Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling

Leslie K. John[1], George Loewenstein[2], and Drazen Prelec[3]

[1]Marketing Unit, Harvard Business School; [2]Department of Social & Decision Sciences, Carnegie Mellon University; and [3]Sloan School of Management and Departments of Economics and Brain & Cognitive Sciences, Massachusetts Institute of Technology

Abstract

Cases of clear scientific misconduct have received significant media attention recently, but less flagrantly questionable research practices may be more prevalent and, ultimately, more damaging to the academic enterprise. Using an anonymous elicitation format supplemented by incentives for honest reporting, we surveyed over 2,000 psychologists about their involvement in questionable research practices. The impact of truth-telling incentives on self-admissions of questionable research practices was positive, and this impact was greater for practices that respondents judged to be less defensible. Combining three different estimation methods, we found that the percentage of respondents who have engaged in questionable practices was surprisingly high. This finding suggests that some questionable practices may constitute the prevailing research norm.

Although cases of overt scientific misconduct have received significant media attention recently (Altman, 2006; Deer, 2011; Steneck, 2002, 2006), exploitation of the gray area of acceptable practice is certainly much more prevalent, and may be more damaging to the academic enterprise in the long run, than outright fraud. Questionable research practices (QRPs), such as excluding data points on the basis of post hoc criteria, can spuriously increase the likelihood of finding evidence in support of a hypothesis. Just how dramatic these effects can be was demonstrated by Simmons, Nelson, and Simonsohn (2011) in a series of experiments and simulations that showed how greatly QRPs increase the likelihood of finding support for a false hypothesis. QRPs are the steroids of scientific competition, artificially enhancing performance and producing a kind of arms race in which researchers who strictly play by the rules are at a competitive disadvantage. QRPs, by nature of the very fact that they are often questionable as opposed to blatantly improper, also offer considerable latitude for rationalization and self-deception.

Concerns over QRPs have been mounting (Crocker, 2011; Lacetera & Zirulia, 2011; Marshall, 2000; Sovacool, 2008; Sterba, 2006; Wicherts, 2011), and several studies—many of which have focused on medical research—have assessed their prevalence (Gardner, Lidz, & Hartwig, 2005; Geggie, 2001; Henry et al., 2005; List, Bailey, Euzent, & Martin, 2001;

Martinson, Anderson, & de Vries, 2005; Swazey, Anderson, & Louis, 1993). In the study reported here, we measured the percentage of psychologists who have engaged in QRPs.

As with any unethical or socially stigmatized behavior, self-reported survey data are likely to underrepresent true prevalence. Respondents have little incentive, apart from good will, to provide honest answers (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 made it possible to elicit alternative estimates, from which we inferred the upper and lower boundaries of the actual prevalence of QRPs. Across QRPs, even raw self-admission rates were surprisingly high, and for certain practices, the inferred actual estimates approached 100%, which suggests that these practices may constitute the de facto scientific norm.

Corresponding Author:
Leslie K. John, Harvard Business School—Marketing, Morgan Hall 169, Soldiers Field, Boston, MA 02163
E-mail: ljohn@hbs.edu

Method

In a study with a two-condition, between-subjects design, we e-mailed an electronic survey to 5,964 academic psychologists at major U.S. universities (for details on the survey and the sample, see Procedure and Table S1, respectively, in the Supplemental Material available online). Participants anonymously indicated whether they had personally engaged in each of 10 QRPs (*self-admission rate*; Table 1), and if they had, whether they thought their actions had been defensible. The order in which the QRPs were presented was randomized between subjects. There were 2,155 respondents to the survey, which was a response rate of 36%. Of respondents who began the survey, 719 (33.4%) did not complete it (see Supplementary Results and Fig. S1 in the Supplemental Material); however, because the QRPs were presented in random order, data from all respondents—even those who did not finish the survey—were included in the analysis.

In addition to providing self-admission rates, respondents also provided two impersonal estimates related to each QRP: (a) the percentage of other psychologists who had engaged in each behavior (*prevalence estimate*), and (b) among those psychologists who had, the percentage that would admit to having done so (*admission estimate*). Therefore, each respondent was asked to provide three pieces of information for each QRP. Respondents who indicated that they had engaged in a QRP were also asked to rate whether they thought it was defensible to have done so (0 = *no*, 1 = *possibly*, and 2 = *yes*). If they wished, they could also elaborate on why they thought it was (or was not) defensible.

After providing this information for each QRP, respondents were also asked to rate their degree of doubt about the integrity of the research done by researchers at other institutions, other researchers at their own institution, graduate students, their collaborators, and themselves (1 = *never*, 2 = *once or twice*, 3 = *occasionally*, 4 = *often*).

Table 1. Results of the Main Study: Mean Self-Admission Rates, Comparison of Self-Admission Rates Across Groups, and Mean Defensibility Ratings

Item	Self-admission rate (%)		Odds ratio (BTS/control)	Two-tailed p (likelihood ratio test)	Defensibility rating (across groups)
	Control group	BTS group			
1. In a paper, failing to report all of a study's dependent measures	63.4	66.5	1.14	.23	1.84 (0.39)
2. Deciding whether to collect more data after looking to see whether the results were significant	55.9	58.0	1.08	.46	1.79 (0.44)
3. In a paper, failing to report all of a study's conditions	27.7	27.4	0.98	.90	1.77 (0.49)
4. Stopping collecting data earlier than planned because one found the result that one had been looking for	15.6	22.5	1.57	.00	1.76 (0.48)
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	.58	1.68 (0.57)
6. In a paper, selectively reporting studies that "worked"	45.8	50.0	1.18	.13	1.66 (0.53)
7. Deciding whether to exclude data after looking at the impact of doing so on the results	38.2	43.4	1.23	.06	1.61 (0.59)
8. In a paper, reporting an unexpected finding as having been predicted from the start	27.0	35.0	1.45	.00	1.50 (0.60)
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	.16	1.32 (0.60)
10. Falsifying data	0.6	1.7	2.75	.07	0.16 (0.38)

Note: Items are listed in decreasing order of rated defensibility. Respondents who admitted to having engaged in a given behavior were asked to rate whether they thought it was defensible to have done so (0 = *no*, 1 = *possibly*, and 2 = *yes*). Standard deviations are given in parentheses. BTS = Bayesian truth serum. Applying the Bonferroni correction for multiple comparisons, we adjusted the critical alpha level downward to .005 (i.e., .05/10 comparisons).

The two versions of the survey differed in the incentives they offered to respondents. In the Bayesian-truth-serum (BTS) condition, a scoring algorithm developed by one of the authors (Prelec, 2004) was used to provide incentives for truth telling. This algorithm uses respondents' answers about their own behavior and their estimates of the sample distribution of answers as inputs in a truth-rewarding scoring formula. Because the survey was anonymous, 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 BTS scoring system. By inducing a (correct) belief that dishonesty would reduce donations, we hoped to amplify the moral stakes riding on each answer (for details on the donations, see Supplementary Results in the Supplemental Material). Respondents were not given the details of the scoring system but were told that it was based on an algorithm published in *Science* and were given a link to the article. 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. (For details on the effect of the size of the incentive on response rates, see Participation Incentive Survey in the Supplemental Material.)

The three types of answers to the survey questions—self-admission, prevalence estimate, admission estimate—allowed us to estimate the actual prevalence of each QRP in different ways. The credibility of each estimate hinged on the credibility of one of the three answers in the survey: First, if respondents answered the personal question honestly, then self-admission rates would reveal the actual prevalence of the QRPs in this sample. Second, if average prevalence estimates were accurate, then they would also allow us to directly estimate the actual prevalence of the QRPs. Third, if average admission estimates were accurate, then actual prevalence could be estimated using the ratios of admission rates to admission estimates. This would correspond to a case in which respondents did not know the actual prevalence of a practice but did have a good sense of how likely it is that a colleague would admit to it in a survey. The three estimates should converge if the self-admission rate equaled the prevalence estimate multiplied by the admission estimate. To the extent that this equality is violated, there would be differences between prevalence rates measured by the different methods.

Results

Raw self-admission rates, prevalence estimates, prevalence estimates derived from the admission estimates (i.e., self-admission rate/admission estimate), and geometric means of these three percentages are shown in Figure 1. For details on our approach to analyzing the data, see Data Analysis in the Supplemental Material.

Truth-telling incentives

A priori, truth-telling incentives (as provided in the BTS condition) should affect responses in proportion to the baseline (i.e., control condition) level of false denials. These baseline levels are unknown, but one can hypothesize that they should be minimal for impersonal estimates of prevalence and admission, and greatest for personal admissions of unethical practices broadly judged as unacceptable, which represent "red-card" violations.

As hypothesized, prevalence estimates (see Table S2 in the Supplemental Material) and admission estimates (see Table S3 in the Supplemental Material) were comparable in the two conditions, but self-admission rates for some items (Table 1), especially those that were "more questionable," were higher in the BTS condition than in the control condition. (Table 1 also presents the p values of the likelihood ratio test of the difference in admission rates between conditions.)

We assessed the effect of the BTS manipulation by examining the odds ratio of self-admission rates in the BTS condition to self-admission rates in the control condition. The odds ratio was high for one practice (falsifying data), moderate for three practices (premature stopping of data collection, falsely reporting a finding as expected, and falsely claiming that results are unaffected by certain variables), and negligible for the remainder of the practices (Table 1). The acceptability of a practice can be inferred from the self-admission rate in the control condition (baseline) or assessed directly by judgments of defensibility. The nonparametric correlation of BTS impact, as measured by odds ratio, with the baseline self-admission rate was –.62 ($p < .06$; parametric correlation = $-.65$, $p < .05$); the correlation of odds ratio with defensibility rating was –.68 ($p < .03$; parametric correlation = $-.94$, $p < .001$). These correlations were more modest when Item 10 ("Falsifying data") was excluded (odds ratio with baseline self-admission rate: nonparametric correlation = $-.48$, $p < .20$; parametric correlation = $-.59$, $p < .10$; odds ratio with defensibility rating: nonparametric correlation = $-.57$, $p < .12$; parametric correlation = $-.59$, $p < .10$).

Prevalence estimates

Figure 1 displays mean prevalence estimates for the three types of responses in the BTS condition (the admission estimates were capped at 100%; they exceeded 100% by a small margin for a few items). The figure also shows the geometric means of all three responses; these means, in effect, give equal credence to the three types of answers. The raw admission rates are 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 geometric means of the three variables that nearly 1 in 10 research psychologists has introduced false data into the scientific record (Items 5 and 10) and
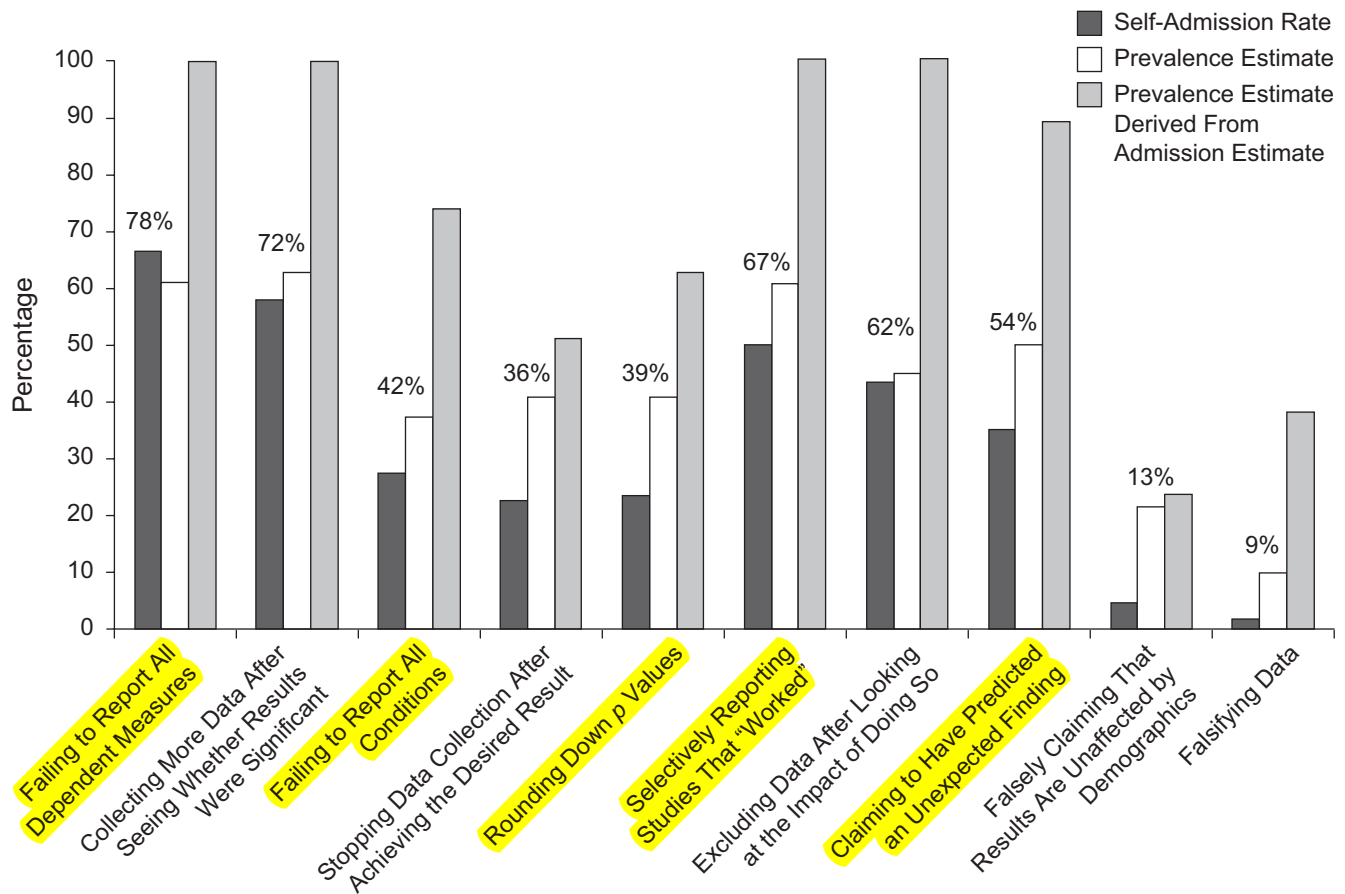
Fig. 1. Results of the Bayesian-truth-serum condition in the main study. For each of the 10 items, the graph shows the self-admission rate, prevalence estimate, prevalence estimate derived from the admission estimate (i.e., self-admission rate/admission estimate), and geometric mean of these three percentages (numbers above the bars). See Table 1 for the complete text of the items.

that the majority of research psychologists have engaged in practices such as selective reporting of studies (Item 6), not reporting all dependent measures (Item 1), collecting more data after determining whether the results were significant (Item 2), reporting unexpected findings as having been predicted (Item 8), and excluding data post hoc (Item 7).

These estimates are somewhat higher than estimates reported in previous research. For example, a meta-analysis of surveys—none of which provided incentives for truthful responding—found that, among scientists from a variety of disciplines, 9.5% of respondents admitted to having engaged in QRPs other than data falsification; the upper-boundary estimate was 33.7% (Fanelli, 2009). In the present study, the mean self-admission rate in the BTS condition (excluding the data-falsification item for comparability with 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 (compared with 91.4% in the control

condition). The self-admission rate in our control condition (33.0%) mirrored the upper-boundary estimate obtained in Fanelli's meta-analysis (33.7%).

Response to a given item on our survey 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 respondent had also engaged in 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 or she responded affirmatively—was .80 (high values indicate close agreement; items are considered to form a Guttman scale if reproducibility is .90 or higher; Guttman, 1974). This finding suggests that researchers' engagement in or avoidance of specific QRPs is not completely idiosyncratic. It indicates that there is a rough consensus among researchers about the relative unethicality of the behaviors, but large variation in where researchers 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, our findings showed that respondents who admitted to a QRP tended to think that their actions were defensible. The overall mean defensibility rating of practices that respondents acknowledged having engaged in was 1.70 ($SD = 0.53$)—between possibly defensible and defensible. Mean judged defensibility for each item is shown in Table 1. Defensibility ratings did not generally differ according to the respondents' discipline or the type of research they conducted (see Table S4 in the Supplemental Material).

Doubts about research integrity

A large percentage of respondents indicated that they had doubts about research integrity on at least one occasion (Fig. 2). The degree of doubt differed by target; for example, respondents were more wary of research generated by researchers at other institutions than of research 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 own set of collaborators), it is noteworthy that approximately 35% of respondents indicated that they had doubts about the integrity of their own research on at least one occasion.

Frequency of engagement

Although the prevalence estimates obtained in the BTS condition are somewhat higher than previous estimates, they do not enable us to distinguish between the researcher who routinely engages in a given behavior and the researcher who has only engaged in that behavior once. To the extent that self-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 133 attendees of an annual conference of behavioral researchers whether they had engaged in each of 25 different QRPs (many of which we also inquired about in the main study). Using a 2×2 between-subjects design, we manipulated the wording of the questions and the response scale. The questions were either phrased as a generic action ("Falsifying data") or in the first person ("I have falsified data"), and participants indicated whether they had engaged in the behaviors using either a dichotomous response scale (yes/no, as in the main study) or a frequency response scale (*never, once or twice, occasionally, frequently*).

Because the overall self-admission rates to the individual items were generally similar to those obtained in the main study, we do not report them here. Respondents made fewer affirmative admissions on the dichotomous response scale ($M = 3.77$ out of 25, $SD = 2.27$) than on the frequency response scale ($M = 6.02$ out of 25, $SD = 3.70$), $F(1, 129) = 17.0$, $p < .0005$. This
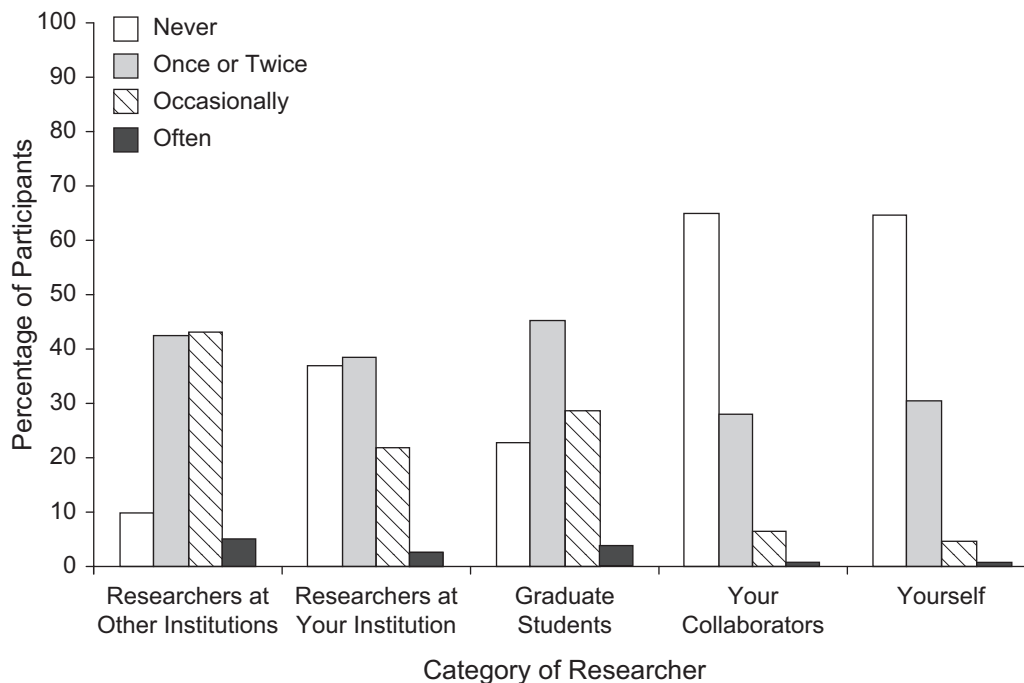

Fig. 2. Results of the main study: distribution of responses to a question asking about doubts concerning the integrity of the research conducted by various categories of researchers.

result suggests that in the dichotomous-scale condition, some nontrivial fraction of respondents who engaged in a QRP only a small number of times reported that they had never engaged in it. This suggests that the prevalence rates obtained in the main 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 (Fig. 3). Among the affirmative responses in the frequency-response-scale condition (i.e., responses of *once or twice, occasionally*, or *frequently*), 64% (i.e., .153/ (.151 + .062 + .023)) of the affirmative responses fell into the *once or twice* category, a nontrivial percentage fell into *occasionally* (26%), and 10% fell into *frequently*. This result suggests that the prevalence estimates from the BTS study represent a combination of single-instance and habitual engagement in the behaviors.

Subgroup differences

Table 2 presents self-admission rates as a function of disciplines within psychology and the primary methodology used in research. Relatively high rates of QRPs were self-reported among the cognitive, neuroscience, and social disciplines, and among researchers using behavioral, experimental, and laboratory methodologies (for details, see Data Analysis in the Supplemental Material). Clinical psychologists reported relatively low rates of QRPs.

These subgroup differences could reflect the particular relevance of our QRPs to these disciplines 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 participants in the main study, which asked them to rate two aspects of the same 10

QRPs. First, they were asked to rate the extent to which each practice applies to their research methodology (i.e., how frequently, if at all, they encountered the opportunity to engage in the practice). The possible responses were *never applicable, sometimes applicable, often applicable*, and *always applicable*. Second, they were asked whether it is generally defensible to engage in each practice. The possible responses were *indefensible, possibly defensible*, and *defensible*. Unlike in the main study, in which respondents were asked to provide a defensibility rating only if they had admitted to having engaged in a given practice, all respondents in the follow-up survey were asked to provide these ratings. We counterbalanced the order in which respondents rated the two dimensions. There were 504 respondents, for a response rate of 35%. Of respondents who began the survey, 65 (12.9%) did not complete it; as in the main study, data from all respondents—even those who did not finish the survey—were included in the analysis because the QRPs were presented in randomized order.

Table 2 presents the results from the follow-up survey. The subgroup differences in applicability ratings and defensibility ratings were partially consistent with the differences in self-reported prevalence: Most notably, mean applicability and defensibility ratings were elevated among social psychologists—a subgroup with relatively high self-admission rates. Similarly, the items were particularly applicable to (but not judged to be more defensible by) researchers who conduct behavioral, experimental, and laboratory research.

To test for the relative importance of applicability and defensibility ratings in explaining subfield differences, we conducted an analysis of variance on mean self-admission rates across QRPs and disciplines. Both type of QRP ($p < .001$, $\eta_p^2 = .87$) and subfield ($p < .05$, $\eta_p^2 = .21$) were highly significant predictors of self-admission rates, and their significance and effect size were largely unchanged after controlling for applicability and defensibility ratings, even though both of the
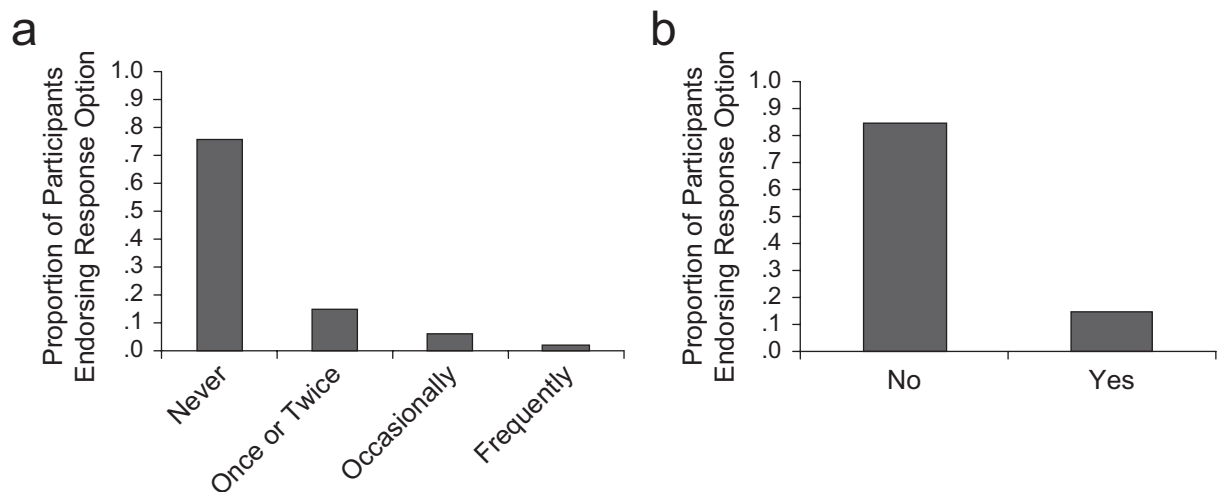

Fig. 3. Results of the follow-up study: distribution of responses among participants who were asked whether they had engaged in 25 questionable research practices. Participants answered using either (a) a frequency response scale or (b) a dichotomous response scale.

Table 2. Mean Self-Admission Rate, Applicability Rating, and Defensibility Rating by Category of Research

Category of research	Self-admission rate (%)	Applicability rating	Defensibility rating
Discipline			
Clinical	27*	2.59 (0.94)	0.56 (0.28)
Cognitive	37***	2.75* (0.93)	0.64 (0.23)
Developmental	31	2.77** (0.89)	0.66 (0.27)
Forensic	28	3.02* (1.12)	0.52 (0.29)
Health	30	2.56 (0.94)	0.69 (0.31)
Industrial organizational	31	2.80 (0.63)	0.73 (0.30)
Neuroscience	35**	2.71 (0.92)	0.61 (0.21)
Personality	32	2.65* (0.92)	0.66 (0.36)
Social	40***	2.89*** (0.85)	0.73** (0.31)
Research type			
Clinical	30	2.61 (0.99)	0.56 (0.27)
Behavioral	34*	2.77** (0.88)	0.63 (0.28)
Laboratory	36***	2.87*** (0.86)	0.66 (0.29)
Field	31	2.76** (0.88)	0.63 (0.28)
Experimental	36***	2.83* (0.87)	0.66* (0.29)
Modeling	33	2.74 (0.89)	0.62 (0.26)

Note: Self-admission rates are from the main study and are collapsed across all 10 items; applicability and defensibility ratings are from the follow-up study. Applicability was rated on a 4-point scale (1 = *never applicable*, 2 = *sometimes applicable*, 3 = *often applicable*, 4 = *always applicable*). Defensibility was rated on a 3-point scale (0 = *no*, 1 = *possibly*, 2 = *yes*). For self-admission rates, random-effects logistic regression was used to identify significant effects; for applicability and defensibility ratings, random-effects ordered probit regressions were used to identify significant effects.
*p < .05. **p < .01. ***p < .0005.

latter variables were highly significant independent predictors of mean self-admission rates. Similarly, methodology was also a highly significant predictor of self-admission rates (p < .05, η_p^2 = .27), and its significance and effect size were largely unchanged after controlling for applicability and defensibility ratings (even though the latter were highly significant predictors of self-admission rates).

The defensibility ratings obtained in the main study stand in contrast with those obtained in the follow-up survey: Respondents considered these behaviors to be defensible when they engaged in them (as was shown in the main study) but considered them indefensible overall (as was shown in the follow-up study).

Discussion

Concerns over scientific misconduct have led previous researchers 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, we designed this study to provide accurate estimates of the prevalence of QRPs that are specifically applicable to research psychologists. In addition to being one of the first studies 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 percentage of psychologists admit to having engaged in QRPs. The effect of the BTS manipulation on self-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. This 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 two primary components to the BTS procedure—both a request and an incentive to tell the truth—and we were unable to isolate their independent effects on disclosure. 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 was based on the premise 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; Lensvelt-Mulders, Hox, van der Heijden, & Maas, 2005; Tourangeau & Yan, 2007; Warner, 1965). This assumption is generally accepted, provided that the behaviors in question are sensitive 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 gray area of acceptable practices. Although falsifying data (Item 10 in our study) is never justified, the same cannot be said for all of the items on our survey; 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 self-admissions represent scientific felonies, or even misdemeanors; some respondents provided perfectly defensible reasons for engaging in the behaviors. Yet other respondents provided justifications that, 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 and thus increased the likelihood of publication). It is worth noting, however, that in the follow-up survey—in which participants rated the behaviors regardless of personal engagement—the defensibility ratings were low. This suggests 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. Furthermore, most of the respondents in our study believed in the integrity of their own research and judged practices they had engaged in to be acceptable. However, given publication pressures and professional ambitions, the inherent ambiguity of the defensibility of "questionable" research practices, and the well-documented ubiquity of motivated reasoning (Kunda, 1990), researchers may not be in the best position to judge the defensibility of their own behavior. This 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 generate a post hoc explanation to justify removing nuisance data points than it is to justify outright data falsification, even though both practices produce similar consequences.

Given the findings of our study, it comes as no surprise that many researchers have expressed concerns over failures to replicate published results (Bower & Mayer, 1985; Crabbe, Wahlsten, & Dudek, 1999; Doyen, Klein, Pichon, & Cleeremans, 2012; Enserink, 1999; Galak, LeBoeuf, Nelson, & Simmons, 2012; Ioannidis, 2005a, 2005b; Palmer, 2000; Steele, Bass, & Crook, 1999). In an article on the problem of nonreplicability, Lehrer (2010) discussed possible explanations for the "decline effect"—the tendency for effect sizes to decrease with subsequent attempts at replication. He concluded that conventional accounts of this effect (regression to the mean, publication bias) may be incomplete. In a subsequent and insightful commentary, Schooler (2011) suggested that unpublished data may help to account for the decline effect. By documenting the surprisingly large percentage of researchers who 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. Simmons and his colleagues (2011) went further by showing how easily QRPs can yield invalid findings and by proposing reforms in the process of reporting research and accepting scientific manuscripts for publication.

QRPs can waste researchers' time and stall scientific progress, as researchers fruitlessly pursue extensions of effects that are not real and hence cannot be replicated. More generally, the prevalence of QRPs raises questions about the credibility of research findings and threatens research integrity by producing 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 would effectively reduce the prevalence of QRPs, they not only would bolster scientific integrity but also could reduce the pressure on researchers to produce unrealistically elegant results.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Supplemental Material

Additional supporting information may be found at http://pss.sagepub .com/content/by/supplemental-data

References

Altman, L. K. (2006, May 2). For science gatekeepers, a credibility gap. *The New York Times*. Retrieved from http://www.nytimes .com/2006/05/02/health/02docs.html?pagewanted=all

Bower, G. H., & Mayer, J. D. (1985). Failure to replicate mood-dependent retrieval. *Bulletin of the Psychonomic Society*, *23*, 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.

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

Crocker, J. (2011). The road to fraud starts with a single step. *Nature*, *479*, 151.

Deer, B. (2011). How the case against the MMR vaccine was fixed. *British Medical Journal*, *342*, 77–82.

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*, 14–27.

Doyen, S., Klein, O., Pichon, C.-L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLoS ONE*, *7*(1), e29081. Retrieved from http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0029081

Enserink, M. (1999). Fickle mice highlight test problems. *Science*, *284*, 1599–1600.

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS One*, *4*(5), e5738. Retrieved from http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0005738

Galak, J., LeBoeuf, R. A., Nelson, L. D., & Simmons, J. P. (2012). Correcting the past: Failures to replicate psi. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2001721

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

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

Guttman, L. L. (1974). The basis for scalogram analysis. In G. M. Maranell (Ed.), *Scaling: A sourcebook for behavioral scientists* (pp. 142–171). New Brunswick, NJ: Transaction.

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

Ioannidis, J. P. A. (2005a). Contradicted and initially stronger effects in highly cited clinical research. *Journal of the American Medical Association*, *294*, 218–228.

Ioannidis, J. P. A. (2005b). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. Retrieved from http://www.plosmedicine.org/article/info:doi/10.1371/journal.pmed.0020124

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

Lacetera, N., & Zirulia, L. (2011). The economics of scientific misconduct. *The Journal of Law, Economics, & Organization*, *27*, 568–603.

Lehrer, J. (2010, December 13). The truth wears off. *The New Yorker*. Retrieved from http://www.newyorker.com/reporting/2010/12/13/101213fa_fact_lehrer

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: Thirty-five years of validation. *Sociological Methods & Research*, *33*, 319–348.

List, J., Bailey, C., Euzent, P., & Martin, T. (2001). Academic economists behaving badly? A survey on three areas of unethical behavior. *Economic Inquiry*, *39*, 162–170.

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

Martinson, B. C., Anderson, M. S., & de Vries, 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. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366.

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*, 366–369.

Steneck, N. H. (2002). Assessing the integrity of publicly supported research. In N. H. Steneck & M. D. Scheetz (Eds.), *Investigating Research Integrity: Proceedings of the First ORI Research Conference on Research Integrity* (pp. 1–16). Washington, DC: Office of Research Integrity.

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

Sterba, S. K. (2006). Misconduct in the analysis and reporting of data: Bridging methodological and ethical agendas for change. *Ethics & Behavior*, *16*, 305–318.

Swazey, J. P., Anderson, M. S., & Louis, K. S. (1993). Ethical problems in academic research. *American Scientist*, *81*, 542–553.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, *133*, 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.

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, *60*, 63–69.

Wicherts, J. (2011). Psychology must learn a lesson from fraud case. *Nature*, *480*, 7.