Sealed Envelope Submissions Eliminate Bias in Science[♣]

Martin Dufwenberg[♥] & Peter Martinsson[♠]

February 24, 2019

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. We argue that researchers who are inherently honest and who would not have cheated anyway will not be hurt by our protocol, but rather be helped by being protected.

Keywords: scientific misconduct; editorial policy; incentives; sealed-envelope-submissions; backward induction; registered reports; ethics; honesty

JEL codes: A19, B49

[♣] The first version of this paper, then titled "Keeping Researchers Honest: The Case for Sealed-Envelope-Submissions," was completed in the fall of 2013. The manuscript got its current form in connection with a workshop on *Emotions & Ethics: Insights from Behavioral Economics*, organized by the *Revue Economique* in honor of Marie Claire Villeval (Paris, February 1, 2019); we thank participants for their comments. We have previously benefited from comments by Douglas Bernheim, Chris Chambers, Georg Kirchsteiger, John List, Zach Maniadis, Larry Samuelson, Joel Sobel, Marcella Veronesi, and participants at two events organized by Bocconi University: *Science Communication & Information* (March 6, 2015) and *Incentivizing Experimental Research* (a BELSS event, December 1, 2014).

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

[♠] University of Gothenburg; peter.martinsson@economics.gu.se

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.

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, and 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 could be useful tools to thwart "harking," i.e., hypothesizing after results are known. However, besides being burdensome to formulate – pre-registration does not solve the issue that if certain results are more publishable than others researchers will still have incentives to fabricate such results while flagging for them beforehand, or that pre-registration will take place after data has been collected (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:

> *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 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.

We would like to make a few further comments regarding this solution:

First, 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.

Second, the proposal has a lot in common with writing grant proposals. It is essentially already in place for researchers who need to find funds to conduct their research. We hence propose to extend those commonly accepted principles to editorial policy.

Third, one may legitimately worry that people may try to game the sealed envelope system such that they follow up their very careful description of research questions and design with a sloppy data analysis, since that analysis would appear only inside the envelope and so seem un-incentivized. We propose, therefore, that any acceptance decision is still conditional on a check that the data analysis is of sufficient quality, and we would assume that editors are able to enforce such a standard in an unbiased way.

3. Honesty

If all researchers were perfectly honest, intrinsically motivated to do good science rather than to score a good publication no matter how, the problem we have described would go away. The sealed envelope proposal would be redundant. There is indeed plenty of experimental evidence indicating that many people do not like to lie, or to break promises or informal

agreements, or to cheat when reporting data.[1] Nevertheless, these experiments typically document that some subjects behave badly, or at least that some cheat to some degree.

It is hard to quantify how prevalent cheating is in science, since undetected fraud (tautologically) cannot be observed. Yet, empirical researchers tried, using various indirect methods. Their findings are consistent with the experimental ones, i.e., some scientists do seem to cheat, some more than others. For example, in a recently published paper in *PNAS*, Fanelli, Costas & Ioannidis (2017) conduct a systematic review to assess bias in research by performing meta-analyses. On the positive side they report on average small biases. In their conclusions, however, they write: "A link between pressures to publish and questionable research practices cannot be excluded, but is likely to be modulated by characteristics of study and authors, including the complexity of methodologies, the career stage of individuals, and the size and distance of collaborations […]. The latter two factors, currently overlooked by research integrity experts, might actually be growing in importance, at least in the social sciences" (p. 3718).[2]

It is important to understand the incentive structure for researchers in order to evaluate their behavior and to suggest countermeasures. A model explaining an individual's behavior may contain three key elements: (i) extrinsic motivation, (ii) intrinsic motivation and (iii) self-image, where most of us would attach some weight to all three elements.[3] Perhaps the findings reported

[1] See e.g. Gneezy (2005), Charness & Dufwenberg (2006), Dufwenberg, Servatka & Vadovic (2017), Fischbacher & Follmi-Heusi (2013), Garbarino, Slonim & Villeval (2016). See also Olken (2015) who discusses how many researchers are likely to be honest, or at least not "nefarious."

[2] There is also some conclusion-wise not dissimilar previous work by economists. List, Bailey, Euzent & Martin's (2001) conduct a 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 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 only duplicated the results in 39% of them. Camerer et al. (2016) report 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%.

[3] Compare with e.g. Bénabou & Tirole (2006). One example of a purely intrinsically motivated researcher may be the Russian mathematician Grigori Perelman who has turned down many prestigious awards and job offers.

in Fanelli *et al.* (2017) may be linked to how the elements (i), (ii), and (iii) apply differentially to different researchers, in different situations, and at different stages in their careers. For example, extrinsic motivation may be relatively more important for (pre-tenure) early career researchers and the seal envelope proposal might have more influence in such an environment where a publication seems more likely if results are significant.

Note that a sealed envelope approach could set the incentives correct. The point we wish to make here is that even if a significant number of researchers are (to some degree) honest they will not be hurt by this protocol, but rather be protected. The incentive to approach important and relevant research questions, without focusing on statistical significance, is the same whether researchers are extrinsically or intrinsically motivated. Thus, dishonest researchers will therefore gain no advantage over honest ones.

4. Papers & results

We propose that publishability should not depend on data. We do not suggest that all results are equally interesting. Obviously, they are not! The results of Nobel Laureate Barry Marshall (who swallowed H pylori to test his hypothesis that gastric ulcer had bacterial cause) would hardly have gotten him his Prize if they had-not been positive.

A downside of our proposal may thus be to fill journals with long boring articles with null findings. But, imagine having a section of the journal called "Papers" and another called "Results," the two equally meritorious as regards outlet prominence.[4] The latter articles are note-length summaries of methods and results that refer to online appendices for details. Now suppose the review process involves two stages. The first is the one we described above, a result-free assessment that determines, based on methods alone, whether the results would be worth reporting and publishing. In the second stage the editor opens the envelope and makes an assessment of the results. The editor then decides whether the paper is accepted for the "Papers" or "Results" section based on the interest level in the results. That way results don't determine whether the paper gets published, but they do determine how much space it gets. And the editor cannot complain about methods when making the space decision – the justification has to be based on intrinsic interest of the results, which is much less controversial.

[4] We thank Douglas Bernheim for this excellent suggestion.

5. Galileo

Related proposals, involving sealed envelope submission, were discussed in the past. However, they concerned avoiding publication bias or project selection rather than eliminating incentives for misconduct (see 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.[5] 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? Faked 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!

Postscript

After we finished the first version of this paper, in the fall of 2013, Chris Chambers alerted us to the "Registered Reports" (RR) initiative started by the journal *Cortex* (Chambers 2013).

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

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 164 (February 23, 2019).[6] In some cases, starting with *Journal of Business & 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 necessarily preclude harking or "p-hacking" (i.e., fiddle with data to achieve a desired significance level) if researchers believe this will help attract 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.

Time may tell what is best. With the exciting RR initiative underway there is hope, although the evidence is still too limited (across time and journals, and as regards extent to which it is applied 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

Bénabou, R & J. Tirole (2006), "Intrinsic and Extrinsic Motivation," *Review of Economic Studies* 70, 489–520.

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,

[6] Economics is represented (only) by *Journal of Development Economics*, which launched a RR initiative in 2018.

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.

Charness, G. & M. Dufwenberg (2006), "Promises & Partnership," *Econometrica* 74, 1579-1601.

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, Martin (2015), "Maxims for Experimenters", in *Handbook of Experimental Economic Methodology* (II.7), G. Fréchette & A. Schotter (eds.), ch. 7, Oxford University Press, pp. 141-44.

Dufwenberg, M., M. Servátka & R. Vadovič (2017), "Honesty & Informal Agreements," *Games & Economic Behavior* 102, 269-85.

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

Fanelli, D, Costas, R, Ioannidis JPA (2017), "Meta-Assessment of Bias in Science," *Proceedings of the National Academy of Science* 114, 3714-3719.

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.

Fischbacher, U. & F. Föllmi-Heusi (2013), "Lies in Disguise – An Experimental Study on Cheating," *Journal of the European Economic Association* 11, 525-547.

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

Garbarino, E., R. Slonim & M. C. Villeval (2016), "Loss Aversion and Lying Behavior: Theory, Estimation and Empirical Evidence, unpublished manuscript.

Gächter, S. & J. Schulz (2016), "Intrinsic Honesty and the Prevalence of Rule Violations Across Societies, *Nature* 531, 496-499.

Gneezy, U. (2005), "Deception: The Role of Consequences," *American Economic Review* 95, 384-394.

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.

Olken, B. (2015), "Promises and Perils of Pre-Analysis Plans," *Journal of Economic Perspectives* 29, 61-80.

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.