UNIVERSITY OF Southampton
School of Social Sciences

Economics Division
University of Southampton
Southampton SO17 1BJ, UK

Discussion Papers in
Economics and Econometrics

Title: Evaluating Solutions to the Problem of False Positives

Authors : Thomas Gall (University of Southampton) & Zacharias Maniadis (University of Southampton)

No. 1504

This paper is available on our website
http://www.southampton.ac.uk/socsci/economics/research/papers

ISSN 0966-4246

Abstract

The 'credibility crisis' in science has led to various suggestions for reform by researchers in several scientific disciplines. We use game theory and general equilibrium arguments to evaluate recent proposals to increase transparency by strengthening disclosure requirements when an empirical study is submitted for publication. We find that a policy that prohibitively increases the cost of less severe questionable research practices, such as selective reporting, tends to decrease the overall rate of researcher misconduct, because the rate of 'felonies', such as fabrication, also tends to decrease. Accordingly, proposals that aim to prevent lying by omission (Simmons et al., 2011, Landis et al., 2012, Fanelli, 2013) are likely to be effective in reducing researcher misconduct. Blunt measures such as government audits, which can be used to counteract pure fraud, do not seem equally effective.

Keywords: Researcher Misconduct, Fabrication, False-Positives
JEL Codes: B41, D02, C90

Evaluating Solutions to the Problem of False Positives

Thomas Gall and Zacharias Maniadis*

June 6, 2015

1 Introduction

Is the model of self-correcting science and cumulative scientific growth in accordance with the contemporary world of research? Serious concerns have been voiced recently (Lehrer, 2010, Ioannidis, 2012) that cast doubt on this view, suggesting that it is overly optimistic. In fact, it has been argued that there is a 'confidence crisis' in several scientific fields, such as psychology,[1] management (Bettis, 2012), and several branches of the biomedical sciences (Ioannidis, 2005, Jennions and Møller, 2002). A natural way to measure the extent of the problem is the Rate of False Positives (RFP). It is defined as the fraction of newly discovered findings – regarding associations between phenomena – that correspond to associations that are false in reality (Ioannidis, 2005).[2] That is, a field can be thought to be in a crisis of confidence when the scientific community believes that the RFP is unacceptably large, and it is incumbent upon scientific researchers to find a way to decrease it.

*Department of Economics, School of Social Sciences, University of Southampton, UK.

[1]The journal *Perspectives on Psychological Science* had a special issue on the problem in 2012. Since then, the same journal has introduced a special 'Registered Replication Reports' article type, and a 'Many Labs' replication project has been established (Klein et al., 2014) in order to examine the degree of replicability in the discipline.

[2]In other words, given a new empirical finding, the RFP in a discipline captures the likelihood that this result does not correspond to the truth.

The study of the processes that operate within the system of scientific knowledge accumulation is a scientific discipline of its own, and there is a clear need to study possible solutions to ameliorate the problem of overly high RFP. Despite the fact that the problem has attracted major attention, there is a dearth of rigorous evidence, needed to adequately evaluate the numerous proposals for reform. Ioannidis (2012) emphasizes: "... it is essential that we obtain as much rigorous evidence as possible, including experimental studies, on how these practices perform in real life and whether they match their theoretical benefits [...]. Otherwise, we run the risk that we may end up with worse scientific credibility than in the current system."

In this study we shall make an attempt to contribute both to the understanding of the emergence of Questionable Research Practices (QRP) that tend to distort scientific knowledge, and to the evaluation of proposals for tackling particular forms of these practices. Proposals that target practices related to incomplete revelation of relevant information have been made for science in general (Fanelli, 2013), but also for specific fields such as medicine (Landis et al., 2012) and psychology (Simmons et al., 2011).[3] The remedies suggested typically take the form of guidelines that improve transparency in reporting all relevant information for evaluating research results. However, an important problem is that direct formal evidence in favor of the specific proposals is lacking, but for good reasons. The empirical study of QRP is difficult, because it is a sensitive issue plagued by measurement problems and other methodological complications (Fanelli, 2009). In light of all these problems, the sceptical scientific audience often needs to become convinced based on verbal reasoning alone.

In what follows we shall employ formal (game) theory with an aim to tackle this formidable problem. A rigorous game-theoretic analysis of the proposals is largely absent from the debate, though we believe it has a lot to offer if one is interested in the analysis of scientists' behavior in a highly structured environment with high stakes. Formal tools employed in economics and political science can contribute constructively to assessing the efficacy of appropriate institutional policy changes in lowering the rate of

[3]These authors show in a prominent study that under current scientific practices the rate of false positives may rise to unacceptably high levels.

false positives, because these disciplines focus on the effects of policy changes on incentives and thus behavior. The proposed solutions critically rest on changing researchers' incentives, as Nosek et al. (2012) emphasize: "... the solution requires making incentives for getting it right competitive with the incentives for getting it published".

It should be noted that by rigorously examining the interaction among agents who have idiosyncratic incentives within an institutional system, formal theory often reaches useful counter-intuitive conclusions. This is particularly important for the issue of QRP, as an attempt to eradicate a given form of QRP will generally affect incentives to pursue other forms of misconduct.[4] That is, simple policies may have unintended side effects. Game theoretic modeling provides a means to examine effects of policies in complex patterns of behavioral interactions. Indeed, the overall effect of under-theorized factors, such as the trade-offs that researchers face and the general interdependencies among individuals in the publication system, may be the opposite of what is desired by the proponents of a given proposal.

Our approach will assume that people behave rationally and consider the strategic situation and the incentives that they face. Individual behavior is modeled in terms of a game between players with different incentives. This is a common approach in economics and political science, and there is also increasing acceptance in the psychological literature that researchers respond rationally to institutionalized incentives (Nosek et al., 2012). In our game-theoretic model fraud increases the originality of one's research and thus the likelihood of publication. The reward of fraud depends, however, on the scientific competition, that is, on the extent of fraud committed by one's peers. Hence, if QRP are widespread, engaging in such practices at least to some extent may be necessary to maintain chances of successful publication, as a form of self-defense. Hence, in a sense, a best response to widespread

[4]In many disciplines there is a tradition to consider extreme forms of misbehaviour as the domain of particularly distorted personalities. Scholars from such disciplines might find it hard to believe that extreme behaviour is responsive to incentives. Yet, recent experience has shown that when the incentives are high enough, condemnable practices such as direct falsification of evidence may occur, even at the highest academic level. For instance, the journal *Science* has very recently retracted a political science article where this type of misconduct is believed to have occurred.

fraud may be to engage in even more fraud.

Our game-theoretic model will show that this is indeed the case under plausible assumptions, and as a consequence policies that aim at tackling 'mild' forms of QRP (defined as practices, such as selective reporting, which can be reasonably self-justified – unlike fabrication) may also remove the desire or necessity to resort to even stronger forms of QRP such as pure fraud and fabrication. To our knowledge the possibility of such a mechanism has not yet been part of the scientific discussion.

2 Describing the Research Environment

In the remainder of this paper, QRP will refer to any malpractices employed by researchers that tend to distort the scientific evidence.[5] A necessary first step in conceptualizing and modeling the problem is to capture the essential properties of these QRPs. In our view, the relevant QRPs can be categorized in terms of the degree to which they are acceptable practices, and thus can be self-justified. John et al. (2012) offer a comprehensive list of such practices, drawing from data in the field of psychology. At the top of this 'pyramid of misdeeds' (in terms of the difficulty of ex post justification) one will find data fabrication, which is universally considered unacceptable. Slightly below rank data alteration and falsification, which would be unacceptable in almost all circumstances. Practices such as rounding off p-values will follow, and so on. At the pyramid's bottom one will find practices such as collecting more data after one has obtained a non-significant result, a practice that more than 50% of psychologists admit they have engaged in, as John et al. (2012) show.

The evidence indicates that practices that are not considered uniformly morally condemnable by the scientific community might be tolerated and thus more prevalent.[6] John et al. (2012) and Meyer and McMahon (2004) provide

[5]This does not include ethical problems such as plagiarism and phantom authorship, which do not distort the published evidence, at least in the short run.

[6]Nosek et al. (2012) clearly state this: "At the extreme, we could lie: make up findings or deliberately alter results. However, detection of such behavior destroys the scientist's reputation. This is a strong incentive against it, and - regardless of incentives - most resist

survey evidence that less ethically defensible behaviors are self-admitted and observed (as committed by other scientists) less than more defensible behaviors, in a roughly monotonic fashion. Fabrication/falsification is typically self-admitted to by about 2 percent of respondents (Fanelli, 2009), while other types of QRP (such as selective reporting) are admitted by about half of respondents (John et al., 2012).

A second important aspect of the relevant environment we wish to model is its *tournament* character. In particular, the space for prestigious publications is limited and all researchers have incentives to publish as high as possible. This reflects a more fundamental characteristic of the scientific profession: the scientific community's capacity of attention is limited, with a myriad of research agendas and results competing for attention. Since QRP tend to increase the originality of one's results and thus the likely attention that peers will devote to them, the payoffs of engaging in any form of QRP will depend not only on the research quality of fellow scientists, but also on their choice of (mis-)conduct. This tournament character can lead to a rat race, where scientists use QRP partly as a self-defense against their peers' use of QRP, and science and society as a whole lose.[7]

To model the different forms of QRP with different degrees of defensibility we shall categorize the possible QRPs in only two tiers, *severe* QRP and *mild* QRP. A key virtue of a theoretical model is simplicity, and this simple assumption will suffice to enable a sufficiently rich level of analysis. To reflect the tournament character of the scientific game of publishing, we shall assume

such behavior because it is easy to identify as wrong (Fanelli, 2009). We have enough faith in our values to believe that we would rather fail than fake our way to success. Less simple to put aside are ordinary practices that can increase the likelihood of publishing false results, particularly those practices that are common, accepted, and even appropriate in some circumstances."

[7]This concern is explicitly raised by John et al. (2012): "QRPs are the steroids of scientific competition, artificially enhancing performance and producing a kind of arms race in which researchers who strictly play by the rules are at a competitive disadvantage [...] the prevalence of QRPs raises questions about the credibility of research findings and threatens research integrity by producing unrealistically elegant results that may be difficult to match without engaging in such practices oneself. This can lead to a 'race to the bottom' with questionable research begetting even more questionable research."

that there is a fixed number of results that will be published,[8] and that if other researchers engage in 'more severe' forms of misconduct, they gain an advantage in publishing.

Already this very simple environment can be used to some profit to examine some key issues regarding the proposals for transparent reporting (Simmons et al., 2011, Landis et al., 2012, Fanelli, 2013). In particular, if enforced, such guidelines can increase the cost of 'mild' forms of QRP but they cannot rule out more condemnable forms, such as data falsification or fabrication. If changing the cost of some type of QRP affects the rate in which other types occur, the overall effect might not be the desired one. It seems very plausible that ruling out mild forms of misconduct might lead to an increase in outright fabrication. Even if we assume that each set of proposals will effectively increase the cost of the QRPs it targets, how do we know that an increase in harder forms of QRP is not likely to happen and counteract any beneficial effects?

Our simple model can thus be used to assess the logic of the proposed solutions. The model shows that under a very simple set of assumptions, ruling out mild forms of misconduct has an unambiguously beneficial effect, since is tends to eradicate the occurrence of severe forms of QRP as well. This somewhat surprising result is due to the fact that reducing *mild* QRP would lower the publication standards with respect to the degree of perfection and significance needed for scientific success. This would lower the incentives for outright fraud, since it has very high cost, and publishing without resorting to QPR is now more likely to be rewarded, and thus more frequent. Secondly, the model indicates that measures for tackling severe forms of QRP, such as government audits of data and statistical techniques for detecting fabrication (Simonsohn, 2012) are not the most effective way for reducing the overall rate of QRP as long as milder forms of QPR are used.

[8]Of course, this can be thought of as the set of all papers in journals that are relevant for career advancement purposes. We note that the payoff structure is also consistent with a world in which the rank of one's result in one's peer group matters, for instance, because of career concerns.

3 A Simple Model

Although it necessarily rests on simplifying assumptions, a mathematical model can help clarify ideas and guide future research. If a given set of assumptions seems questionable, the appropriate response is to see what the model predicts under alternative assumptions. If the predictions are robust across what seems like the domain of logical and empirically plausible assumptions, then the model is useful for guiding practice and policy. Our model will be able to capture 'general equilibrium' effects resulting from strategic interaction, that is, fully accounting for changes in payoffs for certain behavior brought about by the strategic response to changes in institutions. This is a key virtue of economic theory.

To illustrate the main point, we shall use a very simple setting, and employ equilibrium analysis. This means that we shall examine aggregate behavior – aggregate behavior simply comprises individuals' behaviors – such that each individual's behavior is 'the best possible' according to the actor's payoff, given the behavior of others. Therefore, our analysis should be interpreted as predicting what will happen in the long run, when all effects of learning have taken place and behavior has stabilized. This is important in order to evaluate the set of environments to which the theory has predictive power. We believe that scientists receive regular feedback on their performance relative to others, and are capable of eventually learning how society behaves. These are conditions that generally ensure that behavior in the long run will represent a Nash Equilibrium (Fudenberg and Levine, 1998). If these conditions do not describe well the environment of interest, the Nash equilibrium predictions will lack plausibility.

We shall illustrate how the long-run predictions about the prevalence of misconduct change under alternative assumptions. First we focus on homogeneous researchers, and examine the efficacy of alternative policies in reducing the overall rate of QRPs. In Part 4 we tackle the more realistic but also complex case where researchers differ in their psychic cost of misconduct (this is equivalent to letting the rewards vary, which would capture heterogeneity in research ability – researcher types will capture an aggregate of

psychic cost of cheating and research ability).[9] As we will show the model is capable of delivering strong messages about the differential effects of different policies, and these messages are relatively robust to the different settings of the model.

3.1 Homogeneous Researchers

Suppose there are two researchers who each have obtained a research result. Both the results are of comparable interest to the scientific community, but as a matter of fact, they could be more surprising and interesting. To remedy this deficiency each researcher has the opportunity to tune up their work, exerting creative effort e. To keep things simple suppose there are three possibilities:

(i) Report the results as they are, i.e. choose effort $e = 0$, which does not incur any cost.

(ii) Carefully omit some of the more "boring" parts of the results, to bring out the original parts (i.e., suppressing evidence), choosing $e = \underline{e}$.

(iii) Creatively improve the results to increase the novelty of the research (i.e., fabrication), corresponding to a choice $\overline{e}$.

We assume that the possible levels of creative effort $0 < \underline{e} < \overline{e}$ correspond to the cost of unease when performing the respective cosmetic action.[10] After exerting creative effort (or not) the researchers present their results to an

[9] More generally, one would want to allow for the possibility that the psychic cost of misconduct, for instance through feelings of guilt, is a fixed characteristic of the researcher, while the reward from QRP will depend on the results of the current research project relative to expectation (if the true result is exceptional, there will be little need of tampering), which may vary over projects. Here we focus on a static setting, and translating our results into empirical predictions will therefore require controlling for project quality.

[10] This cost may have several interpretations. It can simply correspond to a psychological aversion to cheating, it could capture the probability of detection and punishment, and finally it could represent the intrinsic difficulty in cooking the evidence. Our setting thus relates to the theoretical literature on strategic communication with lying costs (Kartik, 2009), where researchers are 'senders' of information that try to affect the behavior of an editor–'receiver'.

interested scientific audience by submitting the results to a journal. Assume there is only one journal with a capacity of one article.[11] The journal decides to publish the result that appears more original and novel, and tosses a fair coin if indifferent.

Thus the probability of publication of a researcher's result can be described by a function $P(e, e')$ that depends on own creative effort e and that of the competition, e'. A researcher has payoff $R > 0$ from publishing his work in the distinguished journal. Hence, a researcher's expected total payoffs from exerting creative effort e given e' is given by:

$$u = P(e, e')R - e.$$

Note that an equivalent model could set the reward $R = 1$ but explicitly model a common cost deflator e/k. Assume that

$$0 < \underline{e} < R/2 < \overline{e} < R/2 + \underline{e}. \tag{1}$$

This assumption is best interpreted as saying that a moderate level of misbehavior increases the publication probability sufficiently to offset psychic cost if the opponent does not cheat. Severe misconduct is inducing a too high psychic cost to make its use profitable if the opponent does not cheat at all. If one's opponent is employing mild misconduct, however, engaging in severe QRP increases the publication probability enough to compensate for the increase in psychic cost. This setup with two players who have three possible actions each can be represented by a 3×3 matrix of payoffs of the game in the standard way. Notice that the game is symmetric, in that possible actions and payoffs have the same form for both players. Working with the matrix game it is then easy to see that the best reply function is the following:

$$e^*(e') = \begin{cases} \underline{e} & \text{if } e' = 0 \\ \overline{e} & \text{if } e' = \underline{e} \\ 0 & \text{if } e' = \overline{e} \end{cases}$$

This implies that there is no Nash equilibrium in pure strategies, as there is no pair of actions such that the actions are mutually best replies to each

[11] In Part 4 we shall show that this assumption can be easily generalized. What is crucial is that only a limited number of results will receive rewards in form of public recognition.

other. A Nash equilibrium in mixed strategies is guaranteed by standard arguments, however. Therefore we turn to examine equilibria in mixed strategies.

Let $\underline{q}$ and $\overline{q}$, associated to effort levels $\underline{e}$ and $\overline{e}$ respectively, denote the opponent's mixing distribution: i.e., a player's opponent chooses $\underline{e}$ with probability $\underline{q}$ and $\overline{e}$ with probability $\overline{q}$. Then in a Nash equilibrium in mixed strategies researcher i must be indifferent between all actions chosen with positive probability. Hence, if a researcher uses all three actions with positive probability, the following will necessarily hold:

$$(1-\underline{q}-\overline{q})R/2 = (1-\underline{q}-\overline{q})R + \underline{q}R/2 - \underline{e} = (1-\overline{q})R + \overline{q}R/2 - \overline{e}.$$

This states the simple fact that in order for all strategies of player i to be played in equilibrium, they must give the same expected payoff to the player. The above equations can be solved to pinpoint the mixing probabilities of the opponent, yielding:

$$\underline{q} = \frac{2\overline{e}}{R} - 1 \text{ and } \overline{q} = 1 - \frac{2\underline{e}}{R}.$$

By symmetry, this distribution fully characterizes the Nash equilibrium in mixed strategies for both players. Note that $\underline{q}+\overline{q} < 1$ under our assumption above. Indeed the mixing probabilities can be interpreted as frequencies in games played in a large population (also see the continuum version). The equilibrium seems counterintuitive in that the equilibrium frequency of each form of QPR depends on the cost of the *other* form of QPR (as in any game). The intuition is that when the cost of a given action increases, for a short amount of time this action will tend to be played less in the population. However, this will change the expected payoff of the other action. Eventually, the frequency of this other action adjusts enough to counteract the initial cost increase. So, in the long run, the change in cost of a given action will not affect its frequency of being played.

The equilibrium in our game has an interesting property: there is an asymmetry in the effect of increasing costs for the different forms of QRP. The frequency of mild misbehavior $\underline{q}$ will increase if the cost of severe misbehavior $\overline{e}$ goes up, while the rate of severe misbehavior $\overline{q}$ decreases in the cost of mild misbehavior $\underline{e}$. This means that severe misconduct, which generates a large

advantage at high cost, crowds out mild misconduct, since the presence of severe misconduct drives down the expected return of mild misconduct. Mild misconduct begets severe misconduct, however, since severe misconduct is optimal only if one's opponent employs mild QRP.

It is important to emphasize that achieving this mixed strategy equilibrium requires very mild assumptions. We do not need to assume that any person knows the payoffs of other researchers, or that individuals actually use randomization. The game can be interpreted as capturing interaction among a large number of researchers, who are randomly matched with each other. The equilibrium is expected to occur only after a large number of interaction has taken place, and each individual has accumulated enough experience. The mixing probabilities then can be understood as the frequency of each strategy in the population, while each individual chooses a strategy with probability 1. Moreover, the individuals need not know the exact structure of the game. Indeed it suffices that enough feedback is given each time the game is played: each player should know what the opponent has chosen every time the game is played (Fudenberg and Levine, 1998).

In this paper we are interested in the total aggregate misconduct ($\overline{q} + \underline{q}$). Note first that according to the equilibrium strategies this sum is equal to $2(\overline{e} - \underline{e})/R$. This means that aggregate misconduct is increasing in the cost of severe misconduct and decreasing on the cost of mild misconduct. As we shall see, the result that increasing the cost of mild misbehaviour is effective in reducing total misconduct is very robust. Let us consider a policy of prohibitively increasing this cost. The fact that as severe misconduct becomes easier the return to mild misconduct decreases gives rise to the following observation.

Lemma 1. *Suppose there is a policy that completely prevents some action, either* $\overline{e}$ *or* $\underline{e}$*. If cheating a lot is prevented, the aggregate frequency of cheating increases, i.e.,* $\underline{q} = 1$ *in the unique Nash equilibrium. If cheating a little is prevented, the aggregate frequency of cheating decreases, i.e.,* $\overline{q} = 0$ *in the unique Nash equilibrium.*

To see this suppose first that severe misconduct (action $\overline{e}$) becomes impossible. In the new game with actions 0 and $\underline{e}$ mild misconduct is now a

dominant strategy, since removing $\overline{e}$ has increased the likely return of mild misconduct. Therefore the unique Nash equilibrium is $(\underline{e}, \underline{e})$. Now suppose that mild cheating (action $\underline{e}$) becomes impossible. In the new game with actions 0 and $\overline{e}$ it is now a dominant strategy not to misbehave at all, since removing $\underline{e}$ has removed the only action that severe misconduct was a best reply to!

Lemma 2. *Increasing the reward for publication from R to $R' > R$ with $R' < 2\overline{e}$ yields more severe misbehavior ($\overline{q}' > \overline{q}$) and less mild misbehavior ($\underline{q}' < \underline{q}$). Moreover, the overall misbehavior declines: $\overline{q}' + \underline{q}' < \overline{q} + \underline{q}$.*

This lemma states that moderately increasing the reward of publication (e.g. in form of career concerns or importance for tenure) has a surprising effect: it crowds out mild misbehavior by increasing incentives for severe misbehavior and reduces the aggregate prevalence of QRP. This is in contrast with simple intuition which says that an increase in the publish-or-perish culture is likely to lead to more biased research. However, notice that if R increases enough to achieve $2\overline{e} \leq R'$, then severe misbehaviour becomes the dominant strategy. It therefore seems that an increase in R is not detrimental only if it is mild.

Robustness Checks

To assess the role of our assumptions let us now dispose of Assumption 1. That is, $\underline{e}$ and $\overline{e}$ may take any values such that $0 < \underline{e} < \overline{e}$. Indeed the following result shows that prohibitively increasing the cost of mild misbehavior is a minimum-risk option, in the sense that it cannot increase the overall frequency of misconduct.

Lemma 3. *For any parameters $0 < \underline{e} < \overline{e}$, removing action $\underline{e}$ from the action set will (weakly) decrease the aggregate frequency of misconduct.*

To see this, consider first the case where $\underline{e} < R/2$ and $R/2 + \underline{e} < \overline{e}$. It is easy to verify that $\overline{e}$ is dominated because it is too costly, and the unique Nash equilibrium is now $(\underline{e}, \underline{e})$. Removing $\overline{e}$ will not change the equilibrium. However, removing $\underline{e}$ will yield $(0, 0)$ as the new Nash equilibrium, meaning that ruling out mild misbehavior improves things.

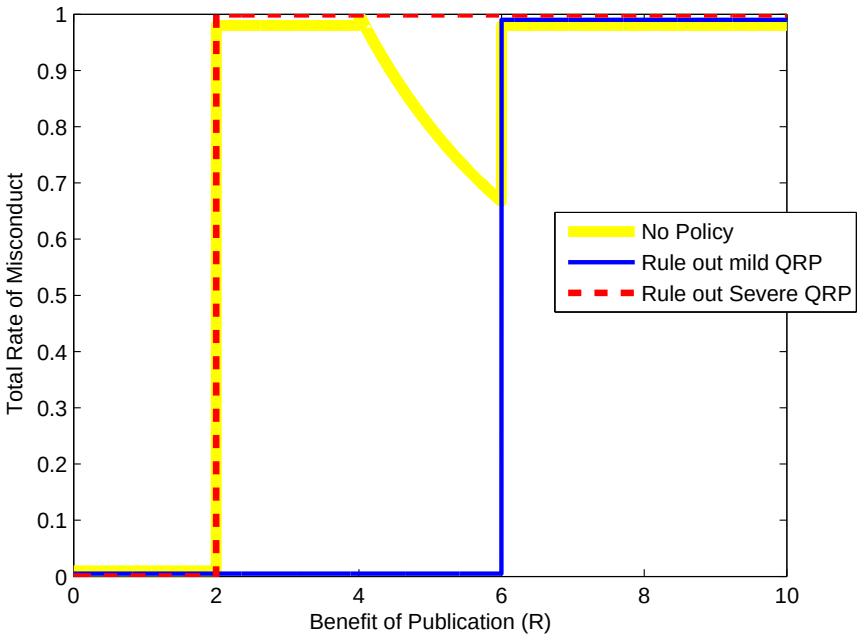

Figure 1: Frequency of total misconduct for the three alternative policies for various levels of rewards.

If $\underline{e} < \overline{e} < R/2$, $\overline{e}$ is a best response to both forms of QRP, and the unique Nash equilibrium is $(\overline{e}, \overline{e})$. Removing $\underline{e}$ does not change this, but removing $\overline{e}$ yields $(\underline{e}, \underline{e})$ as the new equilibrium.

Finally, if $R/2 < \underline{e} < \overline{e}$ any kind of misconduct is too costly compared to sincere reporting and the unique equilibrium is $(0, 0)$. Since for any game that results from ruling out some form of misconduct choosing $e = 0$ is a dominant strategy, no policy can change the equilibrium in this case.

Figure 1 illustrates the total level of cheating for various levels of rewards R when $\underline{e} = 1$ and $\overline{e} = 3$. It is clear that the policy of ruling our mild misbehaviour results in a lower rate of total misconduct compared either with the absence of intervention or the policy of striking down severe misbehaviour. This shows that this policy has a relatively low risk of backfiring, even when

there is relative uncertainty about key variables such as the benefits of publication.

The analysis so far has shown us that in a very simple setting the implementation of a transparency regime is robustly successful in reducing (or non-increasing) the overall rate of misconduct. In the next section we examine a more general case, allowing for heterogeneous individuals and a large number of possible outlets for publication, and find that our results carry over even in this case (see Proposition 2).

4 A Model with Heterogeneous Researchers

Suppose now that there is a continuum of researchers, endowed with probability mass 1. Researchers are characterized by their type θ, which reflects the marginal cost of misconduct: a researcher incurs a disutility, or psychic cost, θe from choosing a level $e \in \{0; \underline{e}; \overline{e}\}$ of misconduct. Let θ follow a uniform distribution on $[0, 1]$ for illustrative purposes.[12]

Researchers' rewards are given by the expected quality of the outlet in which the results can be published. There is a number of distinguished journals that are relevant for the researchers' career concerns. Suppose again that being published yields reward $R > 0$. To keep matters tractable suppose that these distinguished journals publish a mass $\kappa < 1$ of articles. As above suppose that bar any scientific misconduct all researchers' result are equally original.

Hence, a researcher's payoff from not committing any misconduct is κR, if the share of agents who engage in misconduct is zero. Denote the share of agents who choose $\underline{e}$ by $\underline{q}$ and of those who choose $\overline{e}$ by $\overline{q}$. Then a researcher's payoff from abstaining from misconduct is

$$E[u] = \max\{0; (\kappa - \underline{q} - \overline{q})/(1 - \underline{q} - \overline{q})\}R.$$

A researcher who chooses $\underline{e}$ has expected payoff

$$E[u] = \min\{1; \max\{0; (\kappa - \overline{q})/\underline{q}\}\}R - \theta\underline{e},$$

[12] Simulations suggest that our results do not depend on the distributional assumption on θ.

and a researcher who chooses $\overline{e}$ has expected payoff

$$E[u] = \min\{1; \kappa/\overline{q}\}R - \theta\overline{e}.$$

These payoffs define cutoffs $\theta^*(e_{low}, e_{high})$ for binary comparison of action choices, such that agents of types $\theta < \theta^*(.)$ prefer the choice e_{high} involving more severe misconduct, and the opposite is true for types $\theta > \theta^*(.)$:

$$\begin{aligned}
\theta^*(0,\underline{e}) &= \left(\min\left\{1; \max\left\{0; \frac{\kappa-\overline{q}}{\underline{q}}\right\}\right\} - \max\left\{0; \frac{\kappa-\underline{q}-\overline{q}}{1-\underline{q}-\overline{q}}\right\}\right)\frac{R}{\underline{e}},\\
\theta^*(0,\overline{e}) &= \left(\min\left\{1; \frac{\kappa}{\overline{q}}\right\} - \max\left\{0; \frac{\kappa-\underline{q}-\overline{q}}{1-\underline{q}-\overline{q}}\right\}\right)\frac{R}{\overline{e}},\\
\theta^*(\underline{e},\overline{e}) &= \left(\min\left\{1; \frac{\kappa}{\overline{q}}\right\} - \min\left\{1; \max\left\{0; \frac{\kappa-\overline{q}}{\underline{q}}\right\}\right\}\right)\frac{R}{\overline{e}-\underline{e}}.
\end{aligned}$$

The distribution of θ and these cutoffs then determine measures $\underline{q}$ and $\overline{q}$.

Proposition 1. *Suppose $R \leq 2\underline{e}$ or $\kappa \in [0, 7/8]$. Then the equilibrium is unique and there are three possible equilibrium regimes:*

1. *for $R/\overline{e} \geq \kappa$, $\underline{q} = 0$ and $\overline{q} \geq \kappa$, a share $\overline{q} = \sqrt{\frac{\kappa R}{\overline{e}}}$ engages in severe misconduct.*

2. *for $R/\overline{e} < \kappa < R/\underline{e}$ both $\underline{q} > 0$ and $\overline{q} > 0$, and $\overline{q} < \kappa < \underline{q} + \overline{q}$.*

3. *for $\kappa \geq R/\underline{e}$, $\overline{q} = 0$ and $\underline{q} < \kappa$, a share $\underline{q} = \frac{1}{2}\left(1 - \sqrt{1 - 4(1-\kappa)\frac{R}{\underline{e}}}\right)$ engages in mild misconduct.*

Proof. We start by noting that there are three different, mutually exclusive possible equilibrium regimes: (i) $\underline{q} = 0$, which implies that $\overline{q} \geq \kappa$ (since otherwise $\underline{e}$ would dominate $\overline{e}$ for all types), (ii) $\overline{q} = 0$, which implies that $0 < \underline{q} \leq \kappa$ (since otherwise types θ close to 0 would prefer $\overline{e}$ to $\underline{e}$), and (iii) $\underline{q} > 0$ and $\overline{q} < \kappa$, which implies that $\underline{q} + \overline{q} > \kappa$ (since otherwise $\underline{e}$ dominates $\overline{e}$ for all types).

Case (i): Note that both 0 and $\underline{e}$ yield a publication probability of 0. Hence, $\underline{e}$ is dominated by 0 and $\underline{q} = 0$. The relevant cutoff is therefore

$$\theta^*(0,\overline{e}) = \frac{\kappa}{\overline{q}}\frac{R}{\overline{e}}.$$

With a uniform distribution of θ, $\overline{q} = \theta^*(0, \overline{e})$. This implies

$$\overline{q} = \sqrt{\frac{\kappa R}{\overline{e}}}.$$

This regime requires $\kappa \leq \overline{q}$, that is, $\kappa \leq R/\overline{e}$.

Case (ii): Note that both $\overline{e}$ and $\underline{e}$ yield a publication probability of 1. Hence, $\overline{e}$ is dominated by $\underline{e}$, and $\overline{q} = 0$. The relevant cutoff is therefore

$$\theta^*(0, \underline{e}) = \frac{1-\kappa}{1-\underline{q}} \frac{R}{\underline{e}}.$$

With a uniform distribution of θ, $\underline{q} = \theta^*(0, \underline{e})$. This implies a negative root

$$\underline{q} = \frac{1}{2}\left(1 - \sqrt{1 - 4(1-\kappa)\frac{R}{\underline{e}}}\right).$$

This regime requires $\kappa \geq \underline{q}$, that is, $\kappa \geq R/\underline{e}$, which also implies that the term under the root is positive. The positive root satisfies $\kappa \geq \underline{q}$ only for $\kappa \leq R/\underline{e}$. To have a positive term under the root, $1 - \underline{e}/(4R) \leq \kappa \leq R/\underline{e}$. Moreover, $\underline{q} > 1/2$, therefore $R \geq 2\underline{e}$, which implies $\kappa \geq 7/8$.

Case (iii): Now both $\underline{q} > 0$ and $\overline{q} > 0$. Transitivity of preferences over actions is only consistent with $\theta^*(0, \underline{e}) > \theta^*(0, \overline{e}) > \theta^*(\underline{e}, \overline{e})$. Equilibrium shares are thus $\overline{q} = \theta^*(\underline{e}, \overline{e})$ and $\underline{q} = \theta^*(0, \underline{e}) - \theta^*(\underline{e}, \overline{e})$. The cutoff types are given by

$$\theta^*(0, \underline{e}) = \tfrac{\kappa - \overline{q}}{\underline{q}} \tfrac{R}{\underline{e}},$$
$$\theta^*(\underline{e}, \overline{e}) = \left(1 - \tfrac{\kappa - \overline{q}}{\underline{q}}\right) \tfrac{R}{\overline{e} - \underline{e}}.$$

Equilibrium shares of agents are therefore given by

$$\overline{q} = \tfrac{(\kappa - \underline{q})\frac{R}{\overline{e}-\underline{e}}}{\frac{R}{\overline{e}-\underline{e}} - \underline{q}},$$
$$\underline{q}^2 = \left((\kappa - \overline{q})\tfrac{\overline{e}}{\underline{e}} - \underline{q}\right) \tfrac{R}{\overline{e}-\underline{e}}.$$

Solving the system of equations yields either $\underline{q} = 0$ and $\overline{q} = \kappa$ (a contradiction, as it implies that $\theta^*(0, \underline{e}) \to \infty$ and thus the relevant cutoff becomes $\theta^*(0, \overline{e})$, so that case (i) obtains), or

$$\underline{q} = \sqrt{\tfrac{R^2 - R\kappa\overline{e}}{\underline{e}(\underline{e} - \overline{e})}} = \tfrac{R}{\overline{e}-\underline{e}} \sqrt{1 - \tfrac{\overline{e}}{\underline{e}}\left(1 - \kappa\tfrac{\overline{e}-\underline{e}}{R}\right)}.$$

The term under the root is positive if $\kappa \geq R/\overline{e}$. Note that $\underline{q} = 0$ for $\kappa = R/\overline{e}$ and strictly increases in κ. $\overline{q} = \kappa$ for $\kappa = R/\overline{e}$, strictly decreases in $\underline{q}$ for $\kappa < R/(\overline{e} - \underline{e})$, and strictly increases for $\kappa > R/(\overline{e} - \underline{e})$.

For $\underline{q} + \overline{q} > \kappa$ it is needed that $\kappa < R/\underline{e}$. To see this use the expressions for $\overline{q}$ and $\underline{q}$ above, distinguish the cases of $\sqrt{1 - \frac{\overline{e}}{\underline{e}}\left(1 - \kappa\frac{\overline{e}-\underline{e}}{R}\right)} \gtrless 1$ for $\kappa \gtrless R/(\overline{e} - \underline{e})$, and solve for the threshold value of κ, which is $R/\underline{e}$ in both cases. □

While the assumption $\kappa \in [0, 7/8]$ appears plausible in the context of our application the equilibrium behavior when it is violated may still be of interest. In that case, for $1 - \underline{e}/(4R) \leq \kappa \leq R/\underline{e}$ there is another equilibrium regime with

$$\underline{q} = \frac{1}{2}\left(1 + \sqrt{1 - 4(1-\kappa)\frac{R}{\underline{e}}}\right)$$

and $\overline{q} = 0$ along the one described in the Proposition (both $\underline{q} > 0$ and $\overline{q} > 0$, and $\overline{q} < \kappa < \underline{q} + \overline{q}$).

Policy A: Preventing $\overline{e}$

Suppose now that the cost for engaging in severe misconduct $\overline{e}$ is prohibitively high, so that $\overline{q} = 0$, but $\underline{e}$ remains unchanged. The equilibrium is now determined by

$$\theta^*(0, \underline{e}) = \left(\min\{1; \kappa/\underline{q}\} - \max\{0; (\kappa - \underline{q})/(1 - \underline{q})\}\right)\frac{R}{\underline{e}}, \tag{2}$$

and

$$\underline{q} = \theta^*(0, \underline{e}),$$

where we used the uniform distribution. Distinguishing between the possible regimes $\underline{q} < \kappa$ and $\underline{q} \geq \kappa$, the share $\underline{q}$ has to satisfy

$$\underline{q} = \left(1 - \frac{\kappa - \underline{q}}{1 - \underline{q}}\right)\frac{R}{\underline{e}} \text{ and } \underline{q} = \frac{\kappa}{\underline{q}}\frac{R}{\underline{e}},$$

respectively. Therefore $\underline{q} = \sqrt{\kappa R/\underline{e}}$ if $\kappa \leq R/\underline{e}$, $\underline{q} = 1/2 - \sqrt{1/4 - (1-\kappa)R/\underline{e}}$ if $\kappa > R/\underline{e}$ and $\underline{q} = 1/2 + \sqrt{1/4 - (1-\kappa)R/\underline{e}}$ if $1 - \underline{e}/(4R) \leq \kappa \leq R/\underline{e}$. Again $R \leq 2\underline{e}$ or $\kappa \in [0, 7/8]$ implies uniqueness of the equilibrium allocation.

Policy B: Preventing $\underline{e}$

Suppose now that the cost for engaging in mild misconduct $\underline{e}$ is prohibitively high, so that $\underline{q} = 0$, but $\overline{e}$ remains unchanged. The equilibrium is now determined by

$$\theta^*(0, \overline{e}) = \left(\min\{1; \kappa/\overline{q}\} - \max\{0; (\kappa - \overline{q})/(1 - \overline{q})\}\right) \frac{R}{\overline{e}}, \tag{3}$$

and

$$\overline{q} = \theta^*(0, \overline{e}).$$

Analogously to Policy A, equilibrium measures are $\overline{q} = \sqrt{\kappa R/\overline{e}}$ if $\kappa < R/\overline{e}$, $\overline{q} = 1/2 - \sqrt{1/4 - (1-\kappa)R/\overline{e}}$ if $\kappa > R/\overline{e}$ and $\overline{q} = 1/2 + \sqrt{1/4 - (1-\kappa)R/\overline{e}}$ if $1 - \overline{e}/(4R) \leq \kappa \leq R/\overline{e}$. Again $R \leq 2\overline{e}$ or $\kappa \in [0, 7/8]$ implies uniqueness of the equilibrium allocation.

Denote by $\underline{q}^*$ and $\overline{q}^*$ the measures associated to a laissez faire equilibrium, and by $\underline{q}^A$ and $\overline{q}^B$ the ones associated to policies A and B.

Proposition 2. *Suppose $R \leq 2\underline{e}$ or $\kappa \in [0, 7/8]$. Then the equilibrium is unique under all regimes considered. Preventing $\overline{e}$ (policy A) will not decrease the equilibrium frequency of misconduct, while preventing $\underline{e}$ (policy B) will not increase the equilibrium frequency of misconduct.*

Proof. Note first that comparing equilibrium cutoffs (2) and (3) reveals that $\overline{q}^B < \underline{q}^A$, i.e., the total incidence of misconduct is lower under policy B than under A.

Suppose $\kappa \leq R/\overline{e}$. Then $\overline{q}^B = \overline{q}^* < \underline{q}^A$ by equilibrium cutoffs (2) and (3).

Suppose $\kappa \geq R/\underline{e}$. Then $\underline{q}^A = \underline{q}^* > \overline{q}^B$ by equilibrium cutoffs (2) and (3).

Suppose that $R/\overline{e} < \kappa < R/\underline{e}$. Tedious calculations reveal that in this case $\underline{q}^* + \overline{q}^* < \underline{q}^A$ as defined above. Since $\underline{q}^* + \overline{q}^* > \kappa$ in this regime, $\overline{q}^B \leq \kappa$, which is quickly verified, is sufficient for $\overline{q}^B < \underline{q}^* + \overline{q}^*$. □

Numerical Example

Suppose that $\kappa = 1/3$, $R = 1$. Let $\underline{e} = 2$ and $\overline{e} = 4$ so that half the population would cheat a little to achieve probability 1 of success instead of 0, and a quarter would cheat a lot for this.

Then $\underline{q} = .2887$ and $\overline{q} = .1057$ under laissez faire. Preventing $\overline{e}$ results in an increase of 3.5% in the prevalence of cheating ($\underline{q}' = .4082$), while preventing $\underline{e}$ results in a decrease of 53.6% in the frequency of cheating ($\overline{q}' = .2113$).

Figure 2 shows the incidence of cheating for varying capacity κ under the three different policies. Policy A is captured by the dotted line, coinciding with laissez faire for higher values of κ. Policy B is captured by the broken line coinciding with laissez faire for low values of κ. Solid and dashed lines correspond to the laissez faire outcome, giving the total frequency of cheating (solid line), and the frequency of $\underline{e}$ (dashed line). As can be seen from Figure 2, Policy B leads to a lower overall level of misconduct - relative to the absence of any policy - for any level of capacity κ. On the other hand, Policy A, which rules out severe misconduct, performs worse in terms of the overall rate of misconduct, for small values of κ. Once more, our main result, that a regime of increased transparency that prohibitively increases the costs of 'lying by omission' will result in weakly less overall misconduct, is remarkably robust.

5 Discussion

The extent of the recent public debate about QRP, further stimulated by a number of high profile incidents, generates a clear mandate to assess the effectiveness of the alternative proposals for tackling the problem of false positives. Any form of investigation into fraudulent behavior is, by the nature of its subject, bound to face severe difficulties. In this paper we suggested a simple game-theoretic framework that may help us in conceptualizing the problem and identifying some key trade-offs. Our results are important for providing rigorous theoretical guidance on where the public should direct its efforts for improving the credibility of science. In particular, our model indicates that targeting mild forms of misconduct is indeed more efficient than targeting severe ones, such as outright fraud.

This is good news in a sense, as the prevention of outright fabrication of results may in fact be very difficult[13] and explicit audits by an outside body

[13]Nosek et al. (2012) argue that "Notably, it is difficult to detect deliberate malfeasance.

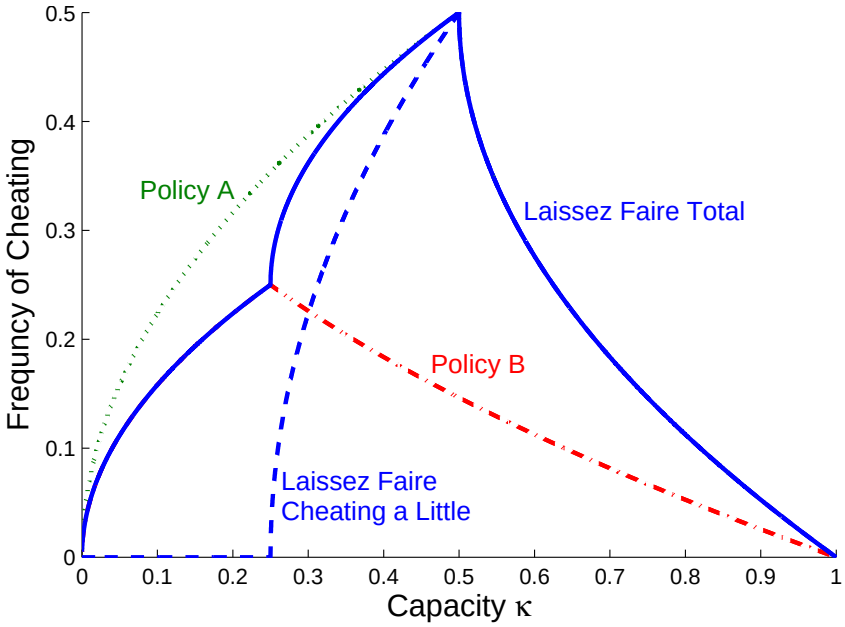

Figure 2: Frequency of cheating for the different policies

are very costly and highly unpopular among scientists.[14] Our model indicates that such extreme measures may not be necessary, at least if researchers face a high cost of engaging in practices of severe misconduct, be it by way of psychic cost, e.g., feelings of guilt, or high penalties incurred in the unlikely, but possible event of being found out by chance.[15] Indeed this reasoning appears similar to "broken windows" theories. The main difference is that our argument does not rely on an erosion of social norms, but rather on the erosion of possible rewards when not engaging in misconduct, focusing on the use of QRP as a form of self-defence.

It should be strongly emphasized that our study focuses on the overall rate of misconduct, treating the two forms of QRP as 'similarly bad'. This approach is more reasonable than what seems at first glance. Although ethically these two practices differ, we are interested in the overall credibility of research results. It is not well understood how much each type of misbehaviour affects the credibility of the published evidence. Is performing multiple studies and revealing the most 'interesting' one less detrimental for discovering the truth about Nature than inventing the relevant data? Possibly yes, in the sense that at least in the former case (which corresponds to mild QRP) the data do come from Nature, although not via a random sample. If a typical case of severe QRP distorts the evidence much more than a typical case of mild QRP, then our focus on the total rate of misbehaviour might be misplaced. In this case, in order to generate valid insights one needs to consider a value function weighting each form of misconduct differently, and possibly in a non-linear way. More research is needed.

The three most prominent cases in psychology's recent history - Karen Ruggiero, Marc Hauser, and Diederik Stapel - were not identified by disconfirmation of their results in the published literature (though, in Hausers case, there was some public skepticism for at least one result). The misbehavior was only identified because colleagues - particularly junior colleagues - took considerable personal risk by voicing concerns about the internal practices of the laboratory."

[14]Greenberg and Goldberg (1994) find that less than 16 percent of surveyed environmental and research economists found any usefulness in any form of government audit or intervention.

[15]The panoply of harsh criticisms against Michael J. LaCour, the lead author of a recent political science study that was retracted by *Science*, and the attention drawn to the incident point to a potentially very high expected cost.

As any theory, our theory has to be evaluated in terms of the degree to which it captures the essential aspects of the problem, which we believe to be a modicum of rational behavior by scientists, the competitive character of the game of publication, and a clear hierarchy of QRPs in terms of their moral defensibility. Furthermore, as any model, our model critically depends on its assumptions, but the results are quite robust. In particular, even in environments where the cost of engaging in severe misconduct is relatively low, the policy of preventing mild QRPs will not do harm. Importantly, the assumptions about the relative magnitude of the cost of engaging in each type of QRP can be tested, and this is precisely what we plan to do next by conducting empirical studies. Also note that in order for the transparency proposals to work as our theory predicts, they have to substantially increase the cost of engaging in mild QRP - which they target. This is another hypothesis that should and can be tested empirically.

References

Richard A Bettis. The search for asterisks: compromised statistical tests and flawed theories. *Strategic Management Journal*, 33(1):108–113, 2012.

Daniele Fanelli. How many scientists fabricate and falsify research? a systematic review and meta-analysis of survey data. *PLOS one*, 4(5):e5738, 2009.

Daniele Fanelli. Redefine misconduct as distorted reporting. *Nature*, 494 (7436):149, 2013.

Drew Fudenberg and David K. Levine. *The theory of learning in games*, volume 2. MIT press, 1998.

Michael Greenberg and Laura Goldberg. Ethical challenges to risk scientists: an exploratory analysis of survey data. *Science, Technology & Human Values*, 19(2):223–241, 1994.

John PA Ioannidis. Why most published research findings are false. *PLoS medicine*, 2(8):e124, 2005.

John PA Ioannidis. Why science is not necessarily self-correcting. *Perspectives on Psychological Science*, 7(6):645–654, 2012.

Michael D Jennions and Anders P Møller. Relationships fade with time: a meta-analysis of temporal trends in publication in ecology and evolution. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 269(1486):43–48, 2002.

Leslie K John, George Loewenstein, and Drazen Prelec. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science*, 23(5):524–532, 2012.

Navin Kartik. Strategic communication with lying costs. *The Review of Economic Studies*, 76(4):1359–1395, 2009.

Richard A Klein, Kate A Ratliff, Michelangelo Vianello, Reginald B Adams Jr, Štěpán Bahník, Michael J Bernstein, Konrad Bocian, Mark J Brandt, Beach Brooks, Claudia Chloe Brumbaugh, et al. Investigating variation in replicability: A many labs replication project. *Social Psychology*, 45(3):142, 2014.

Story C Landis, Susan G Amara, Khusru Asadullah, Chris P Austin, Robi Blumenstein, Eileen W Bradley, Ronald G Crystal, Robert B Darnell, Robert J Ferrante, Howard Fillit, et al. A call for transparent reporting to optimize the predictive value of preclinical research. *Nature*, 490(7419): 187–191, 2012.

Jonah Lehrer. The truth wears off. *The New Yorker*, 13:52, 2010.

Michael J Meyer and Dave McMahon. An examination of ethical research conduct by experienced and novice accounting academics. *Issues in Accounting Education*, 19(4):413–442, 2004.

Brian A Nosek, Jeffrey R Spies, and Matt Motyl. Scientific utopia ii. restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, 7(6):615–631, 2012.

Joseph P Simmons, Leif D Nelson, and Uri Simonsohn. False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, 22(11):1359–1366, 2011.

Uri Simonsohn. Just post it: The lesson from two cases of fabricated data detected by statistics alone. *Available at SSRN 2114571*, 2012.