Social Psychological and Personality Science
2016, Vol. 7(1) 45-52

DOI: 10.1177/1948550615612150
spps.sagepub.com
SAGE

Questionable Research Practices Revisited

Klaus Fiedler[1] and Norbert Schwarz[2]

Abstract

The current discussion of questionable research practices (QRPs) is meant to improve the quality of science. It is, however, important to conduct QRP studies with the same scrutiny as all research. We note problems with overestimates of QRP prevalence and the survey methods used in the frequently cited study by John, Loewenstein, and Prelec. In a survey of German psychologists, we decomposed QRP prevalence into its two multiplicative components, proportion of scientists who ever committed a behavior and, if so, how frequently they repeated this behavior across all their research. The resulting prevalence estimates are lower by order of magnitudes. We conclude that inflated prevalence estimates, due to problematic interpretation of survey data, can create a descriptive norm (QRP is normal) that can counteract the injunctive norm to minimize QRPs and unwantedly damage the image of behavioral sciences, which are essential to dealing with many societal problems.

Keywords

ethics/morality, language, research methods, survey methodology, research practices

In a recent, widely circulated article, John, Loewenstein, and Prelec (2012) concluded that 10 "questionable research practices" (QRPs) are common in psychology, with occurrence rates up to 100%. This message found attention in the national (Carey, 2011) and international (van Maanen, 2012) press and at scientific conferences. Apparently, academic psychologists estimated QRP to be common among their colleagues and admitted that they themselves engage in these practices more often than one might expect. This self-critical analysis of psychology found wide agreement as a timely and responsible contribution to the current debate on the quality of science (Lilienfeld, 2010).

An article that is deeply concerned with (violations of) good scientific practice deserves itself to be treated at the highest level of scientific scrutiny, if only to avoid the ironic effect that communicating an unfortunate descriptive norm (almost everybody violates norms of good scientific practice anyway) undermines a desirable injunctive norm (scientists must not violate rules of good scientific practice). The side effects and the collateral damage caused by descriptive norms have been vividly demonstrated by Cialdini and colleagues (Cialdini, 2007; Schultz, Nolan, Cialdini, Goldstein, & Griskevicius, 2014).

Critical examination of the John et al. (2012) survey reveals problems with its internal and external validity, casting its strong conclusions into doubt. In the present article, we first explain the nature of these validity problems. With respect to established rules of survey methodology, we point out that many survey items were ambiguous and that, according to the logic of conversation, respondents could not fully deny having ever engaged in any behavior and they had no chance to communicate good reasons for committing some behaviors. Moreover, with respect to the logically sound interpretation of empirical data, we show that mistaking the proportion of individuals who ever engaged in a behavior as a measure of the behavior's prevalence can lead to misunderstandings.

We then report the results of a new survey, in which less ambiguous questions and less misleading response formats led to radically lower prevalence estimates for the same 10 QRPs. Because overestimating QRPs can be counterproductive and harmful, we believe it is important, and an ethical act in its own right, to rectify any misleading inferences on a topic as serious as QRPs.

Problems With QRP Questions

Logic of conversation. An essential criterion for valid survey data is that questions must be unambiguous and the response format must not obscure the intended communication of self-reports. Several QRP questions used by John et al. (2012) do not meet this criterion. The behavioral references of some items remain ambiguous, and the question format does not give participants a chance to communicate the reasons for committing certain behaviors that may not constitute questionable practices per se. To illustrate, *failing to report all of a study's dependent*

[1] University of Heidelberg, Heidelberg, Germany
[2] University of Southern California, Los Angeles, CA, USA

Corresponding Author:
Klaus Fiedler, University of Heidelberg, Hauptstrasse 47-51, Heidelberg, 69117, Germany.
Email: kf@psychologie.uni-heidelberg.de

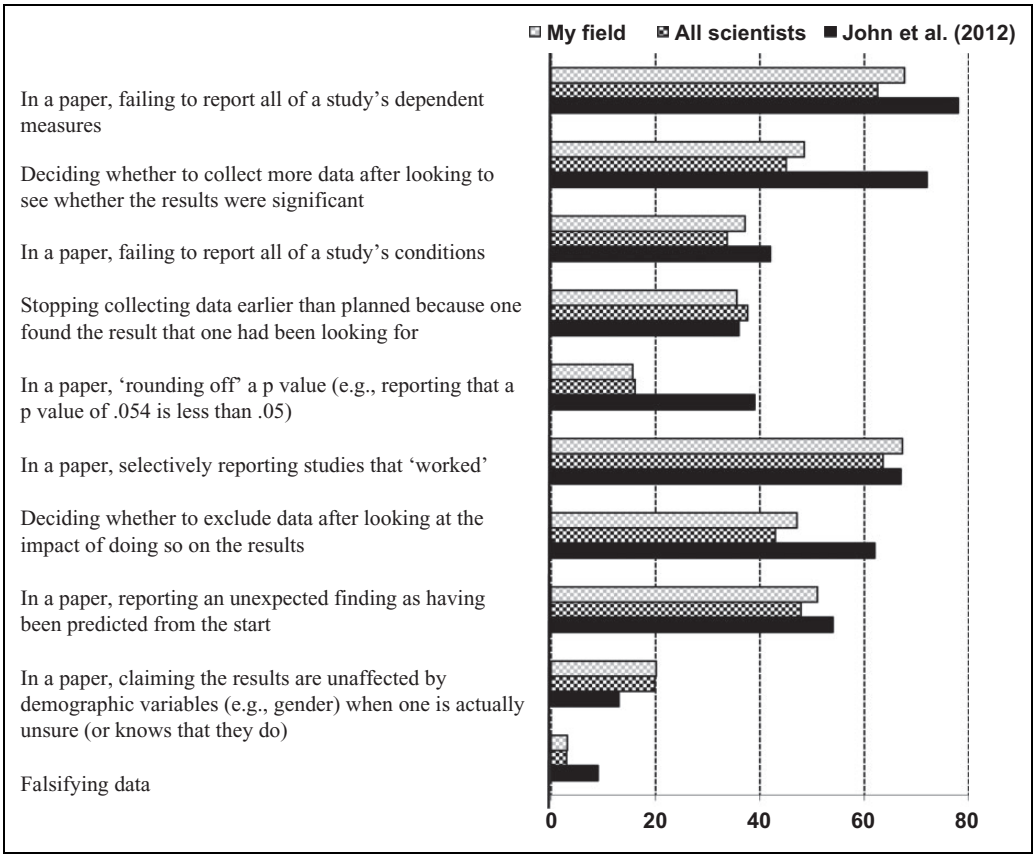

Figure 1. Proportions of respondents admitting to have engaged in questionable research practices (QRPs) at least once in the original by John, Loewenstein, and Prelec's (2012) survey (black bars) along with Dutch graduate students' estimates of QRP prevalence (across all studies conducted) by all scientists (dark dotted bars) and by scientists in their own field (light dotted bars).

measures, the item with the highest admission rate (cf. black bars in Figure 1) may indeed refer to intentionally concealing unwanted results regarding the tested hypothesis. However, it may as well refer to not reporting subsidiary results of a post-experimental interview, results of irrelevant extra analyses broken down by demographic variables, invariance test for monotonic transformations, or conscientious checks on reliability, scale level, factor structures, or an endless list of variables generated in computer simulations. Some respondents—indeed the most careful and idealistic ones—may want to express proudly that they always run a large number of extra analyses, many more than can be included in any article, providing a "Yes" response to this item and not believing that what they are doing will ever count as QRP. They may indeed be convinced that every careful researcher is obliged to arrange and analyze the data in as many ways as possible, thus producing countless derived measures that cannot all be published. To reduce the equivocality, a minimal amendment would be to reformulate the item as *failing to report all dependent measures that are relevant for a finding* (see Table 1).

Still, such a reformulation would by no means prevent all "innocent" researchers from responding yes. Being a cooperative communication partner, a conscientious respondent may know—and know that the survey researcher knows—that it is impossible to publish every dependent measure, for many reasons unrelated to the motive to augment one's empirical results. Neither editors nor reviewers nor the readership of a journal want to read a boring report of all subsidiary analyses conducted in the course of a study. The respondent will hardly feel the need to indicate external reasons for not reporting everything. But if they had tried to do so in a comment box, these data would have been ignored by John et al. (2012). They rather coded every yes response to the question as evidence for QRP, regardless of the reference behavior and its internal or external attribution.

The item *selectively reporting studies that "worked,"* which also solicited a very high admission rate, looks like a lie-scale item, because hardly any honest respondent could say "no." Even the most renowned researchers will be given no journal space to report studies that did not work, for whatever reason. Every skilled experimenter will have conducted pilot studies to try out manipulations and instructions, instruments, and dependent measures. To be sure, responding yes may of course refer to an act of negatively motivated concealment. However, it may as well refer to completely normal practices that are part and parcel of all empirical research. To reduce equivocality, one might at least modify the item to *selectively reporting studies related to a specific finding that "worked."*

Table 1. Wording of Questionnaire Items Used to Assess QRPs.

Original Items Used by John, Loewenstein, and Prelec (2012)	Translated and Edited Wording	Type of Equivocality to be Reduced by Modification
[In a paper,] failing to report all of a study's dependent measures *that are relevant for a finding*	Nicht alle abhängigen Messungen berichten, *die für einen Befund relevant sind*	Fully "innocent" unreported measures (e.g., incidental side effects of careful data analysis)
[Deciding whether to] collect more data *in order to render non-significant results* [after looking to see whether the results were] significant	Zusätzliche Daten erheben, *um noch nicht signifikante Ergebnisse zu erwünschtem Befund signifikant zu bekommen*	Temporal relation ("after") may be harmless (e.g., in the context of sequential analysis (Wald, 1947)
[In a paper,] failing to report all of a study's conditions *that are relevant for a finding*	Nicht alle Bedingungen einer Untersuchung berichten, *die für einen Befund relevant sind*	There may be good reasons not to report conditions (e.g., from later stages) that are irrelevant
Stopping collecting data earlier than planned because [one found the result that one had been looking for] the expected result concerning a specific finding were already obtained	Die Datenerhebung eher als geplant abbrechen, weil das erwartete Ergebnis *zu einem bestimmten Befund* schon erreicht war	Avoiding presupposition that the researcher is looking for particular results rather than observing open-mindedly
[In a paper,] 'rounding off' a *p* value (e.g. reporting that a *p* value of .054 is less than .05)	Den berichteten *p*-Wert abrunden (z.B. einen p-Wert von .054 als .05 berichten)	
[In a paper,] selectively reporting studies *related to a specific finding* that 'worked'	Selektiv solche Studien *zu einem bestimmten Befund* berichten, die 'funktioniert' haben	Introduction and discussion sections always concentrate on prior studies that worked
Deciding whether to exclude data after looking at the impact of doing so *on the desired results*	Über den Ausschluss von Daten entscheiden, nachdem man die Auswirkung dieser Maßnahme *auf den erwünschten Befund* geprüft hat	Checking on exclusion effects is not problematic, unless exclusion is contingent on its impact on desired results
[In a paper,] reporting an unexpected finding as having been predicted from the start	Einen unerwarteten Befund so berichten, als ob er von Beginn an vorhergesagt worden wäre	
In a paper, claiming that results are unaffected by demographic variables (e.g. gender) [when] although one is actually unsure (or knows that they do)	Die Unabhängigkeit eines Befundes von demografischen Variablen (z.B. Geschlecht) mitteilen, obwohl man das eigentlich nicht weiß (oder gar das Gegenteil weiß)	The purpose for replacing "when" by "although" is to clarify that respondents are actually aware of their being unsure
Falsifying data	Daten fälschen	

Note. The original items (left column, including the phrases in brackets and excluding the phrases in italics) were translated to German and modified to include the phrases in italics and to exclude the phrases in brackets (middle column). Reasons for modification are summarized in the right column. QRP = questionable research practice.

Likewise, some reflection on the third item with a very high rate of yes responses, *deciding whether to collect more data after looking to see whether the results were significant*, leads to the insight that one cannot honestly respond no. That further data collection depends on previous results is a truism, and there can be no logical and moral rule to stop data collection after a single nonsignificant test. The item not even distinguishes between data collection within the same study or in a new study. Whether the referent behavior constitutes a QRP or not is unclear. It is possible but by no means necessary.

In standard survey practice, such ambiguities would have been identified at the questionnaire development stage (Schwarz, 1994; Schwarz & Sudman, 1996; Willis, 2005) and rectified. Moreover, the state-of-the-art methodology would not allow survey researchers to equate yes responses with positive (aggravating) evidence for QRP, such that a QRP-free respondent must provide a constant no response across all 10 items. The tacit norms of cooperative conversation (Grice, 1975) discourage respondents from providing the same invariant response to every item of a questionnaire (Schwarz, 1994), further increasing the likelihood that cooperative respondents will respond yes to a sizable subset of items.

Which "Prevalence"?

However, the aforementioned notes on survey design and logic of conversation are only the first part of our critique. The most serious validity problem arises in the interpretation of the quantitative measures assessed by John et al. (2012). Although the crucial dependent measure is the *proportion of people* who admits to have committed specific behaviors once in their life, these proportions are then interpreted as if they reflect the *prevalence of those behaviors*, as indicated in the title of the target article. From high ''proportions of respondents that have [once in their life] engaged in these practices,'' John et al. draw pessimistic conclusions that ''some questionable practices may constitute the prevailing research norm'' (p. 524) and that ''the prevalence of QRPs raises questions about the credibility of research findings and threatens research integrity'' (p. 531).

It remains unexplained, though, how the prevalence of behaviors can be inferred from proportions of people who showed

behaviors at least once in their life. The two quantities are entirely distinct. There is no logically sound rule to derive the prevalence of behaviors from the proportions of people who ever engaged in these behaviors, and the two statistics can diverge by an order of magnitudes. What does the proportion of people who ever told a lie in their life reveal about the prevalence of lying? How can the proportion of people who ever entered a church (presumably 100%) be used to infer the prevalence of church attendance? The failure to distinguish these fundamentally different quantities may have promoted estimates of QRP prevalence that are inflated to an unknown degree.

Logically, to estimate QRP prevalence, it would be necessary to multiply the proportion of respondents who committed some QRP at least once with another measure, namely, the rate with which researchers have repeated this behavior across all studies they have conducted. To illustrate, if the proportion of people who ever told a lie in their life is 100% and the rate of lies among all utterances is, say, 1%, then the prevalence of lying is not 100% but the product of 100% × 1% amounts to only 1% (or, on a probability scale, $1 \times .01 = .01$). By analogy, if the proportion of researchers who did not report all measures in at least one study is 60%, and if that behavior was exhibited in, say, 10% of all conducted studies, then the prevalence is not 60% but 60% × 10% = 6%, an estimate in a different order of magnitude. In any case, the data obtained by John et al. (2012) do not allow any estimates of QRP prevalence simply because they did not assess the logically necessary repetition rates.

To assess the empirical consequences of this important distinction, we conducted a new survey in which respondents were not merely asked to indicate whether or not they ever engaged in the 10 QRPs but also to estimate the probability with which they exhibited these QRPs across all their studies. According to the preceding analysis, we expected the prevalence estimates derived from this new survey to be radically lower than estimates derived from the incomplete John et al. data.

Reporting on Others' Behavior

Whereas the preceding issues pertained to self-reports of respondents' own behavior, the John et al. (2012) survey also included questions about others' behavior. Specifically, respondents were asked to estimate "the prevalence of the practice by estimating the percent of research psychologists that have engaged in the practice on at least one occasion" (John, Loewenstein, & Prelec, 2012, supplemental materials). This request poses a formidable task: How would people go about computing this estimate?

The John et al. (2012) data show a conspicuous convergence of nonzero self-admission rates and other-related prevalence estimates (cf. black and open bars in figure 1 of John et al., 2012). Did respondents have such excellent knowledge of their peers' behavior and their self-admission rates, apparently allowing them to correctly anticipate the entire survey results? More likely, respondents found themselves unable to determine the rate of psychologists with nonzero records and resorted to simpler heuristics, for example, a crude sense of familiarity with the topic of each QRP item (Hintzman & Curran, 1994) or some simulation heuristic used to judge the plausibility of behaviors (Kahneman & Tversky, 1982). If so, this would cast further doubts on the informational value of the reported results.

For a first check on the suspicion that only crude heuristic judgments may be at work, we asked 35 advanced graduate students in the Netherlands, who attended a methods training program, to provide prevalence judgments using task instructions that diverged markedly from the instructions provided by John et al. (2012). Graduate students were explicitly reminded that their task was not to estimate how many scientists have engaged in the various practices at least once. Rather they were instructed to estimate "the prevalence (in %) of this behavior, across all hypothesis tests conducted by all behavioral scientists" and "by all scientists in your field." The precise wording is given in the supplements. In an oral comment, participants' attention was deliberately drawn to this aspect, inviting them to focus on the rate of hypothesis tests rather than the rate of scientists with a nonzero count. This instruction should have led to considerably lower estimates because it increases the reference set in the denominator of the estimation from the number of all researchers to the product of the number of all researchers times the average number of studies they have conducted.

Nevertheless, the resulting mean estimates were again very close to those obtained with different instructions in the original study. Both prevalence estimates obtained here, for "all behavioral scientists" and for "my field" (dark and light dotted bars in Figure 1), were nearly identical to the proportions of researchers who admitted to ever having engaged in the respective behavior in John et al. (2012). This curious convergence seems to support our suspicion that one should be cautious taking the metric properties of the results for granted.

Decomposing QRP-Prevalence Into Its Multiplicative Components: A New Survey Study

Having pointed out the fundamental difference between researcher proportions and behavioral prevalence and having provided some anecdotal evidence that the reported consensus estimates can be hardly trusted, we now present more systematic empirical results from a new survey study. In this survey, we refrained from asking other-referent prevalence questions that may not solicit informed judgments. With reference to all 10 QRPs, we only asked respondents to provide two self-referent judgments: (1) whether they had ever engaged in the corresponding behavior and, if the answer was yes, (2) how frequently they did so. Confining the survey to self-referent knowledge and increasing the sensitivity to task differences by including two judgments in a repeated-measures setting (Fischhoff, Slovic, & Lichtenstein, 1979) should facilitate a better understanding of the task. We did not expect this new

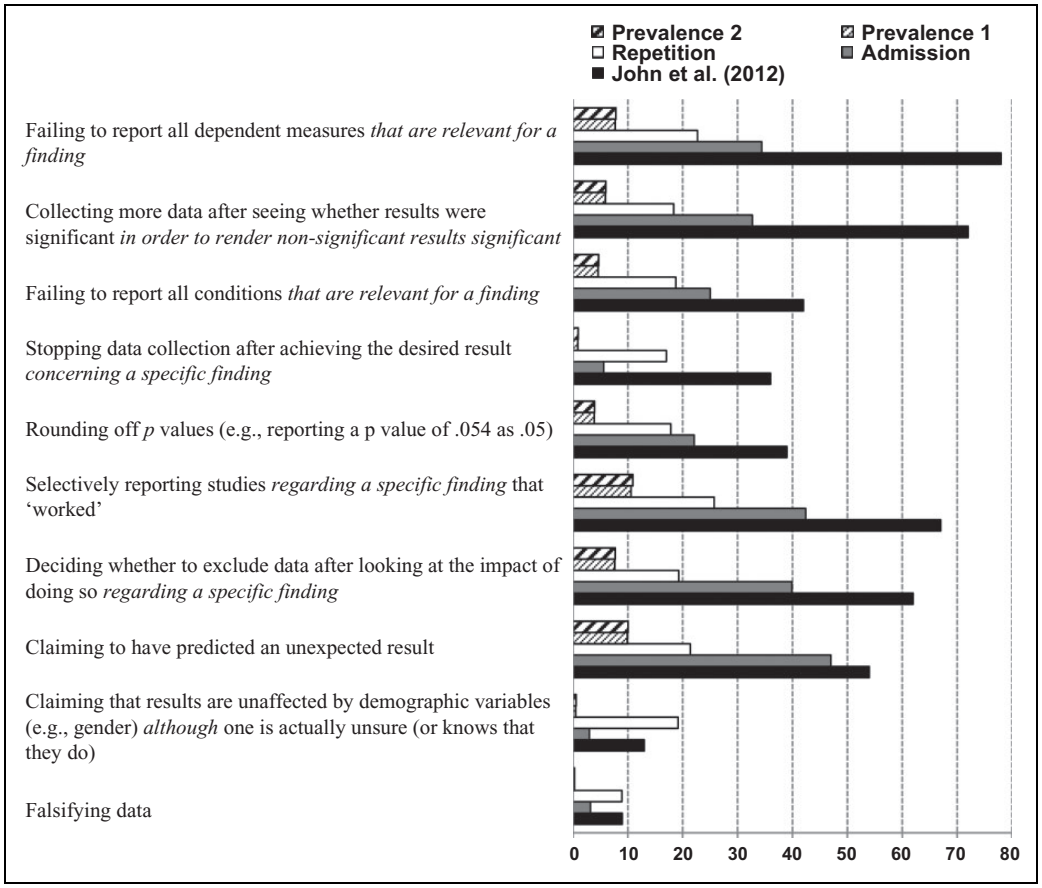

Figure 2. Prevalence indices (shaded bars) derived from admission rates of respondents committing questionable research practices at least once (gray bars) and repetition frequency (white bars), compared to the original John et al. (2012) data (black bars). Modified item wordings appear in italics.

study to reveal true, completely unbiased QRP rates, to be sure. Survey self-reports can never fully rule out ambiguities in meaning, limitations in autobiographical memory, or motivated biases. Regardless of how close the obtained results reflect the true QRP prevalence, the primary purpose here was to demonstrate that a logically appropriate decomposition of two multiplicative components of QRP will result in substantially lower prevalence estimates than those reported by John et al. (2012).

Method

A total of 1,138 members of the Deutsche Gesellschaft für Psychologie (German Psychological Association) followed our invitation to participate in a web survey. This sample encompasses roughly 35% of approximately 3,200 association members.

The complete set of QRP items is given in Table 1. Although 7 of the 10 items were slightly modified to reduce their equivocality, these modifications left the essential meaning of all 10 QRP items unchanged and only served to specify the behavioral domain, typically by adding the phrase "relevant for a specific finding."

Participants were then presented with 1 QRP item at a time on the computer screen, and they were asked to indicate, first, if they had ever engaged in the behavior under question and, second, only if the answer was "yes," in what percentage of all their published findings they had done so. The full instruction text (originally in German) is provided in the supplements, along with English translations.

From the responses to these two questions, QRP prevalence could be estimated as all respondents' average repetition percentage after setting no-responders' percentage to zero (Prevalence 1 in Figure 2). Virtually identical results were obtained by multiplying for each item the proportion of yes responders by the average repetition percentage (Prevalence 2).

Analogous to the "Bayesian truth serum" in the original study, instructions referred to a stochastic lie detector (Moshagen, Musch, & Erdfelder, 2012) used to diagnose dishonest responses. Participants were told that because dishonest responses would make the survey results worthless, "we will ... validate the results using a 'Stochastic lie detector' (Moshagen et al., 2012, *Behavior Research Methods, 44*, 222–231) and discard the results if it turns out that free responses are biased."

Results

Figure 2 provides an overview of the prevalence estimates (shaded bars) along with the corresponding nonzero admission proportions (gray bars) and repetition percentages (white bars) and the original admission rates reported by John et al. (black bars). Several observations are worth highlighting.

First, when QRP prevalence is computed as the product of the admission proportion of yes-responders and the average repetition percentage, the resulting prevalence estimates are indeed much lower than the admission proportions obtained here, or those obtained and interpreted as prevalence by John et al. (2012). For example, 47% of the respondents admit to having at least once claimed to have predicted an unexpected result, but the prevalence estimate is only 10%. Likewise, 22% admit to having rounded down p values at least once, but the corresponding prevalence estimate is only 3.9%. Across all 10 items, the admission rates are about 5 times higher than the prevalence estimates.[1] As expected, then, prevalence estimates diverge—by order of magnitudes—from nonzero admission proportions of respondents who ever in their life engaged in those behaviors.

Secondly, even the nonzero admission proportions obtained in the present study (gray bars) are clearly lower than the corresponding admission rates reported by John et al. (black bars). We refrain from speculating on whether this discrepancy, which appears to arise for both modified (marked in italics) and unmodified items, is due to different researcher populations (United States vs. Germany) or to subtle differences in the questioning procedure. In the absence of a sound explanation, we confine ourselves to pointing out the instability of such survey findings, which have led worldwide to such memorable conclusions about the prevalence of QRPs.

However, in spite of the marked discrepancies in the overall level of both prevalence estimates and nonzero admission rates, a third noteworthy result is that by and large the same practices seem to be identified as most common across all measures assessed in both studies. The correlation between the admission rates obtained here and those obtained in the original study is as high as $r = .84$. However, the conspicuous convergence between self-referent admission judgments (in both studies) and the judgments reported by John et al. (2012) concerning others' behaviors is hard to understand, given that these behaviors are neither precisely defined nor observable in others. In our reading, these convergences suggest that all responses may to some extent reflect the operation of stereotypical clichés about the plausibility of researcher behaviors, rather than actual personal experience.

Fourth, the items with the highest reported prevalence are also the most ambiguous ones, and their behavioral referents do not necessarily imply QRPs. Admitting to having not always reported all dependent measures can reflect justifiable behaviors or attributions to external causes (e.g., editorial requests to drop tangential material). Focusing on studies that worked may reflect the ambiguity of null results as well as the truism that only such studies have a real chance to be published. Similarly, the conditions of excluding data after looking at the impact of doing so are met by merely checking what excluding outliers does to one's conclusions—hardly an undesirable behavior. Likewise, collecting more data after initial analyses may be a sensible response when it turns out that the effect size is smaller than expected, thus requiring a larger N to reach appropriate power. Although this practice undermines the effective α in Fisherian null-hypothesis significance testing, it may also increase the a posteriori likelihood of a correct theoretical inference (Fiedler, Kutzner, & Krueger, 2012; Murayama, Pekrun, & Fiedler, 2014) and thereby reduce the costs of further studies sunk in the file drawer of unpublished research. Moreover, making further data collection contingent on a permanent update of interim results is the rule in an alternative statistical approach called sequential testing (Wald, 1947). In short, many apparent QRPs are judgment calls, and their evaluation requires more detail than the survey items allow for.

Finally, and certainly most seriously, a sizable rate of unambiguous norm violations must not be ignored. The most unpleasant result of our attempt to disambiguate survey data on scientific norm violations is that their prevalence rate is obviously not zero! Yes, some researchers admit to faking and lying, though they claim to have done so only on a small subset of the hypotheses they tested.

Discussion

Scientific fraud exists, so do unwanted research practices. Understanding the prevalence of both is an important step along the way. However, claims about violations of the standards of good science deserve to be held to the high standards they endorse, not the least in light of the damage that misleading inferences can cause. Unfortunately, one of the more widely cited publications on QRPs in psychological research (John et al., 2012) suffers from ambiguities that prohibit the damning conclusions drawn.

First, the research practices addressed in the survey constitute a convenience sample of research practices that may have been selected based on the authors' intuitions about what behaviors might be in conflict with premises of significance testing. As items were not sampled representatively (Dhami, Hertwig, & Hoffrage, 2004), the authors' world knowledge of behaviors that fit the stereotype of what scientists are doing (and what they believe other scientists are doing) may have contributed to selecting items leading to nonzero responses.

Second, the list of behaviors included practices that may or may not be justifiable depending on the specifics of the case. The survey failed to give respondents the opportunity to clarify such ambiguities. Any clarifications respondents may have tried to offer in comment boxes were not considered in the analyses. No attempt was made to ensure that respondents correctly understand the behaviors included in the questionnaire and to rule out that yes-responses may reflect unproblematic instances (e.g., not reporting all dependent measures derived

in computer simulations or as interim results of complex data analyses).

Third, the survey items posed complex memory and estimation tasks. No attempt was made to clarify whether respondents can provide meaningful answers. The observed curious convergences across different reporting tasks, shown in the current Figure 1, cast doubt on the informational value of respondents' estimates.

Fourth, conversational norms discourage repetitive no responses to a long list of items. This may have contributed to the endorsement rate of ambiguous items.

In good survey practice, all of these issues would have been addressed at the questionnaire development stage. Cognitive interviewing and related techniques (for reviews, see Schwarz & Sudman, 1996) would have alerted researchers to ambiguities of the questions, respondents' difficulties in answering them, and the likely inferential strategies used.

Fifth, and most important, the questionnaire presented estimation tasks that respondents were as unlikely to understand as many readers of the survey's results, giving rise to a category mistake at the conceptual level. Just as the prevalence of church attendance is categorically different from the proportion of people who ever went into a church, the prevalence of violations of good research practice is categorically different from the proportion of researchers who violated good practice at least once. On logical grounds, the two quantities (cf. shaded vs. black or gray bars in Figure 2) are likely to differ by an order of magnitude as the present findings confirm. This category mistake is at the heart of most media reports about the John et al. (2012) results. Unfortunately, such erroneous estimates of the prevalence of bad practice must be corrected to prevent them from damaging the public reputation of science. As numerous studies showed across different domains, communicating descriptive norms (It happens all the time) that are at odds with proscriptive norms (You should not do it) can undermine the influence of the proscriptive norm on individuals' behavior (Cialdini, 2007; Schultz et al., 2014).

In closing, we emphasize once more that assessing scientific practice and even more so efforts to improve adherence to norms of good practice are important, no doubt. Indeed, conflicts of interest can lead researchers to see some minor deviations as tolerable, as John et al. (2012) emphasize. As research in moral psychology suggests, the same holds for the moral licensing that comes with laudable goals, including the goal to clean up bad practices in science (for a more detailed discussion, see Fiedler, in press). It is therefore important to hold analyses of poor scientific practice to the high standards they advocate.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research reported in this article was supported by a Koselleck grant awarded by the Deutsche Forschungsgemeinschaft to the first author (Fi 294/23-1).

Note

1. It is interesting to note that although the correlation between nonzero admission rates (gray bars) and repetition rates (white bars) is only $r = .20$, the nonzero admission rates are almost perfectly correlated ($r = .98$) with the prevalence estimates (shaded bars). If respondents based their estimates on personal experience with practice in their own research environment, one would assume that behaviors that are judged as being often committed once are also judged as being often repeated. Admission rates should thus be strongly correlated with repetition rates so that prevalence estimates should bear a quadratic relation to both factors. This is not the case.

Supplemental Material

The online data supplements are available at http://spps.sagepub.com/supplemental.

References

Carey, B. (2011). Fraud case seen as a red flag for psychology research. *New York Times*, *3* November 2011. Retrieved from http://www.nytimes.com/2011/11/03/health/research/noted-dutch-psychologist-stapelaccused-of-research-fraud.html?sq=scientific%20fraud&st=cse&scp=4&pagewanted=print

Cialdini, R. B. (2007). Descriptive social norms as underappreciated sources of social control. *Psychometrika*, *72*, 263–268. doi:10.1007/s11336-006-1560-6

Dhami, M. K., Hertwig, R., & Hoffrage, U. (2004). The role of representative design in an ecological approach to cognition. *Psychological Bulletin*, *130*, 959–988. doi:10.1037/0033-2909.130.6.959

Fiedler, K. (in press). Ethical norms and moral values among scientists: Applying conceptions of morality to scientific rules and practices. In J. P. Forgas, P. van Lange, & L. Jussim (Eds.), *Social psychology and morality*. New York, NY: Psychology Press.

Fiedler, K., Kutzner, F., & Krueger, J. I. (2012). The long way from α-error control to validity proper: Problems with a short-sighted false-positive debate. *Perspectives on Psychological Science*, *7*, 661–669.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1979). Subjective sensitivity analysis. *Organizational Behavior & Human Performance*, *23*, 339–359. doi:10.1016/0030-5073(79)90002-3

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics 3: Speech acts* (pp. 41–58). New York, NY: Academic Press.

Hintzman, D. L., & Curran, T. (1994). Retrieval dynamics of recognition and frequency judgments: Evidence for separate processes of familiarity and recall. *Journal of Memory and Language*, *33*, 1–18. doi:10.1006/jmla.1994.1001

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth-telling. *Psychological Science*, *23*, 524–532. doi:10.1177/0956797611430953

Kahneman, D., & Tversky, A. (1982). The simulation heuristic. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), Judgment under

uncertainty: Heuristics and biases (pp. 201–208). New York, NY: Cambridge University Press.

Lilienfeld, S. O. (2010). Can psychology become a science? *Personality and Individual Differences*, *49*, 281–288.

Moshagen, M., Musch, J., & Erdfelder, E. (2012). A stochastic lie detector. *Behavior Research Methods*, *44*, 222–231. doi:10.3758/s13428-011-0144-2

Murayama, K., Pekrun, R., & Fiedler, K. (2014). Research practices that can prevent an inflation of false-positive rates. *Personality and Social Psychology Review*, *18*, 107–118.

Schultz, P. W., Nolan, J. M., Cialdini, R. B., Goldstein, N. J., & Griskevicius, V. (2014). The constructive, destructive, and reconstructive power of social norms. *Psychological Science*, *18*, 429–434.

Schwarz, N. (1994). Judgment in a social context: Biases, shortcomings, and the logic of conversation. *Advances in Experimental Social Psychology*, *26*, 123–162.

Schwarz, N., & Sudman, S. (1996). *Answering questions: Methodology for determining cognitive and communicative processes in survey research*. San Francisco, CA: Jossey-Bass.

van Maanen, H. (2012). Eén op de tien psychologen vervalst onderzoeksdata. (One of ten psychologists falsifies research data.) *Volkskrant*, February 22, 2012. Retrieved from http://www.volkskrant.nl/wetenschap/een-op-de-tien-psychologen-vervalst-onderzoeksdata~a3195121/

Wald, A. (1947). *Sequential analysis*. Oxford, England: John Wiley.

Willis, G. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks, CA: Sage.

Author Biographies

Klaus Fiedler is a professor of social psychology at the University of Heidelberg, Germany. He is a fellow of various societies, a member of the German national academy of science Leopoldina, and a Leibniz Award winner. He published books and research articles on language and social cognition, judgment and decision making, stereotyping, and the interplay of cognitive and ecological processes. Klaus Fiedler is currently an associate editor of the *Journal of Experimental Psychology: General*.

Norbert Schwarz is a provost professor in the department of psychology and Marshall School of Business at the University of Southern California. He is a member of the American Academy of Arts and Sciences and the German National Academy of Science Leopoldina; other recognitions include the Wilhelm Wundt Medal of the German Psychological Society and the Donald Campbell Award of the Society for Personality and Social Psychology. His research focuses on the context sensitivity of human judgment and decision making and its implications for social science research.