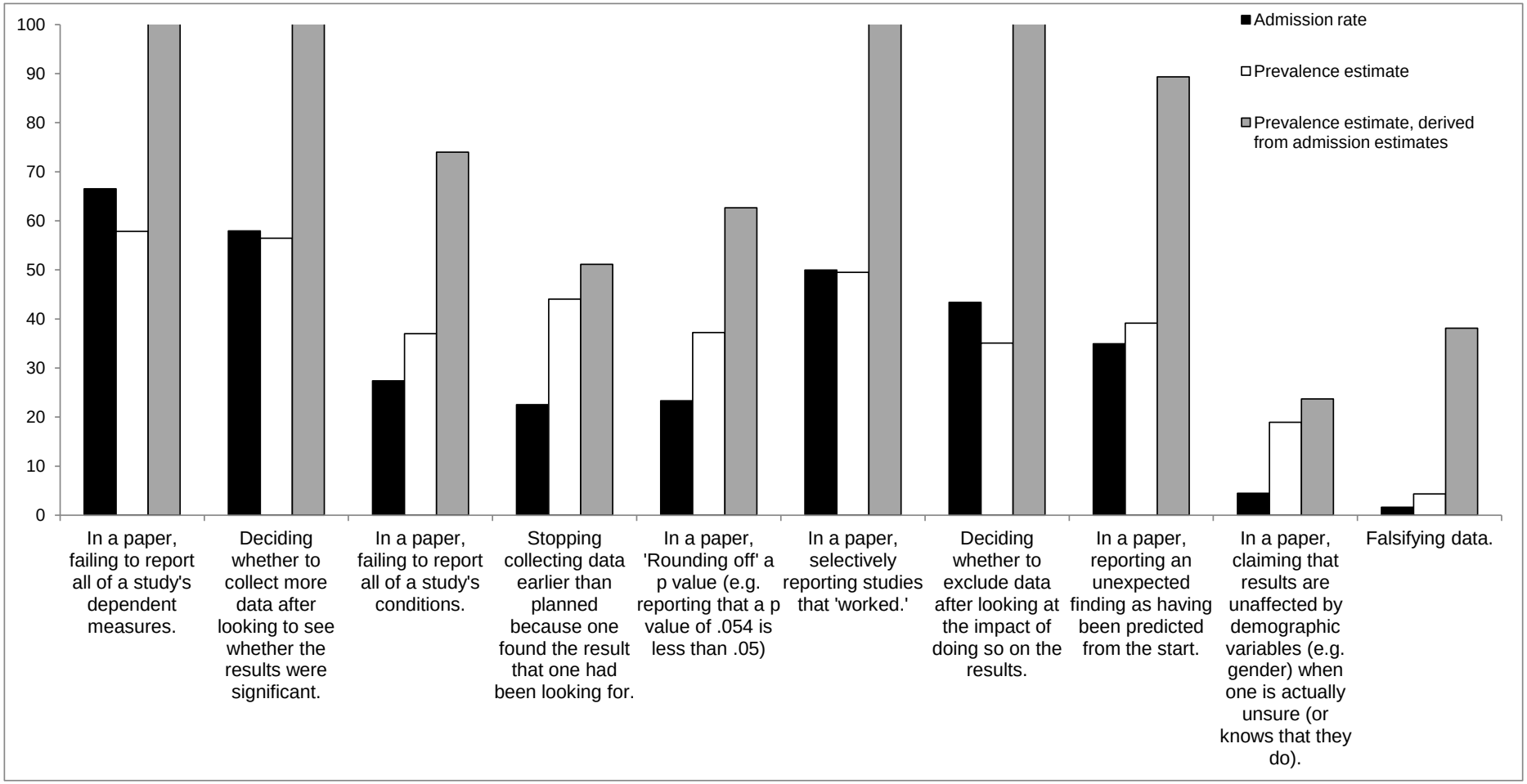

Figure 1. Three measures of the prevalence of questionable research practices, using BTS condition data only: a) affirmative admission rate, b) prevalence estimates, and c) prevalence estimates derived from admission estimates (i.e. admission rate/admission estimate). The bolded numbers in the figure are geometric means of the three estimates.

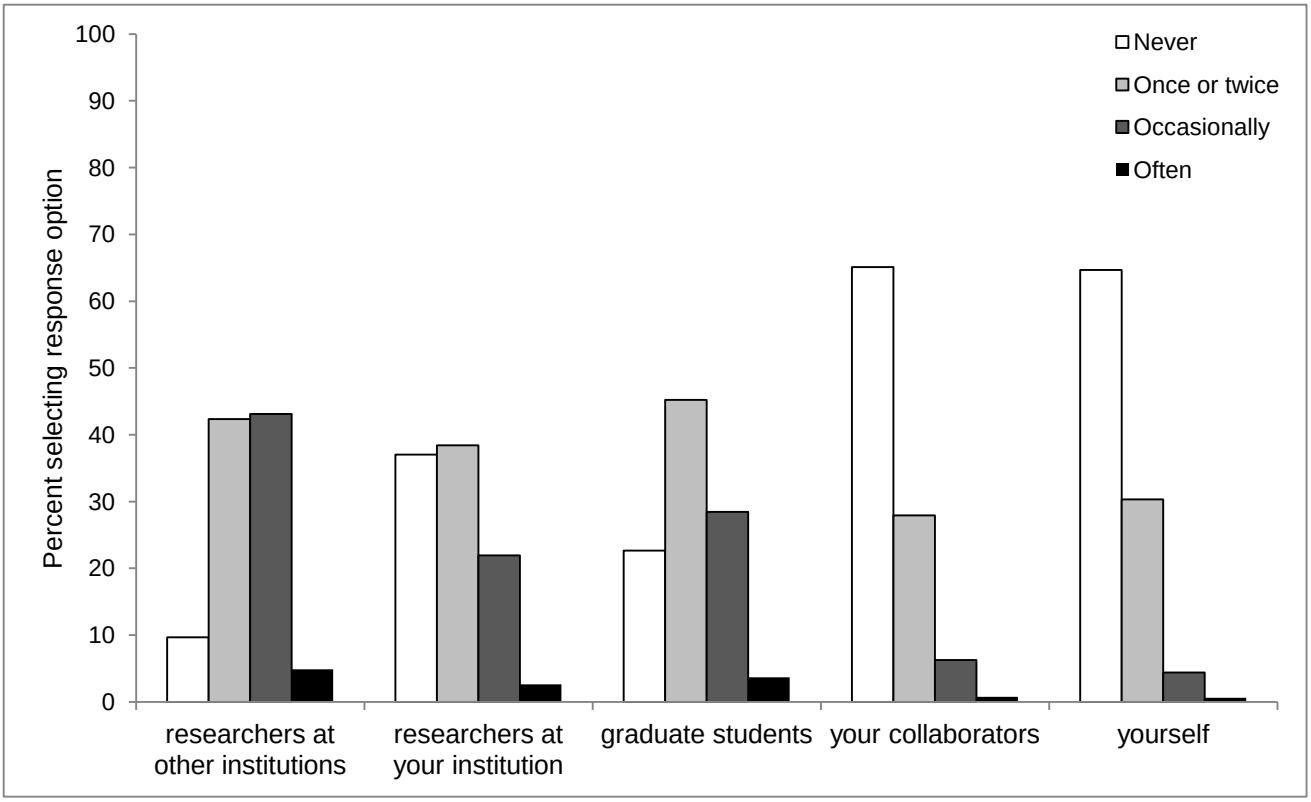

Figure 2. Responses to the question: "Have you ever had doubts about the integrity of the research done by…"

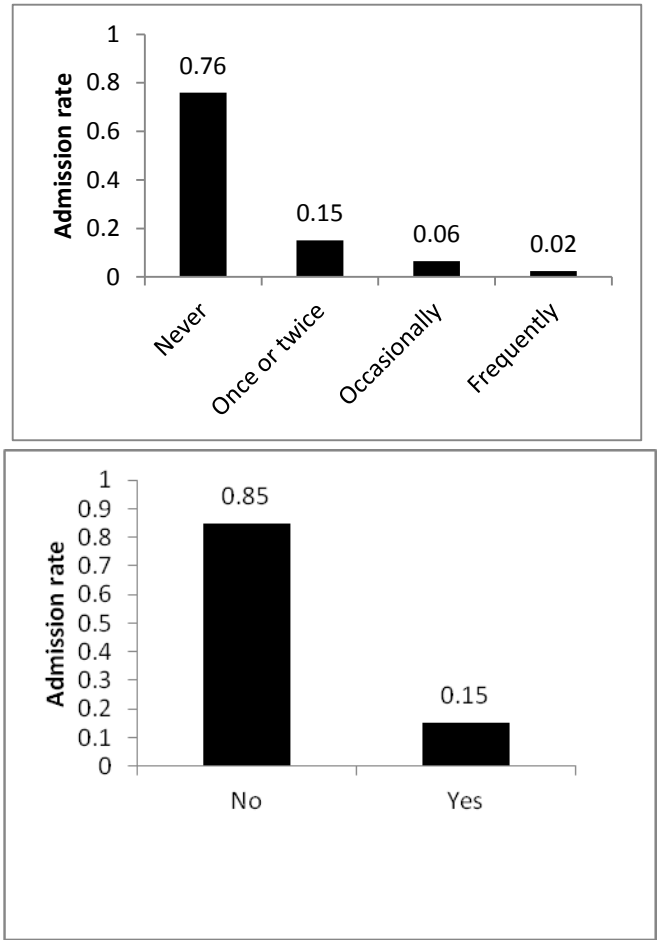

Figure 3. Distribution of responses among participants who answered using a frequency response scale (Figure 3A) versus a dichotomous response scale (Figure 3B).