Providing Incentives for Sound Scientific Practice: The Case for Sealed-Envelope-Submissions[♣]

Martin Dufwenberg[♦] & Peter Martinsson[♥]

April 13, 2017

Abstract: Because journals favor clear stories researchers' may gain by engaging in scientific misconduct, ranging from shady practices like running more sessions hoping for significance to outright data fabrication. To set researchers' incentives straight, we propose sealed-envelope submissions, where editors' and referees' evaluations are based only on the interest of the research question and on the proposed empirical method.

1. Problem

Many worry about questionable scientific practices that bias reported results. There is a spectrum of possibilities, from shady practices like running more sessions hoping for significance to outright data fabrication. By a recent estimate "two-thirds of retracted life-science papers were stricken from the scientific record because of misconduct" (Corbyn 2013, p. 21; cf. Fang, Steen & Casadevall 2013). Couzin-Frankel (2013, p. 68) quotes an anonymous researcher: "We did this experiment a dozen times, got this answer once, and that's the one we decided to publish." Tip of an iceberg? Anecdotes? It is not in a researcher's interest to disclose a shady practice, making it hard to find direct evidence on how widespread scientific misconduct is and on how misleading published results may be.[1]

[♣] We have benefited from comments by Chris Chambers, Georg Kirchsteiger, Zach Maniadis, Nikos Nikiforakis, and Larry Samuelson, as well as participants at a seminar organized by Bocconi University's Experimental Laboratory for the Social Sciences (BELSS).

[♦] University of Arizona, University of Gothenburg, CESifo; martind@eller.arizona.edu

[♥] University of Gothenburg, Linköping University; peter.martinsson@economics.gu.se

[1] Some scholars tried. See List, Bailey, Euzent & Martin's (2001) survey of unethical behavior, using randomized response techniques that encourage honest responses despite the sensitive topic. Brodeur, Lé, Sangnier & Zylberberg (2013, p. 1) report that, 2005-11, three top economics journals (*AER*, *JPE*, *QJE*) published empirical findings with p-values that exhibit "a valley between 0.25 and 0.10 and a bump slightly below 0.05" seen to indicate that many researchers "*inflate* the value of … almost-rejected tests

It is easier to judge the problem by reflecting on the incentives involved. Arguably there is great cause for concern. Suppose journals wish to "cast results as a story that they believe others will want to read" (Couzin-Frankel 2013, p. 68). In response (by backward induction), given the large rewards (grants, tenure, careers!) for publishing well, researchers may gain by tweaking findings (cf. Fanelli & Ioannidis 2013; Lacetera & Zirulia 2011).

Proposals to rectify the problem appeared, though efficacy is doubtful. Whistle-blowing by peers involves "significant risks, and the path is rarely simple" (Young, Ledford & Van Noorden 2013, p. 454). Having senior mentors teach integrity may be useful (Neaves 2012), but the possibility of aligned incentives between junior and senior scholars suggest that relying on such honesty may be wishful thinking. Study registration and pre-analysis plans – besides being burdensome to formulate – do not solve the following issue: if certain results are more publishable then others researchers will still have incentives to fabricate such results while flagging for them beforehand (cf. Humphreys, Sanchez de la Sierra & van der Windt 2013).

We propose a different approach, where the empirical results are submitted in a sealed envelope to the journals.

2. Solution

The problem may be overcome by a drastic change in how articles are submitted and evaluated for publication at journals. We call it a sealed-envelope submissions proposal:[2]

> *Journals should insist that submitted articles do not reveal any empirical results. All the data, along with the statistical analyses, should be submitted in a sealed envelope. The editors and referees should evaluate the submission based only on the interest of the chosen research question and on the relevance of the chosen empirical*

---

by choosing a 'significant' specification." John, Lowenstein & Prelec (2012) report results from a large survey among psychologists on questionable research practices. Their findings indicate that the main activities undertaken are related to selectivity use of data collected and decisions related to collection of data (either to collect more data or stop ongoing data collection). Nosek et. al. (2015) replicated 100 studies published in three psychological journals and they only duplicated the same results in 39% of them. Camerer et al. (2016) report the results from a replication of 18 experimental studies published in two top journals in economics (*American Economic Review* and *Quarterly Journal of Economics*) in 2011-2014. Depending on measure chosen to evaluate replication success, it is found to be in the range of 61% to 78%.

[2] In practical terms, the paper is submitted in two parts in the editorial system, where the result part ("the sealed envelope") is locked until the editor has made her final decision.

method. After making their accept-reject decision, the editors may then open the envelope.

Our diagnosis of the problem was based on a backward induction argument, and so is now our solution. We trace the roots of scientific misconduct to the conditioning of editorial decisions on the nature of data. If one makes editorial decisions blind to the nature of researchers' data then the incentives to engage in questionable research practices may go away.

3. Galileo

Related proposals were floated in the past, but concerned avoiding publication bias or project selection rather than eliminating incentives for misconduct (Sterling 1959, Rosenthal 1966, Walster & Cleary 1970, Feige 1975, Dufwenberg 2014). If results are published only if they tell a clear story (*e.g.,* through statistically significant effects), outlier data get over-represented in published work.[3] These proposals seem to have been largely forgotten or neglected, probably because one can brush off the problem and say that, as long as one is aware of the bias one can adjust one's outlook accordingly. Published data is still real data.

It is much harder to brush off scientific misconduct with an analogous argument. If data are made up, if chosen estimation methods are conditioned on significance, or if reporting is done with spin, how can one tell what's real from what is make-believe? Fake data are *not* real data. Depending on the degree of misconduct, conclusions may vary from dubious to useless. We believe the problem is serious because researchers' incentives are so strong. Furthermore, the risks involved may be rather small. "There is no cost to getting things wrong; the cost is not getting them published," as psychologist Brian Nosek put it when consulted for a recent article on the topic (*The Economist*, 2013). With our proposal, editorial decisions become independent of the nature of the data, so no researcher can gain or lose, in terms of publishability, depending on the nature of the data.

Researchers have reacted to incentives since Galileo, by many considered as the father of science, denounced heliocentricism. While it is easy to sympathize with his decision, modern-day incentives encourage less laudable researcher conduct. The sealed-envelope submission proposal holds promise to set those incentives straight!

[3] Bias either enters directly through editors' decisions, or because researchers do not bother to write up null findings (cf. Franco, Malhotra & Simonovits 2014).

Postscript

After we finished the first version of the above text, Chris Chambers alerted us to the "Registered Reports" (RR) initiative recently started by the journal *Cortex* (Chambers 2013). Under this scheme projects are submitted for review, then accepted or rejected for subsequent publication *before the data has been collected.* It works like our sealed-envelope submission proposal, except there is no envelope. Different versions of RR have subsequently been adopted by an increasing number of journals; currently there are 49 (April 10, 2017). In some cases, starting with *Journal of Business and Psychology* and its editor Steven Rogelberg, a form referred to as "hybrid RR" is used, which is, in fact, our sealed-envelope submission proposal. For more information, including a list of journals that explore related ideas, check out the following URL (hosted by the *Center for Open Science*, founded by Brian Nosek and Jeff Spies):

https://cos.io/rr/

Pondering pros & cons of "full" vs. hybrid RR (=our proposal) is intriguing. Chris Chambers suggested to us that hybrid RR does not prevent hypothesizing after results are known (aka "harking") and that authors may still be motivated to "p-hack" (i.e. fiddle with data to achieve a desired significance level) if they believe that this will help their paper attract more citations. He also conjectured that researchers may use hybrid RR as a vehicle mainly to publish negative or unclear findings, and that editors would probably suspect (at least early on) that all such submissions fall into one of those categories. Against all that, a benefit of hybrid RR may be practical as the refereeing task can be completed right away, while with full RR one has to wait for the researchers to actually go and collect and analyze the data according to RR.

Maybe time will tell what is best. With the exciting RR initiative underway there is hope, although the evidence is still too too limited (across time, across journals, and as regards the extent to which it is applied within journals as, most often, RR-submissions are optional or restricted to special issues) to draw clear conclusions. Relatively few researchers (across all of science) seem to be aware of the movement. We hope the message of our paper is worth repeating and debating.

References

Brodeur, A., M. Lé, M. Sangnier, & Y. Zylberberg (2016), "Star Wars: The Empirics Strike Back," *American Economic Journal: Applied Economics* 8, 1-32.

Camerer, C. F., A. Dreber, E. Forsell, T.-H. Ho, J. Huber, M. Johannesson, M. Kirchler, J. Almenberg, A. Altmejd, T. Chan, E. Heikensten, F. Holzmeister, T. Imai, S. Isaksson, G. Nave, T. Pfeiffer, M. Razen, & H. Wu (2016), "Evaluating Replicability of Laboratory Experiments in Economics," *Science* 351, 1433-1436.

Chambers, C. (2013), Editorial: "Registered Reports: A New Publishing Initiative at *Cortex*," *Cortex* 49, 609-610.

Corbyn, Z (2013), "Misconduct is the Main Cause of Life-Sciences Retractions," *Nature* 490, 21.

Couzin-Frankel, J. (2013), "The Power of Negative Thinking," *Science* 342, 68-69.

Dufwenberg, M. (2014), "Maxims for Experimenters", forthcoming in *Methods of Modern Experimental Economics*, Frechette, G. & A. Schotter (Eds.), Oxford University Press.

Fanelli, D., & J.P.A. Ioannidis (2013), "US studies may overestimate effect sizes in softer research," *PNAS* 110, 15031-15036.

Fang, F.C., R.G. Steen, & A. Casadevall (2013), "Misconduct Accounts for the Majority of Retracted Scientific Publications," *Proceedings of the National Academy of Sciences* 109, 17028-17033.

Feige, E. (1975), "The Consequences of Journal Editorial Policies and a Suggestion for Revision," *Journal of Political Economy* 83, 1291-1296.

Franco, A., N. Malhotra & G. Simonovits (2014), "Publication Bias in the Social Sciences: Unlocking the File Drawer, *Science* 345, 1502-1505.

Humphreys, M. R. Sanchez de la Sierra & P. van der Windt (2013), "Fishing, Commitment, and Communication: A Proposal for Comprehensive Nonbinding Research Registration," *Political Analysis* 21, 1-20.

John, L.K., G. Lowenstein & D. Prelec (2012), "Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling," *Psychological Review 23,* 524-532.

Lacetera, N. & L. Zirulia (2011), "The Economics of Scientific Misconduct," *Journal of Law, Economics and Organization* 27, 568-603, 2011.

List, J., C. Bailey, P. Euzent & T. Martin (2001), "Academic Economists Behaving Badly? A Survey on Three Areas of Unethical Behavior," *Economic Inquiry* 39, 162-170.

Neaves, W. (2012), "The Roots of Research Misconduct," *Nature* 488, 121-122.

Nosek, B.A., G. Alter, G.C. Banks, D. Borsboom, S.D. Bowman, S.J. Breckler, S. Buck, C.D. Chambers, G. Chin, G. Christensen, M. Contestabile, A. Dafoe, E. Eich, J. Freese, R. Glennerster, D. Goroff, D. P. Green, B. Hesse, M. Humphreys, J. Ishiyama, D. Karlan, A. Kraut, A. Lupia, P. Mabry, T. Madon, N. Malhotra, E. Mayo-Wilson, M. McNutt, E. Miguel, E. Levy Paluck, U. Simonsohn, C. Soderberg, B.A. Spellman, J. Turitto, G. VandenBos, S. Vazire, E.J. Wagenmakers, R. Wilson & T. Yarkoni (2015), "Promoting an Open Research Culture," *Science* 348, 1422-1425.

Rosenthal, R. (1966), *Experimenter Effects in Behavioral Research*, New York: Appleton-Century-Croft.

Sterling, T. (1959), "Publication Decision and Possible Effects on Inferences Drawn from Tests of Significance – Or Vice Versa," *Journal of the American Statistical Association* 54, 30-34.

The Economist (2013), "Trouble at the Lab," Issue 19th-25th October.

Walster, G. & T. Cleary (1970), "A Proposal for a New Editorial Policy in the Social Sciences," *American Statistician* 24, 16-19.

Young, E., H. Ledford & R. Van Noorden (2013), "3 Ways to Blow the Whistle," *Nature* 503, 454-457.