


Underreporting in Psychology Experiments: Evidence From a Study Registry

Social Psychological and
Personality Science
2016, Vol. 7(1) 8–12
© The Author(s) 2015
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1948550615598377
spps.sagepub.com


Annie Franco¹, Neil Malhotra², and Gabor Simonovits¹

Abstract

Many scholars have raised concerns about the credibility of empirical findings in psychology, arguing that the proportion of false positives reported in the published literature dramatically exceeds the rate implied by standard significance levels. A major contributor of false positives is the practice of reporting a subset of the potentially relevant statistical analyses pertaining to a research project. This study is the first to provide direct evidence of selective underreporting in psychology experiments. To overcome the problem that the complete experimental design and full set of measured variables are not accessible for most published research, we identify a population of published psychology experiments from a competitive grant program for which questionnaires and data are made publicly available because of an institutional rule. We find that about 40% of studies fail to fully report all experimental conditions and about 70% of studies do not report all outcome variables included in the questionnaire. Reported effect sizes are about twice as large as unreported effect sizes and are about 3 times more likely to be statistically significant.

Keywords

research methods, multiple comparisons, experiments, disclosure, research transparency

Many scholars have raised concerns about the credibility of empirical findings in psychology, arguing that the proportion of false positives reported in the published literature dramatically exceeds the rate implied by standard significance levels (i.e., 5% or 10%; e.g., Schooler, 2011; Simmons, Nelson, & Simonsohn, 2011). The November 2014 issue of *Perspectives on Psychological Science* was almost entirely devoted to this topic. The accumulation of false positive findings is usually attributed to two main sources. First, studies yielding null results are less likely to be written up, submitted, and published than those with positive findings (Franco, Malhotra, & Simonovits, 2014; Rosenthal, 1979). Second, researchers may only report large and statistically significant effects in published articles. For instance, they may exclude outcomes or experimental conditions that yield null results—either by choice or at the request of reviewers and editors to “streamline” manuscripts.

The practice of reporting or performing only a subset of the statistical analyses in a research project is problematic, especially if researchers have incentives to *selectively* report only results that confirm their hypotheses. If this is the case, then the published literature will be biased because large effect sizes will be overrepresented (Simmons et al., 2011). A major challenge to measuring the incidence of these practices is that the exact set of variables available to researchers for analysis—and hence, the universe of statistical tests they could have

conducted—is generally unknown. While survey evidence suggests that selective reporting of studies, experimental conditions, and/or outcomes are common research practices in psychology (John, Loewenstein, & Prelec, 2012), these estimates are based on self-reports and likely underestimate the prevalence of underreporting. Understanding the size and scope of underreporting of findings is crucial to weighing the costs and benefits of instituting reforms to scholarly practice, such as the requirement of filing preanalysis plans in advance of publication (Cumming, 2013; Miguel et al., 2014). Although the existence of publication bias has been documented in psychology (Cooper, DeNeve, & Charlton, 1997) and the social sciences in general (Franco et al., 2014), direct evidence of underreporting in published research is scarce (see Franco, Malhotra, & Simonovits, 2015, for a study of political science).

In this article, we examine the frequency and nature of underreporting in a set of published psychology experiments. We identify a population of studies for which questionnaires

¹ Department of Political Science, Stanford University, Stanford, CA, USA

² Graduate School of Business, Stanford University, Stanford, CA, USA

Corresponding Author:

Neil Malhotra, Graduate School of Business, Stanford University, 655 Knight Way, Stanford, CA 94305, USA.

Email: neilm@stanford.edu

and data are made publicly available. We can therefore approximate the set of analyses researchers intended to perform based on the experimental conditions and measured outcome variables. We first calculate the frequency of underreporting by counting the number of these design elements that were not mentioned in the corresponding journal articles. Excluding these features from published analyses arguably imposes the greatest restriction on the set of possible statistical tests. Authors, perhaps without awareness, might also choose to focus on comparisons that seem likely to reveal significant differences based on a preliminary examination of the raw data. Although this is not selective reporting in a narrow sense, such practices also distort the published literature. To investigate whether reported effects are more likely to be large and statistically significant, we replicate the analyses from papers that underreport outcomes and compare these results to unreported tests implied by their research designs. Taken together, the evidence suggests underreporting in psychology experiments is both prevalent and selective.

Method

The complete experimental design and full set of measured variables are not accessible for most published research. To overcome this problem our analysis leverages Time-sharing Experiments for the Social Sciences (TESS), a National Science Foundation-sponsored program that fields online experiments embedded in surveys of representative samples of the U.S. adult population. TESS collects data at no cost to researchers but requires them to make questionnaires and data corresponding to their experiments public. In essence, we have identified a registry of studies known to have been conducted, and for which we can recover the full set of experimental conditions and outcomes that researchers intended to analyze *before* each study was conducted. An overview of the program can be found in the Online Appendix.

A key feature of TESS is that researchers face strict caps on the number of respondent questions (i.e., the number of questions multiplied by the number of respondents asked each question). Due to these constraints, it is likely that the costly inclusion of an item in the questionnaire reflects an important theoretical expectation. Scarce resources in TESS questionnaires imply that we can reasonably approximate the full set of tests that the researchers planned to conduct, and therefore assess whether underreporting systematically privileges large and statistically significant effects.

One possible concern about using TESS studies for our analysis is that these experiments are clearly not a random sample of all research conducted in the field of psychology. However, it is unlikely that underreporting in general is *less* severe than what is described here. Many empirical studies appearing in psychology journals are based on analyses of convenience samples that are typically cheaper to collect (e.g., undergraduate subjects and Amazon Mechanical Turk workers). For these less expensive studies, it is easier for researchers to include many experimental conditions and outcome variables in the initial

research protocols and exclude them later on. A related concern is that TESS studies may be unrepresentative, given that authors are aware that their questionnaires and data will eventually be made public. Again, this should produce less underreporting than we might see in typical empirical research where the complete data are not public.

Data Collection

Our initial sample consisted of the entire online archive of TESS studies as of January 1, 2014, or 249 studies conducted between 2002 and 2012. The analysis reported below is restricted to the 32 studies belonging to the field of psychology published as of September 15, 2014. We included studies in our final sample if they appeared in peer-reviewed psychology journals or if any of the principal investigators of the study are affiliated with a psychology department. This means that some articles we include were published in interdisciplinary journals such as the *Proceedings of the National Academy of Sciences* and *Political Psychology*. The list of journals in which these studies appeared is reported in Online Appendix Table A1.

We first compare two aspects of experimental design as planned in the questionnaire and reported in the published papers: experimental conditions and outcome variables. We coded the number of conditions as the number of unique values of the manipulated variable for simple experiments, and the number of manipulated variables (i.e., treatment arms or factors) for factorial designs. We coded survey items as “outcomes” if they were asked after an experimental treatment and could plausibly be affected by it. Finally, we coded these design features as being reported if they were mentioned¹ in the text of the published article.

To investigate whether the reporting of experimental tests is selective, we examine 17 studies for which at least one outcome variable was underreported and at least one experimental condition was analyzed. We compare treatment effects on outcomes that were reported in the published article with treatment effects on outcomes that were not reported. While we cannot be certain that all these analyses were actually performed, we conjecture that the authors intended to perform these tests before observing the data (Gelman & Loken, 2014). If underreporting is selective, effect sizes will be smaller among unreported comparisons, and estimated differences will be less likely to be statistically significant. Details of the data collection, coding, and replication procedures are reported in the Online Appendix.

Results

We present our first set of results in Figure 1. For each design feature, we plot the number of items listed in the questionnaire against the number mentioned in the published paper. Observations below the 45° line are indicative of underreporting. As illustrated in Figure 1a, 41% of papers fail to report all experimental conditions in the published paper (six studies do not report any experimental conditions at all).² On average, the

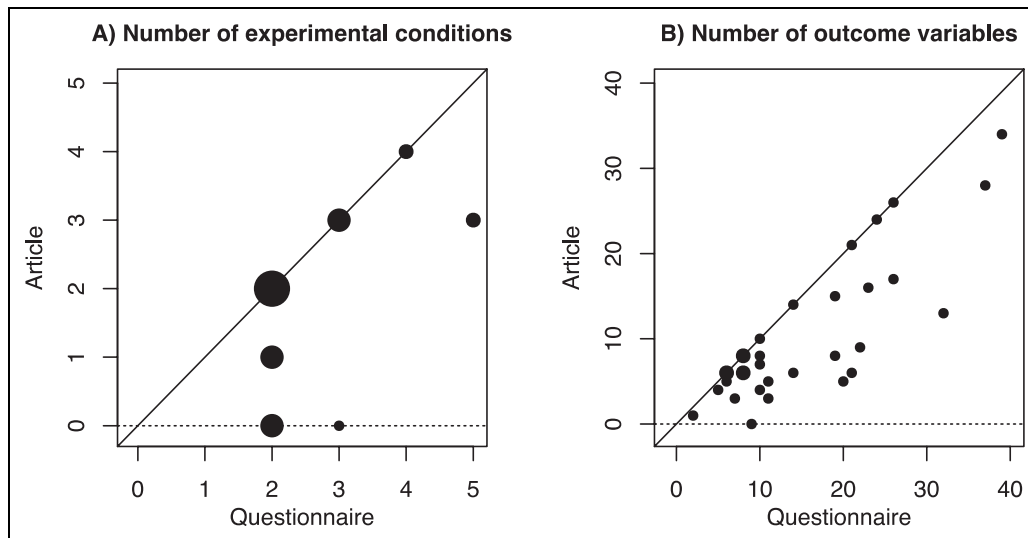


Figure 1. Comparing design features in questionnaires and published results. Note. The sizes of the dots are proportional to the number of studies falling in a particular category. Observations on the diagonal line are those with full reporting.

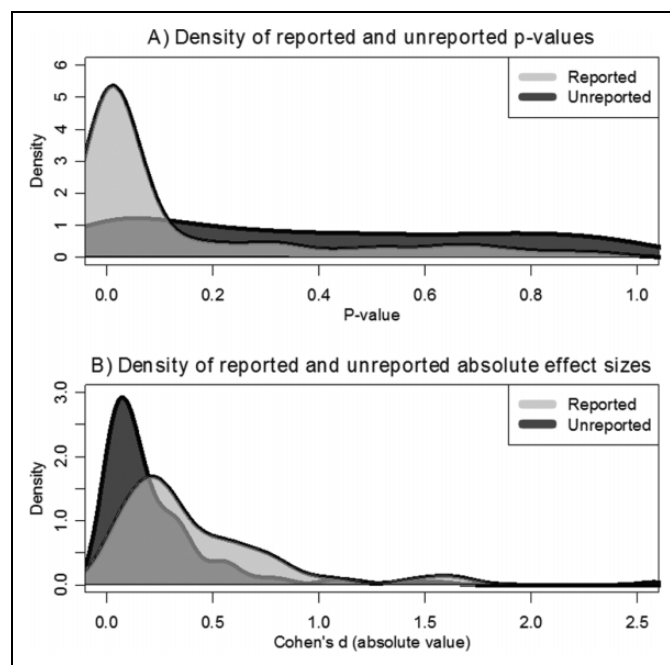


Figure 2. Comparing reported and unreported tests.

questionnaires mentioned 2.5 experimental conditions, whereas the papers only mentioned 1.8 conditions. Figure 1b shows that 72% of papers report fewer outcome variables than those listed in the questionnaire. On average, the questionnaires included 15.4 outcome variables, whereas the articles only reported on 10.4 outcomes. Examining the joint distribution of underreporting across both design features reveals that only a quarter of studies (8 of the 32) reported all experimental conditions and outcome variables.

We also find evidence that the reporting of outcome variables is selective. Figure 2 plots densities of the p values

Table 1. Underreporting Is Selective.

	Median p Value	Median Cohen's d	% Significant at 95%	% Significant at 90%
Unreported tests ($n = 147$)	.35	.13	23	28
Reported tests ($n = 122$)	.02	.29	63	70
Difference	-.33	.16	40	42
Difference (with study fixed effects)	-.24	.14	31	37

Note. The last row reports point estimates from median regressions for the comparisons of medians, and ordinary least squares regression for the comparison of means. All regressions include study fixed effects.

and effect sizes associated with reported and unreported tests. p values associated with reported treatment effects are typically below the .05 threshold, while those from the unreported tests are not (Panel A). Reported effect sizes (the absolute values of Cohen's d 's) are much larger than unreported ones (Panel B). As shown in Table 1, the median reported p value is .02, whereas the median unreported p value is .35. Also, roughly two thirds of the reported tests are significant at the 5% level compared to about one quarter of the unreported tests. We find similar patterns with regard to effect sizes: Reported experimental differences are roughly twice as large as unreported ones. The last row of the table shows that these patterns remain unchanged when we adjust for heterogeneity across studies. Regression estimates from models including study fixed effects—which capture how researchers differentially report tests *within* a given study—reveal similar differences between reported and unreported tests.

Discussion

We have provided evidence that published papers diverge substantially from research protocols, with extensive underreporting of outcome variables and experimental manipulations. Underreporting is selective: Published effect sizes are larger and more likely to be statistically significant compared to the full set of possible comparisons implied by the experimental designs.

Scholars may be motivated to underreport outcomes, conditions, or alternative specifications that do not confirm their hypotheses for several reasons. Blame is often attributed to professional incentive structures that discourage publication of null findings. However, personal motivations and cognitive biases might also play a role. Researchers may not always be consciously aware of their decisions during data analysis (Gelman & Loken, 2014). Regardless of which mechanism underlies the patterns of underreporting we document, the consequences of selective reporting for publication bias remain the same—published effect sizes and the probability of Type I errors will be biased upward.

The evidence presented here suggests preanalysis plan requirements may be the most effective mechanism for reducing selective reporting. Some might argue that the goal of curtailing questionable research practices is achieved at the expense of exploratory research. However, preanalysis plans do not preclude investigators from departing from their planned analyses. They only make clear which analyses were preplanned and which ones are post hoc. A less onerous, but not mutually exclusive, remedy is disclosure (Sagarin, Ambler, & Lee, 2014). If researchers were simply required to report all measures and conditions in a transparent fashion, then readers could make their own determinations about the quality of the evidence.

Psychologists have recently argued that changes to the publication process will also be necessary to shift disciplinary norms (Nosek, Spies, & Motyl, 2012; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012). Indeed, journals are adopting stricter reporting requirements (Eich, 2014; Nosek & Lakens, 2014). However, one potential concern is that reviewers will judge research that reports inconsistent findings harshly unless given explicit instructions to the contrary (Maner, 2014). If this is the case, authors required to report all outcomes might incur a penalty if their data reveal null or mixed support for their hypotheses. But the widespread adoption of preanalysis plan requirements might address this problem. As scholars become more familiar with extended reporting requirements, they may be less willing to evaluate the papers they review on the basis of statistical significance alone.

Authors' Note

All replication data and code are posted on the corresponding author's Dataverse page.

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) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. Note that experimental conditions and outcome variables are sometimes mentioned but not analyzed in published papers. Thus, underreporting is even more severe if we only consider a design element as being reported if corresponding results are presented.
2. Four of these studies report correlational analyses and do not mention the experimental manipulations. One study included a randomization of question order that was not reported. One study employed a factorial design with two factors; one factor was unreported and one was partially reported.

Supplemental Material

The online appendices are available at <http://spps.sagepub.com/supplemental>.

References

- Cooper, H., DeNeve, K., & Charlton, K. (1997). Finding the missing science: The fate of studies submitted for review by a human subjects committee. *Psychological Methods*, 2, 447.
- Cumming, G. (2013). The new statistics: Why and how. *Psychological Science*, 25, 7–29.
- Eich, E. (2014). Business not as usual. *Psychological Science*, 25, 3–6.
- Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, 345, 1502–1505.
- Franco, A., Malhotra, N., & Simonovits, G. (2015). Underreporting in political science survey experiments: Comparing questionnaires to published results. *Political Analysis*, 23, 306–312.
- Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, 102, 460.
- 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.
- Maner, J. K. (2014). Let's put our money where our mouth is: If authors are to change their ways, reviewers (and editors) must change with them. *Perspectives in Psychological Science*, 9, 343–351.
- Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., . . . Van der Laan, M. (2014). Promoting transparency in social science research. *Science*, 343, 30–31.
- Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology*, 45, 137–141.
- Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, 7, 615–631.
- Sagarin, B. J., Ambler, J. K., & Lee, E. M. (2014). An ethical approach to peeking at data. *Perspectives on Psychological Science*, 9, 293–304.
- 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.
- Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86, 638.
- Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, 7, 632–638.

Author Biographies

Annie Franco is a PhD student in the Department of Political Science at Stanford University. She is the coauthor of “Publication Bias in the

Social Sciences: Unlocking the File Drawer,” published in *Science* in 2014.

Neil Malhotra is a professor of political economy in the Graduate School of Business at Stanford University. He is the coauthor of “Publication Bias in the Social Sciences: Unlocking the File Drawer,” published in *Science* in 2014. He is affiliated with the Berkeley Institute for Transparency in the Social Sciences (BITSS) and the Meta-Research Innovation Center at Stanford (METRICS) and has worked on projects with the Center for Open Science (COS).

Gabor Simonovits is a PhD student in the Department of Political Science at Stanford University. He is the coauthor of “Publication Bias in the Social Sciences: Unlocking the File Drawer,” published in *Science* in 2014.