Evaluating Solutions for the Problem of False Positives*

Thomas Gall and Zacharias Maniadis†

October 14, 2017

Abstract

A current challenge for the scientific community is the choice of appropriate policies to reduce the rate of false positives. Existing proposals differ in whether to prioritize tackling omission through transparency requirements, punishing more severe transgressions, or possibly both. We use a formal model to evaluate these possible solutions. We find that a policy that prohibitively increases the cost of 'misdemeanor' types of questionable research practices robustly decreases the overall rate of researcher misconduct, because the rate of 'felonies', such as fabrication, also decreases. Therefore proposals that aim to prevent lying by omission by enforcing reporting guidelines are likely to be effective in reducing researcher misconduct, but measures such as government audits (purported to counteract pure fraud) can backfire. Moreover, we find that an increase in the rewards of publication need not increase overall misconduct.

Keywords: Researcher Misconduct, Reproducibility, False Positives

1 Introduction

Is the model of self-correcting science and cumulative scientific growth still valid in the contemporary scientific world? Serious concerns have been voiced in the last few years (Lehrer, 2010, Ioannidis, 2012) that cast doubt on this view and suggest that it is overly optimistic. In fact, it has been argued that there is a "confidence crisis" in several scientific fields, such as psychology (Collaboration et al., 2015),[1] management and economics (Bettis, 2012, Brodeur et al., 2016), and several branches of the biological and human sciences (Ioannidis, 2005, Jennions and Møller, 2002). The main concern is with the rate of false positives, defined as the fraction of newly discovered findings, i.e. causal associations between different empirical phenomena, that correspond to associations that are false in reality (Ioannidis, 2005).[2] Indeed, a crisis of confidence

*We are grateful to JPA Ioannidis and seminar participants at the University of Southampton and the University of East Anglia.

†Corresponding Author. Department of Economics, School of Social Sciences, University of Southampton, UK. Email: z.maniadis@soton.ac.uk

[1]The journal *Perspectives on Psychological Science* had a special issue on the problem in 2012.

[2]In other words, given a new experimental finding, the rate of false positives in a discipline is the likelihood that this result does not correspond to the truth.

may arise if the rate of false positives is unacceptably large, ultimately making society at large distrust published research findings. It is then incumbent upon scientific researchers to find a way to decrease it and restore public trust.

The concern about the reliability of published scientific findings has already sparked a series of responses, see Maniadis et al. (2015) for a review. One approach targets the excessive degrees of freedom in the (unobserved) choice of empirical and statistical methodology that may contribute to the phenomenon of non-replicable research (see e.g. Simmons et al., 2011, Wagenmakers et al., 2011). In particular, there has been a call for more transparency on research and statistical methods, for instance in the form of checklists of items that research reports should include in order to be eligible for publication (Schulz et al., 2010, Moher et al., 2009, Simmons et al., 2011, Landis et al., 2012, Fanelli, 2013). A number of such measures are being enforced – see Collins and Tabak (2014) and McNutt (2014). In essence, these requirements aim at reducing or eliminating behavior that seeks to "improve" research results by adopting favorable methods, for instance "p-hacking" by unreported multiple testing, censoring of data or adjustment in the number of trials.

Of course, such transparency requirements cannot prevent more extreme versions of cheating, for instance inventing or falsifying data. Since such behavior borders on or constitutes outright fraud, it has been suggested that governments could take a more active role in tackling this problem, using measures such as criminalizing research fraud. For instance, the UK parliament is currently examining whether self-regulating mechanisms are sufficient or not for addressing the problem, or whether regulation (with potentially criminal penalties) is necessary (Houses of Parliament, 2017). Such measures would seem to be effective at curbing or eliminating 'severe transgressions' of good scientific conduct – as opposed to 'mild transgressions' that merely exploit flexibility in methodological choice.

Finally, if research results fail to be replicated with sufficient frequency, a possible remedy would be to encourage replication – e.g. Nosek et al. (2012), Nosek and Lakens (2014). This policy would aim at increasing the cost of any form of questionable research practices. Reliably detecting non-replicable results would deter both pure fraud and lying by omission, since both would likely fail replication and thus invite further audits.

Identifying the optimal policy among the ones proposed is in fact surprisingly difficult, because the policies' possible implications are complex and hard to predict. First, policies that prevent only one class of misbehaviour might lead to an increase in other types of misbehaviour. Secondly, all policies come at a high cost to researchers (in form of red tape or liability), publishing teams (burden of control on referees and editors) and society (enforcement through government agencies). Hence, the question of which policy to choose, if any, has generated substantial public debate.

We contribute to this debate by means of a game-theoretic analysis of the prevalence of different forms of Questionable Research Practices (QRP) that highlights the consequences of different policies on researchers' incentives and behavior. There is much to recommend using a theoretical approach to speak to this debate. For one, the empirical study of QRP is notoriously difficult, because it is a sensitive issue plagued by measurement problems and other methodological complications (Fanelli, 2009). Employing formal (game) theory to examine the behavioural change brought about by a change in the rules and associated incentives will provide guidance on the possible behavioral consequences of different policies. Indeed, all the proposed solutions critically

rely 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". Moreover, scientists operate in a highly structured environment, characterised by concrete processes for publicising results, with high stakes in terms of career concerns ("publish or perish"). All this indicates that game theory is an adequate method for examining incentives and institutions in the domain of science.

Specifically, the career concerns induced by the scientific publication "game" closely resemble what in game-theoretic terminology is called a *tournament* (as suggested by e.g. Stephan, 2012). Researchers spend effort to derive research results that then compete for public attention through the publication process. Success of a given research result, i.e. the attention it generates, will depend on the result's relative originality compared to the literature. Researchers may employ various questionable strategies (at some cost) to enhance the expected originality of their research results. Such strategies range from somewhat defensible (e.g. extending the sample size after initially not getting significant results) to universally condemnable practices, such as outright data fabrication or falsification. Note that ethically less defensible strategies also promise greater "improvement" in the results.

This setup is reminiscent of an "arms race". If many researchers employ QRP to inflate their results, this will increase the intensity of competition and any given researcher will find it harder to generate a result that is likely to attract attention. This in turn will increase the temptation to resort to QRP. Hence, given a high prevalence of QRP, employing questionable research practices becomes more attractive, i.e. a best response. As in an arms race, a researcher may find it optimal to counter the use of QRP by escalating, using even more questionable practices. Our formal, mathematical analysis shows that this is indeed the case for plausible models of academic competition.

It is thus important for the scientific community who wishes to assess reform proposals whether affecting one form of QRP will have an effect on the desirability (for an individual researcher) of another one. For instance, while criminalising scientific fraud can make outright fraud prohibitively expensive, it can also make minor transgressions more attractive. As a result, the overall effect may well be the opposite of what is desired. Our theoretical results show that not all is lost, however. Policies that target and successfully curb the use of only "mild" forms of QRP are more effective than policies that curb more severe forms. This is because the former policies will also reduce the prevalence of severe QRP. The reason for this is that in equilibrium, severe QRP is mainly used as a best response to mild QRP. Likewise, blanket bans, i.e. policies that aim to remove all forms of QRP will likely not be optimal, unless they convey a large cost advantage.

Our analysis thus shows that reforms purported to increase transparency and reduce degrees of freedom in empirical analysis will not worsen the incidence of QRP compared to the status quo. Second, such policies are preferable to policies that target severe misconduct. These results are driven by the assumption that engaging in severe misconduct has a significantly higher personal cost for the researcher than choosing mild misconduct.

To our knowledge, such a formal assessment has not yet been part of the discussion on reforming research practices. Previous contributions in the literature have focused e.g. on scientists' decision on what research to undertake and whether to commit fraud or not (Lacetera and Zirulia, 2009), or on whether to engage in monitoring (Kiri et al., 2015), but less on the type of fraud engaged in and the implications for policy target-

ing certain types of misbehavior. Lacetera and Zirulia (2009) also use a game theory to model the behavior of scientists. Their results suggests that reported fraud is not likely to representative of the true extent, and that both better monitoring and reducing career incentives of scientists may induce researchers to change their type of research and not necessarily reduce the incidence of fraud. Kiri et al. (2015) focus instead on the incentives for scientists to monitor their peers' work. Their game theoretic analysis suggests that peer monitoring will not completely remove questionable research practices and emphasise the possibility of free-riding on the monitoring efforts of others. Policies that reduce monitoring costs will then indeed reduce the prevalence of fraud. Our analysis does not explicitly model monitoring or replication efforts, but entails the possibility that the monitors themselves may employ questionable research practices, for instance to derive a desired, attention grabbing result when replicating other work. Abstracting from the possibility of questionable research practices, Ellison (2002) examines the general problem of allocating researchers' time between original research and work to establish robustness of known results. He focuses on the role of social norms governing the scientists' beliefs about the correct amount of robustness checks, which become more demanding over time, reflecting the idea of self-correction.

2 Capturing the Environment

In the remainder of this paper QRP will refer to any malpractices employed by researchers that will distort the scientific evidence.[3] A necessary first step in conceptualizing and modeling the problem is to capture the essential properties of these QRP. In our view the relevant QRP can be categorized in terms of the degree to which they are acceptable practices, and thus can be self-justified. At the top of this 'pyramid' (of difficulty to ex post justify) one will find data fabrication, which is universally considered unacceptable. Slightly below one will encounter 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 bottom of the pyramid there will be practices such as continuing to collect more data, if one has obtained a non-significant result, a practice that, in a study by John et al. (2012), more than 50% of psychologists admit they have engaged in.

The evidence indicates that practices that are not considered uniformly morally condemnable by the scientific community might be tolerated and thus more prevalent.[4] John et al. (2012) and Meyer and McMahon (2004) provide survey evidence that less ethically defensible behaviors are self-admitted and observed (as committed by other scientists) less than more defensible behaviors. Fabrication/falsification is typically self-admitted 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

[3]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.

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

et al., 2012).

A second important aspect of the relevant environment we wish to model is its *tournament* character (see e.g. Stephan, 2012). A fundamental characteristic of the scientific profession is that all researchers desire to publish in top journals and attract as many citations as possible, while the scientific community's capacity of attention is limited, with a myriad of research agendas and results competing for attention. Hence, there is limited opportunity and fierce competition for prestigious publications. Since QRP tend to increase the strength 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 loses.[5]

To account for the fact that different forms of QRP differ in their degree of defensibility we will categorize the possible QRP in two tiers, *severe* QRP and *mild* QRP. A key virtue of a theoretical model is parsimony and this simple assumption will indeed suffice to generate a rich level of analysis. To reflect the tournament character of the scientific game of publishing, suppose that there is limited capacity or attention, and only a fixed number of results will be published.[6] Our crucial assumption is that if researchers engage in a more severe and less defensible form of misconduct they gain an advantage in publishing.

This simple environment can be used to some profit to examine a number of key issues in the current discussion about possible policy remedies to a perceived rise in the rate of false positives. The different suggestions for possible reforms of scientific institutions can be divided roughly into three categories:

- Transparency requirements: only research results that are accompanied by a report or checklist detailing various methodological choices are considered for publication.
- Replication requirements: policies that make it likely that published research results will be systematically replicated.
- External controls: monitoring scientific conduct by government agencies, increasing the cost of severe scientific misbehaviour.

Note that these policies well reflect actual practices, either already in use or in conception. Collins and Tabak (2014) provide concrete examples of how transparency policies may be implemented. First, the US National Institute of Health has been using checklists to assess their grant applications. Second, similar policies were enforced by the Nature Publishing Group and the journals of the American Association for the

[5]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."

[6]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.

Advancement of Science, who now ensure that "authors use a checklist to facilitate the verification by editors and reviewers that critical experimental design features have been incorporated into the report, and editors scrutinize the statistical treatment of the studies reported more thoroughly with the help of statisticians".

There are two systematic ways of increasing replication. Firstly, referees and editors may simply impose replications (possibly probabilistically) before publication. The recent proposal of Galiani et al. (2017) describes a possible implementation of pre-publication replication in empirical economics: "... authors would submit their data and code after a conditional acceptance. Journals would then verify that all raw data and code (i.e. sample and variable construction, as well as estimation code) are included and executable. They would then commission research associates perform a push button exercise that verifies that the code executes and reproduces the tables and figures in the article. If the code does not execute or reported results are different, editors could either ask authors to correct their errors or choose to re-review the paper. Finally, for a random sample of papers the journal would attempt to re-construct the code from scratch or search the executable code for errors." Galiani et al. (2017) argue that this system can be both incentive-compatible and low-cost. A second avenue consists in the governing bodies of a discipline (associations, funding bodies, etc.) encouraging structured large-scale replication attempts – e.g., Camerer et al. (2016) and the Open Science Collaboration (2015).

The third policy could be operationalized, for instance, by increasing the penalty for extreme cases of research fraud, such as data fabrication, changing the legal environment and criminal law (Houses of Parliament, 2017). In a similar vein, one could increase the likelihood of detection and punishment, which would inflate an individual's expected cost of fraud. To reliably detect data fabrication original collected data will have to be kept safe for later comparison to the reported data for publication. A neutral organization could be formed to save and keep the original data: for instance, some experimental economics labs keep a data archive of conducted experiments. This would need to be combined with regular 'data audits' that perform the required comparison: Shamoo (2013) describes in detail the principles of such data audits for experimental biomedical research. He also cites the work of Glick (1989), who argues that data audits are a cost-efficient policy.

Modelling the Three Policies

In our environment transparency requirements will increase the cost of 'mild' forms of QRP, but they cannot rule out more severe forms, such as data falsification or fabrication. If implemented rigorously, possibly at high cost, such policies could make mild forms of QRP, such as p-hacking, considerably more expensive, possibly prohibitively so. Increasing the likelihood that research results will be replicated will, possibly at a high cost, increase the cost for an individual researcher to engage in any form of QRP, since all forms can be detected by replication. Targeting and penalising severe forms of QRP through government agencies that monitor scientific conduct will increase the cost of engaging in severe QRP such as data fabrication, increasing both detection probability and penalty if caught, again arguably at a high administrative cost to society at large.

Which of these approaches is likely to decrease the prevalence of QRP most? Our simple theoretical model allows to assess the various policy approaches by eliciting the

changes in equilibrium behavior that are brought about by a change of the rules. The model shows that under a very simple set of assumptions a policy approach that relies on rigorous transparency, ruling out mild forms of misconduct, has an unambiguously beneficial effect. This is because such an approach has a knock-on effect, reducing the prevalence 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 QRP 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 QRP are used.

3 A Simple Model

A simple mathematical model can help clarify ideas and guide future research, although it will necessarily rely on simplifying assumptions. If a given set of assumptions seems questionable, the appropriate response is to derive the model's predictions under alternative assumptions. If the predictions are robust across a range of logical and empirically plausible assumptions, then the model is useful for guiding practice and policy. Ideally, one will then use experiments to test the assumptions of the model. Our model will be able to incorporate "general equilibrium" effects resulting from strategic interaction, meaning that the interdependencies across people and institutions are fully accounted for. This is a key virtue of game theory.

Our main point can be illustrated by using a very simple setup and analyzing the properties of its Nash equilibrium. This means that we examine aggregate behavior – aggregate behavior simply comprises individuals' behaviors – such that each individual's behavior is 'the best possible' according to her payoffs and given the behavior of others. Therefore our analysis should be interpreted as predicting what happens in the medium to long run, when all effects of learning have taken place and behavior has stabilized. This is relevant to assess the type of environments, in 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 represents a Nash Equilibrium (Fudenberg and Levine, 1998). If these conditions are not met, then prediction stemming from a Nash equilibrium will lack plausibility.

Key to policy analysis is to assess 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 QRP. In an extension we model the more realistic but also more complex case that researchers differ in their cost of misconduct.[7] This is equivalent to letting the

[7]In our model this cost is meant to capture many different forces and thus consists of several components, such as a psychological aversion to cheating, the probability of detection and punishment, and the actual effort of modifying the evidence. In the latter sense our setting 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'.

rewards vary, which would capture heterogeneity in research ability – researcher types will capture an aggregate of psychic cost of cheating and research ability.[8] As we will show the model is capable of delivering strong results about the differential effects of different policies, and the results are relatively robust to alternative settings.

3.1 Homogeneous Researchers

Suppose there are two researchers who each have obtained a research result. Both results are of comparable interest to the scientific community, but could be made more surprising and interesting. To alter this state of affairs, each researcher has the opportunity to tune up their work, exerting effort e in presenting the results in a biased way. To keep the analysis easily tractable 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 $\overline{e}$.

Note that option (ii) corresponds to lying by omission and option (iii) to lying by commission. We assume that the possible levels of 'creative effort' $0 < \underline{e} < \overline{e}$ correspond to the cost of performing the respective cosmetic action (a cost which, as we have seen, captures several forces). After exerting the level of effort they prefer, researchers present their results to an interested scientific audience by submitting the results to a journal. Assume there is only one journal with a capacity of one article.[9] The journal decides to publish the result that appears more original and novel, and if indifferent it chooses randomly among articles.

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'. Since more extreme misbehaviour yields an advantage we assume that $P(e, e') = 1$ if $e > e'$, $P(e, e') = 1/2$ if $e = e'$, and $P(e, e') = 0$ otherwise. A researcher has payoff R from publishing his work in the distinguished journal. Of course, R could equally be interpreted as the reward of publishing in the distinguished journal as opposed to publishing in a less distinguished, second-tier journal. Hence, a researcher's expected payoff from exerting creative effort e (given e') net of possible cost is given by:

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

Suppose that $R > \overline{e}$, that is, the reward from successful publication outweighs the cost of cheating. This assumption allows the model to capture the interesting case where

[8] More generally, one might allow for the possibility that the 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.

[9] In Part 4 we show that this assumption can be easily generalized. What is important is that the available capacity is lower than the amount of results produced.

at least some researchers cheat; otherwise no cheating would be the – counterfactual – outcome. Note that an equivalent model could set the reward $R = 1$, but explicitly model a common cost deflator e/k instead, or consider publication probabilities $\underline{p}$ and $\overline{p}$ instead of 0 and 1. In addition, one may argue that higher misbehaviour yields publications with higher rewards. Allowing the reward R to be higher under severe misconduct, for instance because results will become disproportionately more surprising, would not change our analysis qualitatively.

Equilibrium

In our setting two players have three possible actions each, which can be represented by a 3×3 matrix of payoffs. Notice that the game is symmetric, in that possible actions and payoffs have the same form for both players. To identify a Nash equilibrium we first derive each player's best reply function to the other player's action. If player i's opponent j chooses not to cheat, i.e. $e_j = 0$, then if i cheats this yields payoff R with certainty, instead of not cheating, which yields R with probability $1/2$. Since $\overline{e} > \underline{e}$, mild cheating will dominate severe cheating. Mild cheating will also dominate no cheating if the gain from mild cheating net its psychic cost is positive, i.e. $R/2 \geq \underline{e}$. Similar reasoning for the other actions of player j yields the following best reply function $e_i^*(e_j)$ for player i (and player j by symmetry):

$$e_i^*(0) = \begin{cases} 0, & \text{if } R/2 < \underline{e} \\ \underline{e}, & \text{if } R/2 \geq \underline{e} \end{cases}$$

$$e_i^*(\underline{e}) = \begin{cases} \underline{e}, & \text{if } R/2 < \overline{e} - \underline{e} \\ \overline{e}, & \text{if } R/2 \geq \overline{e} - \underline{e} \end{cases}$$

$$e_i^*(\overline{e}) = \begin{cases} 0, & \text{if } R/2 < \overline{e} \\ \overline{e}, & \text{if } R/2 \geq \overline{e} \end{cases}$$

The expressions use the assumption that $\underline{e} < \overline{e} < R$.

These best reply functions then determine the Nash equilibrium, depending on the relative cost of cheating compared with the return to publishing. If $\underline{e} > R/2$, there is a pure strategy Nash equilibrium $(0, 0)$. If $\overline{e} < R/2$ there is a pure strategy Nash equilibrium $(\overline{e}, \overline{e})$. If $R/2 + \underline{e} < \overline{e}$ there is a pure strategy Nash equilibrium $(\underline{e}, \underline{e})$. If $R/2 + \underline{e} > \overline{e} > R/2 > \underline{e}$ then is no pure strategy Nash equilibrium. A Nash equilibrium in mixed strategies is, however, guaranteed by standard arguments. Therefore we turn to examining equilibria in mixed strategies.

Let $\underline{q}$ and $\overline{q}$ denote the probability that player j chooses actions $\underline{e}$ and $\overline{e}$ respectively. The remaining probability mass is associated to no cheating. Then in a Nash equilibrium in mixed strategies researcher i must be indifferent between all actions chosen with positive probability. Hence, if researcher i 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 obtain the mixing probabilities of the opponent, yielding:

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

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 – see the continuum version that follows. The equilibrium seems counter-intuitive in that the equilibrium frequency of each form of QRP depends on the cost of the *other* form of QRP.[10] The reason 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.

It is important to emphasize that a mixed strategy equilibrium does not require strong behavioral assumptions. In particular, there is no 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 interactions have 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 or opponents' payoffs. 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).

The following lemma summarises the different Nash equilibria that may occur.

Lemma 1 (Equilibrium Behavior)**.** *Suppose $\underline{e} \geq 0$. Then the Nash equilibrium of the publication game is:*

(i) $(0, 0)$ *– corresponding to $\underline{q} = \overline{q} = 0$, i.e. no researcher cheats – if $R/2 < \underline{e}$.*

(ii) $(\underline{e}, \underline{e})$ *– corresponding to $\underline{q} = 1 > \overline{q} = 0$, i.e. all researchers engage in mild misconduct – if $\overline{e} > R/2 + \underline{e}$.*

(iii) in mixed strategies – such that both researchers engage in mild misconduct with probability $\underline{q}$, severe misconduct with $\overline{q}$ and do no cheat with $1 - \underline{q} - \overline{q} > 0$ – if $R/2 + \underline{e} > \overline{e} > R/2 > \underline{e}$.

(iv) $(\overline{e}, \overline{e})$ *– corresponding to $\underline{q} = 0 < \overline{q} = 1$, i.e. all researchers engage in severe misconduct – if $\overline{e} < R/2$.*

This lemma follows immediately from an analysis of the best response correspondences in each particular range of parameters. In general, cheating behavior depends on the cost of cheating relative to the return to publishing. Figure 1 depicts the different parameter regions that generate the different equilibria. The equilibrium behavior 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 $\underline{q}$ weakly

[10] This is a well-known feature of games such as 'Cops and Robbers', where the equilibrium rate of patrolling depends on the payoffs of Robbers, not Cops, and similarly the equilibrium rate of robbing depends only on the payoff functions of Cops.

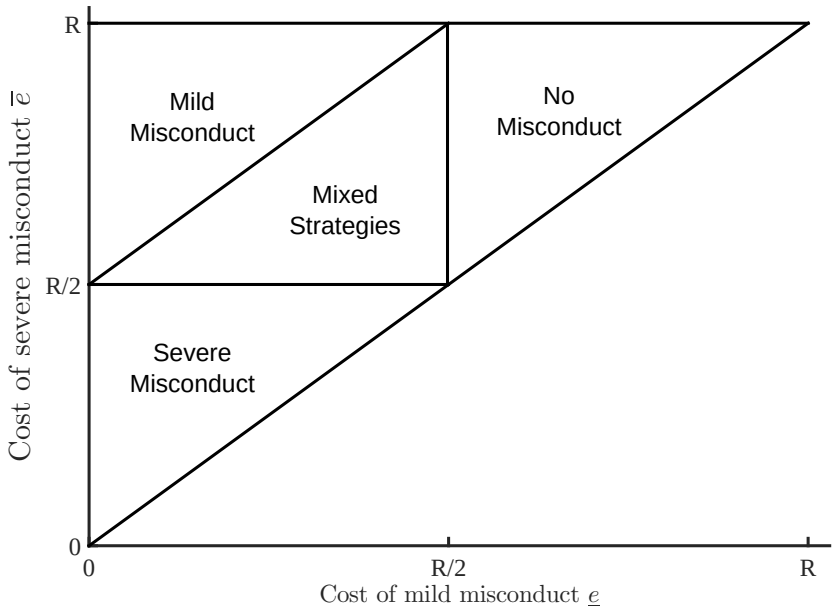

Figure 1: Different equilibria depending on the cost of misconduct.

increases in the cost of severe misbehavior $\overline{e}$, whereas the frequency of severe misbehavior $\overline{q}$ weakly decreases in the cost of mild misbehavior $\underline{e}$, whenever both are played with positive probability. 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 engaging in severe misconduct can be optimal only if one's opponent employs mild QRP.

Policy

In this paper we are interested in the potential policy effects on total aggregate misconduct ($\overline{q} + \underline{q}$). The three policy reform proposals (transparency requirements, replication requirements and external controls) will have differential effects on the cost of mild and severe misbehavior. Transparency requirements will make mild misbehavior more cumbersome, increasing $\underline{e}$, while severe misbehavior such as outright fraud will simply extend to the documentation, leaving $\overline{e}$ virtually unchanged. Conversely, external monitoring by government agency will uncover outright fraud but not mild misconduct, e.g. in the form of ex post choice of sample size. Thus external controls will increase $\overline{e}$ while leaving $\underline{e}$ unchanged. Replication requirements will in principle be able to uncover both types of misconduct, fraud and statistical manipulation, at least if the replication includes an independent re-sampling of data.

This is relevant since changes in the cost of mild and severe misconduct will affect the frequency of misconduct differently. Notice that as the cost of mild misconduct goes up, equilibrium behavior tends to involve less misconduct in aggregate. For instance, starting in a regime that has both researchers engage in mild misconduct with probability one, the equilibrium will first switch to mixed strategies, with a positive probability of no cheating, and eventually to an equilibrium without cheating. If the cost of severe misconduct increases, however, the prevalence of cheating may decrease or increase, depending on the initial conditions.

Equipped with the results derived above we are now in a position to conduct our main exercise: address alternative policies. Using the Lemma it is straightforward to show that given mild misconduct is an equilibrium outcome (i.e. $\underline{e} < R/2 < \overline{e}$), policies that affect either type of misconduct will yield very different outcomes.

Proposition 1 (Increasing the Cost)**.** *Suppose in an equilibrium some player plays mild misconduct with positive probability. Increasing the cost of mild misconduct sufficiently will strictly decrease the equilibrium frequency of misconduct. Increasing the cost of severe misconduct cannot decrease the equilibrium frequency of misconduct.*

By expression (1) it holds that in the parameter range where mixed strategy equilibria occur, the rate of total misconduct $(\overline{q} + \underline{q})$ is decreasing in $\underline{e}$ and increasing in $\overline{e}$. To see that the proposition holds one can use this fact, along with a mere inspection of Figure 1. This result implies that a policy that focuses on the prevention of mild misconduct will dominate a policy that seeks to prevent severe misconduct. Therefore (and assuming that the model is an accurate representation of the environment) the scientific community should support the enforcement of reporting guidelines for empirical articles. Its implementation would diminish the degrees of freedom in reporting empirical results and thus would render lying by omission impossible. Accordingly, researchers would now have to compete having only two available strategies: either be honest or engage in "lying by commission" (pure fraud). Given the great cost of committing pure fraud, all researchers would prefer to report their results honestly in this environment. On the other hand, the research community should not support a policy of criminalizing fraud, as such a policy would be ineffective in reducing the aggregate rate of misconduct. This is because this policy would rule out lying by commission but would leave the potential for lying by omission intact. Given the low cost of the latter strategy, researchers would now turn to it.

Proposition 1 is encouraging, in that it does not depend on the exact parameters of the model used to derive it and thus allows policy conclusions even if the community is uncertain regarding the exact costs of engaging in each form of QRP. As long as mild misbehaviour is perceived to be a problem in the field, a policy that focuses on mild misconduct – such as reporting guidelines – will do no harm and tend to decrease aggregate misconduct, while the opposite is true for policies that focus on severe misbehavior. Later 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 (see Proposition 4). In this more realistic environment a policy of enforcing reporting guidelines still dominates – in terms of minimizing aggregate misconduct – policies that curb severe misconduct.

General Policies

In general, policies may affect both the cost of mild and severe misconduct. Suppose that new costs are given by $\underline{e}' = \underline{e} + \underline{\delta}$ and $\overline{e}' = \overline{e} + \overline{\delta}$. If the policy affects more the cost of mild misconduct, meaning that $\underline{\delta} > \overline{\delta}$ and $\Delta(\overline{e} - \underline{e}) < 0$, this will correspond to a move to the north-east in Figure 3.1, with a slope less than 1. Therefore such a policy will never increase the frequency of misconduct, possibly going from a regime with universal mild or severe misconduct to one with some misconduct of either variety, or to a regime without any misconduct. In contrast, if the policy affects more the cost of severe misconduct, meaning that $\underline{\delta} < \overline{\delta}$ and $\Delta(\overline{e} - \underline{e}) > 0$, this will correspond to

a move to the north-east in Figure 3.1, with a slope greater than 1. The effect is ambiguous and depends on the initial conditions: the frequency of misconduct may decrease, if one starts in a regime with only severe misconduct, or increase if one starts in a regime with mixed strategies.

Consider now the effects of different replication policies: a policy of replication tends to increase the costs of both types of QRP (because of a higher probability of punishment). For instance, replication of experimental procedures generating new data would reveal discrepancies between the published results and the ones of the replication. This would increase the expected cost of both types of QRP, although we are aware of no evidence that could instruct us about the relative magnitude of the increase in these costs. This is why our general analysis above teaches us a few important lessons. First, even replication policies that successfully increase the cost of every type of misbehaviour may backfire, depending on the relative increase in costs and the prevailing conditions. Second, polices that place greater emphasis on preventing mild misconduct appear to be more appropriate to reduce the frequency of misconduct in general.

Competitive Pressure

As the field gets more crowded in most disciplines of science, one might expect the reward from publication R to increase over time as well. This has some interesting effects on misconduct in a mixed strategy equilibrium as the following proposition states. It follows immediately from (1).

Proposition 2 (Increasing the Reward)**.** *Suppose both mild and severe misconduct are played with positive probability in an equilibrium, and increase the reward for publication from R to R' with $R' < 2\overline{e}$. This yields more severe misbehavior ($\overline{q}' > \overline{q}$) and less mild misbehavior ($\underline{q}' < \underline{q}$). Overall misconduct goes down: $\overline{q}' + \underline{q}' < \overline{q} + \underline{q}$.*

That is, increasing the reward of publication (e.g. in form of career concerns or importance for tenure) crowds out mild misbehavior by making severe misbehavior more attractive, which reduces the aggregate prevalence of QRP. Whether the overall effect is desirable depends on how detrimental each type of misconduct is for the accumulation of knowledge. The reduction in overall misconduct is perhaps a little surprising in light of the frequent concern that an increase in the publish-or-perish culture is likely to lead to more biased research across the board. However, notice that if R increases by enough, so that $2\overline{e} \leq R'$, then severe misbehaviour becomes the dominant strategy (see Figure 3.1 where an increase in R is similar to a decrease of both $\underline{e}$ and $\overline{e}$). Hence, one may expect that an increasingly competitive "publish or perish" culture in the academia will not initially lead to more aggregate misbehaviour, but will do so as the competitive pressure increases further over time.

We now turn to the possible effects of alternative policies for different levels of rewards. For simplicity, we only consider policies that prohibitively increase the cost of some form of misbehaviour, (i.e. either mild or severe). Figure 2 illustrates the total level of cheating for various levels of rewards R when (in the absence of any policy) $\underline{e} = 1$ and $\overline{e} = 3$. It is clear that the policy of ruling out mild misbehaviour results in a lower rate of total misconduct compared either with the absence of intervention or the policy of eliminating 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.

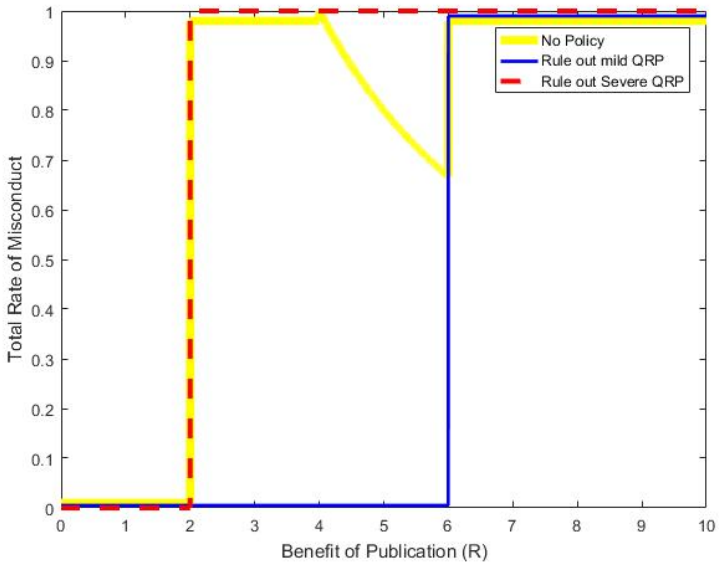

Figure 2: Frequency of total misconduct for the three alternative policies for various levels of rewards when $\underline{e} = 1$ and $\overline{e} = 3$.

4 Continuum of Researcher Types

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 (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.[11]

As above researchers' rewards are given by the expected quality of the outlet where 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

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

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 $\theta^*(e, e')$ for binary comparison of action choices, such that agents of types $\theta < \theta^*(.)$ prefer the option involving the more severe form of

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

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 $\overline{q}$. In a Nash equilibrium each researcher must choose an optimal action given the actions of other researchers, that is given measures $\underline{q}$ and $\overline{q}$. The following proposition characterises the possible equilibria for plausible ranges of the parameter κ, which can be interpreted as the acceptance rate for publication in top journals. The proof can be found in the Appendix.

Proposition 3. *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.*

These equilibrium regimes reflect the ones in Lemma 1: one regime where severe misconduct is a dominant strategy, when its cost is low, one where both forms of misconduct have positive frequency, when the cost difference is sufficiently low, and one where mild misconduct is a dominant strategy, when its cost is low. Since θ is distributed on $[0, 1]$ there will always be some researcher whose cost of committing misconduct is low enough to engage in some form of it, so that no misconduct cannot be a dominant strategy for all types anymore.

Notice that the different equilibrium regimes depend on the cost of mild and severe misconduct, $\underline{e}$ and $\overline{e}$. As in the case of homogeneous researchers, a policy that increases only $\underline{e}$ thus makes the regime with only mild misconduct more likely and decreases the frequency of misconduct. A policy that increases only $\overline{e}$ may lead to a shift from a regime with only severe misconduct to one with both mild and severe misconduct, which entails an increase in the frequency of misconduct. To derive a precise analysis of the likely policy effects on misconduct we will focus on policies that either make mild or severe cheating prohibitively costly to researchers, as in the analysis of Figure 2 above.

Policy A: Preventing $\overline{e}$

Equipped with the properties of the equilibrium absent any policy we turn first to a policy that prevents severe misconduct. Suppose that the cost for engaging in severe

misconduct $\overline{e}$ is prohibitively high, so that $\overline{q} = 0$, but $\underline{e}$ remains unchanged (existence of such an $\overline{e}$ is ensured). 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

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

where F_θ denotes the cumulative density function of θ and 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

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

where the assumption of a uniform distribution has been maintained. Analogously to Policy A, equilibrium measures are $\overline{q} = \sqrt{\kappa R/\overline{e}}$ if $\kappa \leq R/\overline{e}$, $\overline{q} = 1/2 - \sqrt{1/4 - (1-\kappa)R/\overline{e}}$ 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. The proof of the following proposition can be found in the appendix.

Proposition 4. *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.*

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 ($\underline{q}^A = .4082$), while preventing $\underline{e}$ results in a decrease of 53.6% in the frequency of cheating: ($\overline{q}^B = .2113$).

Examining the effect of the two policies for a range of possible values and illustrating it graphically can be very informative. Figure 3 shows the incidence of cheating for

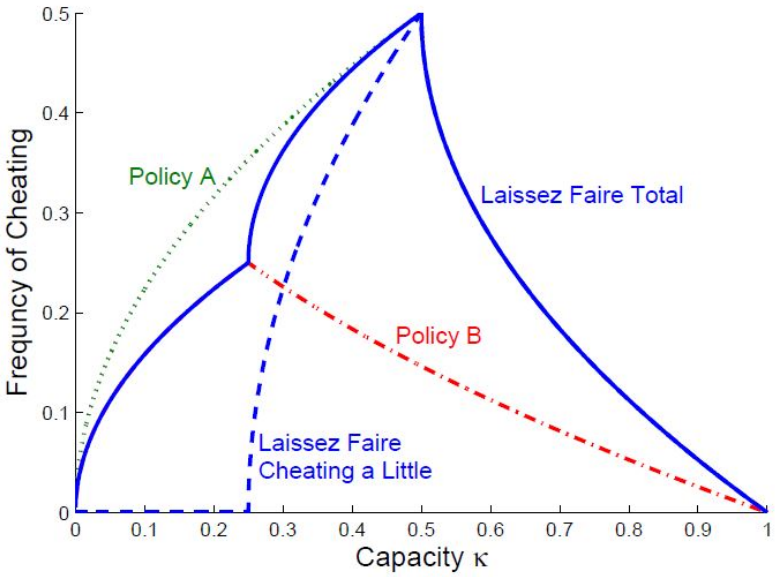

Figure 3: Frequency of cheating for the different policies

varying capacity κ under the three different policies, keeping $R = 1$, $\underline{e} = 2$ and $\overline{e} = 4$ (in the absence of any policy). 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 3, Policy B leads to a weakly 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 than laissez faire, in terms of the overall rate of misconduct, for small values of κ. Once more, our main result appears robust: a regime of increased transparency that prohibitively increases the costs of lying by omission will result in weakly less overall misconduct.

5 Discussion

The extent of the recent public debate about QRP, further stimulated by a number of high profile incidents, necessitates assessing 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 might be useful 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[12] and explicit audits by an outside body are very costly and

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

highly unpopular among scientists.[13] Our model indicates that such extreme measures may not be necessary, at least if researchers face a high cost of engaging in these 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. Indeed this reasoning appears reminiscent of "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 emphasized that our study focuses on the overall rate of misconduct, treating the two forms of QRP as similarly negative. 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 via a non-random sample. If one believes that a typical case of severe QRP distorts the evidence much more than a typical case of mild QRP, then the analysis would have to chance. In this case, in order to generate valid insights we would need to consider a value function weighting each form of misconduct differently, and possibly in a non-linear way. More research is needed.

Our model could also provide insights regarding other policies targeting false positives than the ones discussed here. For instance, tournament logic would suggest that reducing competition by allowing more papers to be published – possibly subject to refereeing to guarantee methodological integrity – will reduce the perceived need to resort to QRP. This argument relies on assuming that a researcher's career concerns only depend on the number of own publications. However, actual research careers seem to depend more on the impact of the research results, i.e. on the attention they generate. Therefore, if attention of one's fellow researchers and the public is limited, there will always be intense competition for it, and enhancing results through the use of QRP is likely to increase impact measured by citations.

It is important to reflect on the implications of our results for different disciplines and types of empirical research. Our model is more applicable to research that generates and reports analysis on original data (i.e. reporting new experimental or survey evidence). It is less applicable to analysis that works on secondary data. The reason is that although in the latter type of research outright fraud is possible, it is less relevant since changing the actual data is impossible. Accordingly, our sharp categorization of types of QRP is less appropriate for secondary data analysis. Furthermore, mild QRP are less costly in new fields of research, where analytical practices are less mature. In mature fields, low flexibility entails less scope for mild misbehaviour, since there are clear codes on reporting practices. Hence, our categorization of QRP is likely to be

most prominent cases in psychologys 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."

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

less relevant. In summary, our results are likely to be more applicable for new scientific fields and for research that generates new data.

As for 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 QRP 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 QRP will not do harm. Moreover, as we pointed out already, some key insights of our model are also obtained by more complex models that explicitly study the behaviour of replicating and verifying researchers – e.g. Lacetera and Zirulia (2009). The general insights thus seem robust to alternative specifications. Finally, 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.

Overall, based on our arguably specific setting, we could suggest to the research community as well as to the UK government and other bodies that promoting checklists for complete reporting seems like a promising policy. This advice, however, is based only on a prediction of possible consequences on the extent of questionable behaviour. The costs of these policies also need to be carefully calculated before implementation. In fact, it seems to us that transparency requirements are the least costly policy: the effort to comply will be noticeable but small in comparison with full blown audits, and will decrease over time as researchers and referees adjust their research practice in line with the transparency 'checklist'. By contrast, policies that rely on audits of scientific data will require setting up new institutions to implement these checks at considerable cost. To this one may need to add an indirect, psychological cost on the researchers' morale: the lack of trust implied by such a policy and fear of intervention by non-scientific bodies that are not necessarily familiar with scientific methods may seriously erode intrinsic motivation.

We conclude by emphasizing the importance of rigorous theory in assessing research policies. We have seen a simple example of how formal theory can reach useful counter-intuitive conclusions. An attempt to eradicate a given form of QRP will generally affect incentives to pursue other forms of misconduct. That is, simple policies may have unintended side effects and game theoretic modeling provides a means to examine effects of policies in complex patterns of behavioral interactions. In general, the overall effect of under-theorized factors, such as the trade-offs which researchers face and the general interdependencies among individuals in the publication system, may well be the opposite of what is desired by the proponents of a given proposal.

The study of factors that affect the credibility of research has evolved into a scientific discipline of its own – i.e. meta-research, see Ioannidis et al. (2015) – and there is a clear mandate to study possible solutions to ameliorate the situation. Despite the major attention the credibility problem has attracted, there is a dearth of rigorous evidence and analysis, which is needed to adequately evaluate the numerous proposals for reform. In a review of how economics tools can play a potentially important role in assessing reforms of practices in biomedical research, Gall et al. (2017) use the results of the basic version of our model – after summarizing them – as an illustration of how basic economic modelling can inform the evaluation of reform proposals in science. As 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."

References

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

Abel Brodeur, Mathias Lé, Marc Sangnier, and Yanos Zylberberg. Star wars: The empirics strike back. *American Economic Journal: Applied Economics*, 8(1):1–32, 2016.

Open Science Collaboration et al. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, 2015.

Francis S Collins and Lawrence A Tabak. Nih plans to enhance reproducibility. *Nature*, 505(7485):612, 2014.

Glenn Ellison. Evolving standards for academic publishing: A q-r theory. *Journal of Political Economy*, 110(5):994–1034, 2002.

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.

Sebastian Galiani, Paul Gertler, and Mauricio Romero. Incentives for replication in economics. Technical report, National Bureau of Economic Research, 2017.

Thomas Gall, John PA Ioannidis, and Zacharias Maniadis. The credibility crisis in research: Can economics tools help? *PLOS Biology*, 2017.

J Leslie Glick. On the potential cost effectiveness of scientific audits. *Accountability in Research*, 1(1):77–83, 1989.

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.

Parliamentary Office of Science & Technology. Houses of Parliament. Integrity in research. POSTnote 544, January 2017. URL `http://researchbriefings.files.parliament.uk/documents/POST-PN-0544/POST-PN-0544.pdf`.

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.

John PA Ioannidis, Daniele Fanelli, Debbie Drake Dunne, and Steven N Goodman. Meta-research: evaluation and improvement of research methods and practices. *PLoS Biol*, 13(10):e1002264, 2015.

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.

Bralind Kiri, Nicola Lacetera, and Lorenzo Zirulia. Above a swamp: A theory of high-quality scientific production. Technical report, National Bureau of Economic Research, 2015.

Nicola Lacetera and Lorenzo Zirulia. The economics of scientific misconduct. *The Journal of Law, Economics, & Organization*, 27(3):568–603, 2009.

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.

Zacharias Maniadis, Fabio Tufano, and John A List. How to make experimental economics research more reproducible: Lessons from other disciplines and a new proposal. In *Replication in experimental economics*, pages 215–230. Emerald Group Publishing Limited, 2015.

Marcia McNutt. Journals unite for reproducibility. *Science*, 346(6210):679–679, 2014.

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.

David Moher, Alessandro Liberati, Jennifer Tetzlaff, Douglas G Altman, Prisma Group, et al. Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *PLoS med*, 6(7):e1000097, 2009.

Brian A Nosek and Daniël Lakens. Registered reports, 2014.

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.

Kenneth F Schulz, Douglas G Altman, and David Moher. Consort 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMC medicine*, 8 (1):18, 2010.

Adil E Shamoo. Data audit as a way to prevent/contain misconduct. *Accountability in research*, 20(5-6):369–379, 2013.

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.

Paula Stephan. *How economics shapes science*. Harvard University Press, 2012.

EricJan Wagenmakers, Ruud Wetzels, Denny Borsboom, and Han L. J. van der Maas. Why psychologists must change the way they analyze their data: The case of psi: Comment on bem (2011). *Journal of Personality and Social Psychology*, 100(3): 426–432, 2011.

A Appendix: Proofs

Proof of Proposition 3

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 $\overline{q} < \kappa < \underline{q} + \overline{q}$).

Proof of Proposition 4

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 $\overline{q}^B < \underline{q}^* + \overline{q}^*$.