P-curve: the key we were looking for?

Assessing benefits and shortcomings of using the p-curve to combat scientific fraud

Master Thesis

Erasmus School of Economics – Department of Economics

19 September 2013

Name: Joost van Gemeren

Student number: 332940

Supervisor: Prof. Dr. A.J. Dur

Displaying statistically significant results can improve the odds of getting a paper published in an academic journal. This induces some researchers to change insignificant results into significant by p-hacking. Simonsohn et al. proposed to use the p-curve to combat this form of fraud. This paper uses economic-theoretical analysis to assess the benefits and limitations of the p-curve in combating fraud in science and promoting social welfare. Its results show that the p-curve induces honest reporting by mildly fraudulent researchers, but it cannot correct the behavior of the heavy fraudsters. The p-curve also induces some fraudsters to switch to other occupations. One should keep in mind that these effects are bounded, because p-hackers have the possibility to switch to other forms of fraud which are harder to detect with the p-curve. With this taken into consideration, the p-curve increases social welfare if fraud is costly enough to society and if the cost to replace leaving scientists is not too large.

JEL-classification: A19, C10, D61, H89, J33, J38, J45, K42

Key words: P-hacking, p-curve, fraud in science, data fabrication, misreporting, publication bias, statistical significance, ethical integrity

16,076 words

1. Introduction

Academia is an industry that produces a public good that is deemed to be one of the main drivers of long-run economic growth: knowledge. Knowledge drives innovation (e.g. Anselin *et al.* 1996) which again creates long-run growth through creative destruction (Schumpeter 1950, Aghion and Howitt 1990). The production of knowledge by the academic world is, however, hampered by moral hazard problems that arise because effort by academics is innately difficult to monitor (Dasgupta and David 1987, Dasgupta 1989). The monitoring issue opens up opportunities for researchers to shirk, which harms academic institutions as well as the society at large.

Designing an incentive compatible reward structure to avoid costly shirking therefore seems appropriate. Such a reward structure requires payments or other desirable benefits to be tied to indicators that measure effort with as little noise as possible (e.g. Lazear and Gibbs 2009). In academics, publications in journals is one of those indicators. For instance, promotions are based on achieving a target number of academic publications (McGrail *et al.* 2006), sometimes in an up-or-out system (Waldman 1990). Other incentive forms are granting monetary payments (Graves *et al.* 1982) and research budgets (McGrail *et al.* 2006) based on academic publications. Intangible benefits, such as a higher academic ranking (RePEc 2013) or more influence on policy making (evidence-based policy, see Nutley *et al.* 2000) can also be a result of past publications. Using journal publications as basis for incentives is conceivable, since academic journals want to publish only high quality papers. If researchers want to get their reward, they therefore will have to exert research effort that results in papers of sufficient quality. Another useful characteristic of these incentive schemes is that they work as a screening device to draw highly able researchers, who are capable of producing more high quality papers than less able ones, into academia.[1]

The decision by academic journals whether to publish a paper does, however, involve some noise. Since at least the late sixties, “a commonly admitted tendency” (Sterling 1959, pp. 33) of journals is to publish a disproportionately large share of papers showing significant results (e.g. Sterling 1959, Easterbrook *et al.* 1991, Ashenfelter *et al.* 1999, Rothstein *et al.* 2005, Dwan *et al.* 2008), a phenomenon often called *publication bias*.

[1] An overview of the other aspects of the reward structure in academics and the rationale for these aspects can be found in Stephan (1996).

Editors', reviewers' and readers' preference for significance is often thought to be a cause of publication bias[2]. These people may dislike studies showing non-significant results, because insignificance could arise from poor research design (Dickersin 1990), too small samples (Freiman *et al.* 1978) or testing of implausible hypotheses (Angell 1989), thus making significance a signal of high research quality. Non-significant results are also less interesting from a policy perspective (Dickersin 1990). Another explanation of the phenomenon is that editors and readers may suffer from confirmation bias, in the way that referees do (Mahoney 1977). If so, they unwittingly favor evidence that is in line with their theoretical premise. For instance, stating a hypothesis that expects the existence of a relationship creates such a premise (Nickerson 1998, a literature review on confirmation bias). Then, finding evidence against it can reduce the odds of publication.

Because they are rewarded for it, researchers focus on significance as well[3], firstly by pointing their research effort at papers that (have a good chance to) find significant results. Non-significant results are metaphorically *file drawered*[4]: thrown away with the justification of being useless (e.g. Rosenthal and Gaito 1963, Rosenthal 1979), with the result that meta-analyses consistently reveal biased effects (e.g. Scargle 2000)[5]. The second response also yields biased effects, albeit more intentionally. Academics have been proven to manipulate insignificant results in such a way that they artificially become significant (Kohn 1986, Simmons *et al.* 2011, Brodeur *et al.* 2013), in order to increase the chance of publication.[6]

Fanelli (2009) proves that fraud is an important issue: two percent of scientists admit to have falsified research and 34 percent admit to have adhered to other "questionable research practices". John *et al.* (2012) found a "surprisingly high" incidence of fraud among

[2] This is also believed to be a reason why effect sizes are overstated, see Ioannidis (2008).

[3] A way to restore proper incentives could be to reward insignificant findings that are published in an academic journal more than significant findings.

[4] Some researchers prefer to call file drawering *selection* or *selection bias*. In order to avoid a mix up with sample selection bias (Heckman 1979), I will use the term file drawering.

[5] Several statistical techniques have been developed to correct for this bias. Rosenthal (1979) was the first to introduce the *fail-safe N* (elaborated upon further by Orwin 1983). Rosenthal shows how many file drawered studies are required to turn the average effect found in meta-analysis non-significant. If this number is very large, one can trust the direction (not the effect size) of the found effect. Begg and Mazumdar (1994) have proposed a rank correlation test to formally show whether or not there is publication bias in a given meta-analysis. Egger *et al.* (1997) and Duval and Tweedie (2000) propose such a test based on funnel plots. Hedges (1992) and Brodeur *et al.* (2013) explicitly model file drawering in order to correct for it. Ioannidis and Trikalinos (2007) is another test based on comparing the expected number of significant findings, derived from statistical power, with the actual number of significant findings. Thornton and Lee (2000), Rothstein *et al.* (2005) and Weiss and Wagner (2011) discuss several other methods and techniques in a detailed way.

[6] Other negative side effects of the focus on publications can be found for instance in Angell (1986), Feigenbaum and Levy (1993) and De Rond and Miller (2005).

psychologists. Given that not all fraudulent researchers admit their sins, the results shown in these papers are likely to be an underestimation of the real incidence of fraud in science.

Fraud-prone researchers have several options to choose from. One is the practice called *data fabrication*, which basically means changing or making up data in order to present significant results. Recently in the Netherlands, Diederik Stapel (Abma 2013) and Dirk Smeesters (Keulemans 2012) were caught doing this. A more subtle form of fraud is exploiting in principle legal flexibility ("researcher degrees of freedom") to present insignificant results as significant, a practice Simonsohn *et al.* (forthcoming) named *p-hacking*[7]. For instance, p-hacking researchers recode variables into new categories, strategically select control variables and use fixed effects regression rather than OLS, in order to find an econometric specification that indicates a significant relationship. Other forms of p-hacking are deleting outliers or adding extra observations again and again, until the results can be presented as significant (*data peeking*). The difference between p-hacking and data fabrication is thus the nature of the manipulation: data fabrication is inherently "bad", while the manipulations in the toolkit of p-hackers are only to be considered a violation if they are used to strategically affect results.

The psychologists Simmons *et al.* (2011) show that using all forms of p-hacking in combination can result in a Type I-error probability of 60.7 percent when a significance level of five percent is used. Thus, in this extreme case and when a relationship in reality is non-existent, p-hackers are more likely to draw false conclusions than correct ones, a result also found by Ioannidis (2005). This is problematic, since government and business officials rely on the accuracy of scientists' conclusions in the policy making process. If science loses its trustworthiness[8], policy makers might base their decisions more on intuition and possibly self-interest, with society bearing the consequences.

Aware of this problem, Simonsohn *et al.* (forthcoming) designed a technique to unmask p-hackers. The so-called *p-curve* is the frequency distribution of significant p-values. They conclude that a researcher whose p-curve is skewed to the left must have been "intensely p-hacking". The economic-theoretical analyses in this paper are dedicated to assessing the benefits and shortcomings of the p-curve in advancing honesty in empirical social science and promoting social welfare. In general, scientists rarely use theoretical models to describe fraud,

[7] The name for the phenomenon p-hacking comes from Simonsohn *et al.* (forthcoming), but the used definition for it originally comes from Simmons *et al.* (2011).

[8] See Ioannidis (2012) for a discussion on how the trustworthiness in science is evolving.

publication bias and its consequences.[9] And since the idea of the p-curve is only just born, this paper is the first in its kind, at least as far as I am aware.

This paper brings forward a model where confirmation bias is the driver of editors' preference for significant results. Because people with higher research ability are assumed to do research about less well-established hypotheses, editors are less biased in favor of significance for the highly able. This induces the highly able to commit less fraud, meaning both p-hacking and data fabricating. More ethical people also commit less fraud, which is intuitive. By assumption, people with higher research ability are more likely to become scientist. Also, less ethical people are more likely to participate, because committing fraud yields unethical people extra utility, a mechanism not available to ethical people. This results in the prediction that ethical people that choose to be researcher must be highly able, otherwise they would not participate in the first place. In this way, showing good ethics can be a signal of high ability.

Introducing the p-curve into the model, modeled by an increase in the chance of apprehension upon p-hacking, affects only the utility of unethical and unable people, who are willing to p-hack. In general, the p-curve decreases the expected utility of being a fraudulent researcher. Therefore, some fraudsters choose to become an honest researcher instead, which is called the incentive effect. Others are induced by the p-curve to leave science: the selection effect. The incentive effect is found effectively stop only the most able and most ethical fraudsters from p-hacking. The p-curve will, however, have no such beneficial incentive effects for the very unethical and very unable, because p-hacking is just too attractive for them, even if the odds of apprehension increase. The selection effect dictates that the least able and most ethical fraudulent researchers leave science. The least able leave because p-hacking initially yielded them just enough extra utility to compensate for their low odds of publication. After introduction of the p-curve, this extra utility no longer suffices. The most ethical fraudsters leave, since introducing the p-curve adds to their already high intrinsic costs of p-hacking. Graphical analysis shows that the selection effect reduces the effectiveness of showing off good ethics in signaling ability.

When the p-curve induces a large increase in the chance to apprehend p-hackers, nobody will p-hack anymore. Instead, fraudsters choose to data fabricate. Therefore, the p-curve alone cannot solve the fraud issue. Other methods aimed at combating data fabrication are required

[9] The only exception to this rule I found was Henry (2009). He proves that it might be optimal if disclosure of research methods is not mandatory. The reason is that researchers will then have to do more research in order to convince their readers that the results presented are real, which could bring society more valuable knowledge.

in combination to it. The p-curve does improve the current situation, if society considers fraud to be costly enough and if the adjustment cost incurred attached to introducing the p-curve is not too large. Included in this adjustment cost are the productivity (or destructiveness) of leaving researchers in their best alternative occupation, the additional wage costs incurred to attract replacements to fill in the vacancies and the productivity of these replacements as a researcher.

This paper proceeds as follows. Section 2 reviews the related literature. In section 3, the workings of the p-curve are discussed in detail. Afterwards, the model is presented, which is solved in section 5. Then, the model will be used to assess the effects of the p-curve. Section 7 discusses several other applications and extensions to the model. The final section concludes.

2. Related Literature

The first part of this review will focus on the economics of crime literature, which has important lessons about fraud in science as well. Since many authors have already reviewed the crime literature, this section is basically a review of literature reviews. After that, the more directly related papers will be discussed: those that put up ideas and methods to combat scientific fraud.

A. *Economics of crime*

1992 Nobel laureate Gary Becker (1968) was the first to introduce crime as an outcome of rational economic decision making in general. He showed that increasing the crime penalty or chance of apprehension stops people from committing crime, a statement now known as the *deterrence hypothesis*. Since the seminal paper of Becker, economics more and more became a tool to solve crime and security issues, as it is in this paper. For an thorough overview on the early law enforcement literature, which emphasizes why and in what circumstances the deterrence hypothesis might be false, see Cameron (1988). One decade after that, the focus is still on refuting the old hypothesis, as noted by Garoupa’s (1997) survey. The author attempts to augment the original Becker model by incorporating the criticisms posed by the authors he cited. Polinsky and Shavell (1999) undertake a similar endeavor. A more textbook style review[10] can be found in Freeman (1999), focusing on empirical evidence from the United

[10] Gottfredson and Hirschi (1990) is a useful standard work for readers also interested in other approaches to crime than the economic one. The other discussed approaches come from criminology, sociology, psychology, cultural anthropology and biology.

States. The more recent papers[11], assessed in Levitt and Miles' (2006) literature review, in particular use (applied) econometrics to identify the causal effects of crime policies. Both Freeman and Levitt and Miles in general find evidence in favor of the deterrence hypothesis. Dills *et al.* (2008), who review the literature in the forty year period after Becker, draw an opposite conclusion. They pose that economists still know little about the empirically relevant determinants of crime, because short-term local "evidence" in favor of the existence of an effect of crime policy, is inconsistent with long-term cross-country data. Either way, as Freeman (1999) puts it, "there is still a lot more to do and learn" (pp. 3563) in the economics of crime. This paper aims to do just that, albeit in a less general setting: fraud in science.

B. Fraud in science

The ideal situation to combat fraud in science would be when everybody has full, costless access to every data point and analysis that every researcher has ever used (Easterbrook *et al.* 1991), or when we already know the true effects to be tested with certainty (Ioannidis 2005). Obviously, both are utopian scenarios. Therefore, science must rely on other, second-best methods to combat fraud. These methods all focus at "restructuring incentives and practices to promote truth over publishability", as the title of the paper by Nosek *et al.* (2012) calls it. Their paper is a very thorough literature review on some, often heard practical solutions to the fraud issue, and it discusses the benefits and limitations of each in turn.[12] This paper's literature review takes more of a helicopter view on the three broad categories of fraud solutions: replication and disclosure, reducing the importance of statistical significance and using formal fraud tests.

Having science control itself is arguably the oldest (and most controversial) solution to the fraud problem. As put forward by Merton (1942), science was thought to be able to eventually ban all errors out of the system itself, by scientists verifying or falsifying each other's claims. A seminal paper in favor of this replication as a means to avoid questionable statistical inference is Cohen (1994). He argues that, in order to make analyses transparent and comparable so that it is ready for replication, science should move towards standardization in measurement. Lykken (1968), also a proponent of replication, identifies three ways of doing this replication: replicating the research completely, replicating only the main parts, and

[11] Another recent strand of economic literature in the enforcement of law is that of behavioral economics. After surveying the literature, Garoupa (2003) argues that behavioral economics of crime is effective at exposing some of the weaknesses of the traditional crime theory. Still, behavioral theory does not surpass the traditional economic approach as the most effective theory to approach law enforcement.

[12] Another useful reference that discusses proposals is Weiss and Wagner (2011).

replicating the research in a different setting. The first is the most useful to detect fraud, but the problem is that "[e]xperiments that are literal replications of previously published experiments are very seldom published" (Nickerson 2000, pp. 283). Perhaps this is because in science, all credit goes to the first one that publishes evidence about a relationship (Stephan 1996). Because one cannot easily replicate results directly, science has to rely on indirect replications to assess fraud. See Ioannidis (2012) for an overview of the other impediments to scientific replication. Despite its limitations, some present-day authors still favor replication, for instance Nickerson (2000), Ioannidis (2008), Roediger (2012) and Simonsohn (2013).

Obviously, replication is useless without transparency on data and methods used. Openness also makes it easier to detect fraud without replication, because fraudulent researchers cannot hide behind the curtain of opaqueness (Nosek *et al.* 2012). A way to achieve transparency is using standardized criteria of disclosure that researchers must adhere to. Simmons *et al.* (2011) design six requirements for researchers and four guidelines for reviewers that are meant to reveal and thus prevent p-hacking. Another such checklist are that of CONSORT for randomized controlled experiments (Ioannidis *et al.* 2004), STROBE for observational studies in epidemiology (Von Elm *et al.* 2007) and STARD for diagnostic research (Bossuyt *et al.* 2003). Brodeur *et al.* (2013) put forward a standardized disclosure of data collection and reporting. Simonsohn *et al.* (forthcoming) also advocate disclosing certain information in papers that use the p-curve empirically. Other papers that argue in favor of disclosure are Ioannidis (2005), Ioannodis (2008), Simonsohn (2013) and many more described in Nosek *et al.* (2012). As discussed in footnote 9, Henry (2009) has argued against more disclosure. Other criticisms are that disclosure is costly and need not be truthful (e.g. Jovanovic 1982).

Reducing the focus on significance tests is also a solution that dates back a few decades, and it is also not uncontroversial. In the sixties, psychologists have argued that journals should allow the researcher to use other techniques than significance testing, such as confidence intervals (Rozeboom 1960), to reduce the reliance by journals on p-values as criterion for acceptance (Bakan 1966) and to base the publication decision more on (subjective) evaluations of research quality (Lykken 1968). The seminal paper by Carver (1978) even argues to completely remove the "corrupt" significance testing from science, and replace it by effect size measures. Cohen (1994) favors using more graphical analyses and effect sizes, instead of hypothesis tests. More recent authors disagree with Rozeboom, Carver and Cohen, since graphs and confidence intervals still incorporate the same information: is the effect larger than zero or not? These authors do still follow Bakan and Lykken by arguing in favor of publishing non-significant papers, see for instance Easterbrook *et al.* (1991), Ioannidis (2005), Ioannidis (2008), Young *et al.* (2008), Gladbury and Allison (2012) and Brodeur *et al.* (2013). According to Nosek *et al.* (2012), however, this desire will likely remain a desire, since it is not incentive compatible (yet) to reduce the preference for significant results.

The last solution discussed is using statistical techniques to uncover fraud. Obviously, the p-curve is one of those, which will be discussed in the next section thoroughly. There are not yet many formal tests for fraud in science, because fraud is hard to distinguish empirically from file drawering (e.g. Ioannidis and Trikalinos 2007). The tests reported in footnote 5 mostly are tests that pool these two together into "publication bias tests". This paragraph focuses on those methods that exclude file drawering as alternative explanation. Al-Marzuki *et al.* (2005) test for fraud simply by using descriptive statistics of randomized controlled experiments. By definition of randomness, group means and variances of control variables cannot differ too much. If they do, the samples may have been changed.[13] On the other hand, randomness does not allow means and standard deviations to be too similar across treatments either (Simonsohn 2013). Toedter (2009) shows how to use Benford's law in detecting fraud, a method also used to reveal accounting fraud for instance. The so-called Benford distribution indicates how often a certain digit should appear as first or second number in regression coefficients and their standard errors. If researchers have manipulated the data, the regression results will no longer be random, and thus the numbers will no longer follow the Benford distribution. The author concludes that the first numbers of at that time recent econometric papers indeed violate Benford's law, but the second numbers do not. Gladbury and Allison (2012) propose a methodology to detect p-hacking that relies on p-values that are just insignificant (between 0.05 and 0.1). The intuition is that p-values closer to 0.05 are easier to turn significant than those closer to 0.1. Therefore, if some p-values between 0.05 and 0.075 are "missing", there is evidence of p-hacking. The method is yet to be applied. Brodeur *et al.* (2013) explicitly model file drawering, in order to correct the distribution of p-values for its influence. After this correction, if there are still "too many" p-values in the significant region, there must have been fraud (they call it *inflation*). The researchers conclude that there are indeed too many p-values just below 0.05 in four main economic journals. However, this methodology is not yet applied to individual researchers' p-values.

[13] Another possible explanation is failure of randomization.

3. The P-curve

The p-curve has multiple uses, but this paper will highlight only its applicability in revealing p-hacking. Interested readers can consult Simonsohn *et al.* (forthcoming) for other uses of and further details about the p-curve.

The p-curve is essentially the density plot of (a selection of) the significant p-values that a researcher has published. So, only p-values lower than or equal to 0.05 are examined. Simonsohn *et al.* (forthcoming) show that it is statistically highly unlikely that the p-curve of an honest researcher shows a left-skewed pattern. To see why, let us hypothesize about the shape of the p-curve of an honest researcher who, without knowing it, only publishes about relationships that in reality are non-existent. Suppose a statistical test in one of his researches yields p-value 0.05. How likely is it that this or an even lower p-value is obtained, knowing that the null hypothesis is in fact true? The p-value measures this probability by definition, so the answer is 0.05. In a similar fashion, the chance to obtain a p-value of at most 0.04 equals 0.04, etcetera. From this, one can infer that our researcher ex ante has a chance of $0.05 - 0.04 = 0.01$ to obtain a p-value between 0.05 and 0.04. Likewise, a p-value between 0.04 and 0.03 is obtained with probability $0.04 - 0.03 = 0.01$, and so forth. If these probabilities were shown in a density plot, a uniform distribution is the result. With enough p-values included, the expected p-curve of our hypothetical researcher will thus be a horizontal line.

But what if that researcher was p-hacking so that some of his insignificant results change into significant? Suppose he observes an insignificant p-value before p-hacking, say $p = 0.06$. If he decides to p-hack, the resulting new p-value will not be a fresh draw from the probability distribution, because both p-values result from similar data and techniques. The new p-value is correlated to the initial p-value, and therefore it is likely that the new p-value is still close to 0.06. Then, if it is assumed that dishonest researchers stop p-hacking once they hit significance, the new p-value is more likely to be near 0.05 than 0.01. P-hacking then results in a left-skewed distribution[14], because p-hacked results are located in the right-end of the p-curve, while the honest results are spread evenly over the curve. Obviously, a left-skewed p-

[14] The formal proof of this result can be found in Simonsohn *et al.* (2013).

curve can arise out of back luck as well. Therefore, the authors propose to test for left-skewness formally using a χ^2 test[15].

The p-curve of an honest researcher who does test relationships that are true in reality, looks somewhat different. Wallis (1942) has shown, albeit in different words, that the p-curve is right-skewed when one tests relationships of which the null hypothesis is false. Simonsohn *et al.* (2013) show that this holds particularly if the statistical power of the research, which depends positively on sample size and true effect size, is large. The intuition behind this comes from Simonsohn *et al.* (forthcoming, pp. 7)[16]:

> *[I]magine a researcher studying the effect of gender on height with a sample size of 100,000. Because men are so reliably taller than women, the researcher would be more likely to find strongly significant evidence for this effect (…) than to find weakly significant evidence for this effect (…). Investigations of smaller effects with smaller samples are simply less extreme versions of this scenario.*

If researchers also test relationships with quite some statistical power, p-hacking becomes more difficult to prove statistically, because the right-skewness that results from the former and the left-skewness resulting from the latter might cancel each other out. Especially if p-hacking is done only sporadically, the right-skewness will dominate. Therefore, for researchers whose statistical tests have high power on average, p-hacking can only be formally proven by the p-curve if they "intensely p-hack". Still, p-hacking might be detectable by the p-curve even when it cannot be proven formally, since p-hacking tends to create a peak in the p-curve at and just below 0.05, which cannot be explained in any other way than by p-hacking (Brodeur *et al.* 2013).

4. Model

Assume we are in a world where agents[17], who are all risk-neutral, vary only in research ability $a \in [0,1]$ and scientific ethical integrity $e \in [0,1]$. a and e[18] are completely independently distributed. In this setting, e can be defined as the degree to which one dislikes

[15] Alternatively, one can test binomially whether there are more p-values in the range $(0.025; 0.05]$ as compared to the range $[0; 0.025]$.

[16] The exact page number is still unknown, since the paper is yet to be published. Instead, the working paper's page number is used. The working paper is accessible via http://papers.ssrn.com.

[17] Throughout the research, "agent" and "researcher" will be used interchangeably.

[18] See, for instance, Tabellini (2008) for a discussion on the origins of integrity/morality, and on incorporating that morality into economic models.

committing scientific fraud. The agent is assumed to dislike it only when he himself commits fraud and does not care when others do. Each agent has to decide whether to become an empirical social scientist (henceforth: researcher) or to take the best possible outside option $\overline{U}$, which is increasing in a[19]. Once the agent has decided to become a researcher, he does empirical research at zero cost[20]. Apart from his wage, any given researcher derives utility J from publishing scientific papers in an academic journal. If he publishes his findings only in working paper format, only W utils are obtained. Working papers yield less utility, since promotions, research budgets and academic rankings cannot be credibly linked to working papers, in contrast to journal publications[21]. This implies that agents will always send their research to academic journals in an attempt to get their work approved for publication. Assume that all submitted papers have a hypothesis in favor of the existence of a relationship, based on previous theoretical or empirical research. It is public knowledge that journals are more likely to publish significant results (probability of publication p) compared to insignificant results (probability π), because of the preference for significance coming from editor confirmation bias[22]. However, confirmation bias is weaker for more creative research topics. This arises because there is less previous research to support the hypothesis of more creative research, making the premise of the editor in favor of a hypothesis less pronounced.[23] This implies that, for more creative research topics, p and π are closer to one another. Assume realistically that more able agents come up with better ideas, which allows them to do more creative research[24]. Creative research also is on average more likely to be published than a paper that retests the same old hypothesis. All in all: $p \geq \pi, d\pi/da, dp/da > 0, d(p - \pi)/da < 0$. Assume without loss of generality that p and π are linear in a and that $\pi(0) = 0$ and $p(1) = \pi(1)$.

Directly after finishing his research, the agent observes whether the results are statistically significant. The probability that this occurs is exogenous and equal to s. In that case, the researcher will immediately send his work to an academic journal for approval, given the

[19] I leave for future research the possibility that outside option utility in- or decreases in e.
[20] This assumption excludes the possibility that agents avoid a possible research cost by making up data.
[21] Journals have no incentives to lie about the quality of the paper, so a journal publication could be a trustworthy signal of high effort. Journal publications are also easy to verify by courts, which allows the superiors to commit to paying out the promised incentive benefit. Rewarding working papers lacks this verifiability. This allows employers to understate paper quality in order to save on incentive costs. Peers cannot take over this quality check, because collusion is a real threat in that case: agent A agrees to give positive quality feedback to agent B, in return for agent B appreciating agent A's work.
[22] Adding other reasons why significance is preferred does not change the qualitative results of the paper.
[23] Researchers strategically revealing information to affect the editor's premise is left for future research.
[24] Creativity should not be mixed up with research quality in this case. This is because research quality interferes with statistical power, which affects the usefulness of the p-curve in detecting p-hacking.

assumption that writing the results in paper format is costless[25]. If the results turn out marginally insignificant, which happens with probability m, the researcher has three options to pick from. Firstly, he can decide to be honest by sending his work to a journal without further adjustment. The second option is to p-hack[26]. P-hacking will turn marginally insignificant results into significant ones, improving the odds of publishing in an academic journal. I assume that fraud cannot be detected by an academic journal or by a referee. This implies that the publication probability of a paper that displays significant results is p, whether or not the results are real. With probability c_h, p-hacking is discovered by an investigator. I assume that this occurs after the agent has been rewarded for his paper. If caught, the world of science will dismiss a fraudster and publically shame him. This burdens the agent with a cost of which the present value is equal to S. Any monetary payments that were given as reward for academic publications cannot be demanded back by the world of science due to limited liability.[27] Besides the extrinsic cost, deliberately misreporting a result also harms agents intrinsically, especially those who have a high integrity level. Particularly, when turning a marginally insignificant result into a significant one, an integrity cost of Ξ_M, which is strictly increasing in e, is incurred.[28] Manually changing data, or data fabricating, is also an option to mask marginally insignificant results. Getting caught doing this, which occurs with probability c_f, also leads to dismissal from the scientific world. When the results initially turn out highly insignificant, which happens with residual probability $1 - s - m$, p-hacking is not possible. There is no econometric specification that yields a significant result and deleting outliers must be done so rigorously that all readers will notice. Adding observations also cannot change the results in such a way that they become significant. The exception to this is adding or deleting only those observations that support significance. However, doing this cannot be called utilizing researcher degrees of freedom to adjust significance. Therefore, this tactic is a form of data fabrication, which is still possible. The integrity cost of doing so equals Ξ_I. This exceeds the integrity cost of masking marginally insignificant results, because one is misreporting the results less “severely” in the latter case. In other words, the integrity cost is increasing in the initial p-value.

For a schematic view, the model is repeated in Figure 1. N and A represent the nature player and the agent, respectively.

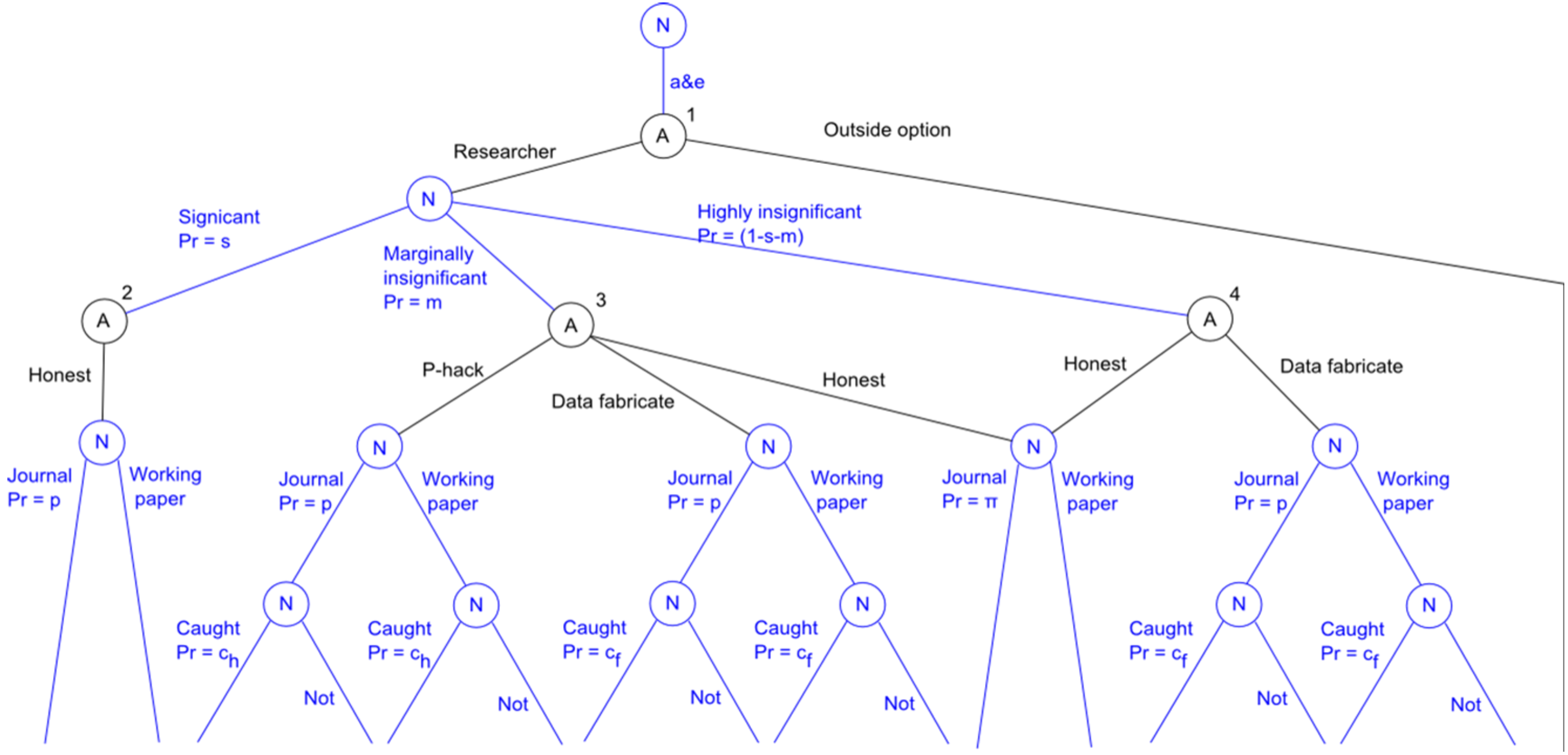
[25] This assumption also rules out the possibility of file drawering.

[26] In this paper, p-hacking is assumed to always be a conscious decision.

[27] All the assumptions about the detection and punishment of crime are without loss of generality.

[28] I assume without loss of generality that agents do not care about the way they change their results. Only the consequences of the change matter.

Figure 1



N
a&e
A
1
Outside option
Researcher
N
Signicant
Pr = s
Marginally
insignificant
Pr = m
Highly insignificant
Pr = (1-s-m)
A
2
A
3
A
4
Honest
P-hack
Data fabricate
Honest
Honest
Data fabricate
N
Journal
Pr = p
Working
paper
Journal
Pr = p
Working
paper
Journal
Pr = p
Working
paper
Journal
Pr = π
Working
paper
Journal
Pr = p
Working
paper
Caught
Pr = c_h
Not
Caught
Pr = c_h
Not
Caught
Pr = c_f
Not
Caught
Pr = c_f
Not
Caught
Pr = c_f
Not
Caught
Pr = c_f
Not

5. Analysis

To solve the model described in the previous section, backward induction is used. After solving the model accordingly, the stage is set to explore the effects of introducing the p-curve. Finally, several extensions to the model will be discussed.

The results are significant (Node 2)

It can be safely assumed that agents will honestly report their results when they turn out significant. Committing fraud has no benefits in terms of a larger probability of journal publication, while it does potentially have extrinsic and integrity costs. The utility of every agent at node 2 will therefore be equal to:

$$U(\text{significant}) = pJ + (1-p)W \tag{1}$$

The results are marginally insignificant (Node 3)

When the results turn out marginally insignificant, one can be honest or decide to commit fraud, either by p-hacking or by data fabricating. The utility outcomes of each of these actions are presented below:

$$U(\text{honest}|\text{marginally insignificant}) = \pi J + (1-\pi)W \tag{2}$$

$$U(\text{p}-\text{hack}|\text{marginally insignificant}) = pJ + (1-p)W - c_h S - \Xi_M \tag{3}$$

$$U(\text{data fabricate}|\text{marginally insignificant}) = pJ + (1-p)W - c_f S - \Xi_M \tag{4}$$

Since the maximization problem is discrete, it must be solved by comparing the utility of the options in pairs of two. Reporting honestly is preferred to p-hacking and data fabrication respectively if and only if these conditions hold:[29]

$$(J-W)(p-\pi) \leq c_h S + \Xi_M \tag{ICC5}$$

$$(J-W)(p-\pi) \leq c_f S + \Xi_M \tag{ICC6}$$

The left-hand side of incentive compatibility constraints ICC5 and ICC6 display the benefits of committing fraud, while the right-hand side reports the associated costs. The inequalities become stricter when the difference in utility between an academic journal and a working paper publication, $(J-W)$, is larger. For instance, installing a bonus per academic journal publication will result in a stronger incentive to commit fraud. Intuitively, when it pays off

[29] I assume agents choose to report honestly when indifferent because it is arguably the risk dominant strategy.

more, more people commit fraud. More people commit fraud also when the preference of journals for significant over insignificant results $(p - \pi)$ is stronger. This also implies that agents with a low a, for whom significance weighs heavily in the decision whether or not to publish, are more inclined to commit fraud than highly able researchers. Actually, agents with $a = 1$ will never commit fraud, since their creativity ensures that journal editors do not suffer from bias in favor of significance at all. When crime is punished harder or with a larger likelihood, fraud becomes less attractive, in line with the deterrence hypothesis. This is why the inequalities become less strict when $c_h(c_f)$ or S increases. Lastly and intuitively, an agent will commit less fraud when he has a higher integrity level, depicted by a larger value for Ξ_M.

There is one comparison left, which tells us what type of fraud the agent will choose when he does not want to be honest. An agent prefers p-hacking over data fabrication if and only if:[30]

$$c_h \leq c_f \tag{7}$$

There is just one difference between the two types of fraud, namely the chance of getting caught. Logically, the agent then chooses the option that is most covert. Assume that condition (7) is satisfied before the p-curve is introduced. This assumption is realistic, since one can defend p-hacking more easily precisely because it makes use of researcher's degrees of freedom. As there are no hard criteria that put a boundary on the use of these techniques, it is difficult to prove that a researcher has indeed premeditated to p-hack. For instance, adding observations can be defended by arguing that the sample size was initially too small, resulting in a lack of statistical power. Displaying a particular specification can be argued to fix endogeneity issues with other specifications. In a similar fashion, one can think of many excuses why some "outliers" are deleted. Data fabricating, on the other hand, cannot be defended in such a way. Another defense of the assumption is that making up data is hard to do in such a way that the data appear to be random. This makes detection by investigators more likely, for instance because they use Benford's law. On the basis of this, condition (7) is assumed to hold. This implies that ICC5 is less strict than ICC6. In other words, data fabrication is strictly dominated by p-hacking before the introduction of the p-curve.

[30] I assume agents choose p-hacking when indifferent. Allowing for any mix between both types of fraud will not affect the qualitative results, however.

The results are highly insignificant (Node 4)

If the initial results are highly insignificant, the agent faces a decision that is similar to that in the previous case. The only differences are that p-hacking is no longer possible and that the integrity cost function of data fabrication is now presented by $\Xi_I > \Xi_M$. Incorporating this, agents decide to report the results honestly if and only if:

$$(J - W)(p - \pi) \leq c_f S + \Xi_I \tag{ICC8}$$

The interpretation is similar to that of ICC5 and ICC6. Incentive compatibility constraint ICC8 is stricter than ICC6 and thus ICC5, implying that fewer agents decide to commit fraud at node 4 compared to node 3. This result follows directly from the assumption that the integrity cost function is increasing in the initial p-value.

Decision strategies

Based on the above, three decision strategies can arise in equilibrium before the introduction of the p-curve. These are described in Table 1. Note that the only relevant conditions are ICC5 and ICC8, since nobody will ever data fabricate when the results are marginally insignificant.

Table 1

Strategy	Condition	Actions
Strategy 1	ICC5 and ICC8 hold	Always reports the results honestly.
Strategy 2	ICC5 does not hold, ICC8 holds	P-hack when the results are marginally insignificant. In any other case, report the results honestly.
Strategy 3	ICC5 and ICC8 do not hold	Report significant results honestly, p-hack when the results are marginally insignificant and data fabricate when the results are highly insignificant.

Since Ξ_M and Ξ_I are positively related to e, ICC5 and ICC8 will depend on the agent's preference for integrity in science. a affects the incentive compatibility constraints via its effect on $(p - \pi)$. If one assumes the existence of an equilibrium in which all three decision strategies are represented, agents with a large value of e or a will adhere to Strategy 1. Agents with little scientific integrity or little ability will follow Strategy 3 and the remaining agents choose “hybrid” Strategy 2. This is described by the following Proposition, assuming Ξ_M, Ξ_I, p and π are continuous and invertible:

PROPOSITION 1: In equilibrium, before introducing the p-curve, the actions of agents who choose to be a researcher are determined by the following decision strategy:

1) Agents with integrity level $e \geq \bar{e}_a$ always report their results honestly.
2) Agents with integrity level $\underline{e}_a \leq e < \bar{e}_a$ report significant results honestly, p-hack to mask marginally insignificant results and report highly insignificant results honestly.
3) Agents with integrity level $e < \underline{e}_a$ report significant results honestly, p-hack to mask marginally insignificant results and data fabricate to mask highly insignificant results.

$\underline{e}_a$ is defined as $\Xi_I^{-1}\left[(J-W)(p-\pi)-c_f S\right]$ and
$\bar{e}_a$ is defined as $\Xi_M^{-1}[(J-W)(p-\pi)-c_h S]$, where Ξ^{-1} denotes the inverse of Ξ.
The three mentioned cases correspond to the strategies mentioned in Table 1.

The payoffs of each strategy are depicted in Figure 2a for the case that Ξ_M and Ξ_I are linear, keeping a fixed at a_0. The slopes of the lines are discussed later on. For now, notice that $\underline{e}_a$ and $\bar{e}_a$ both decrease in a:

$$\frac{d\underline{e}_a}{da} = \frac{d(p-\pi)}{da}(J-W) * \frac{d\Xi_I^{-1}}{d\left[(J-W)(p-\pi)-c_f S\right]} < 0, \tag{9}$$

$$\frac{d\bar{e}_a}{da} = \frac{d(p-\pi)}{da}(J-W) * \frac{d\Xi_M^{-1}}{d[(J-W)(p-\pi)-c_h S]} < 0 \tag{10}$$

This implies that, for more able agents, for whom the benefits of fraud are smaller, e needs to be lower in order to make it optimal to commit fraud. In Figure 2a, an increase in a would be depicted by all lines shifting upwards. The line of strategy 1 (3) shifts up the most (least). I will come back to this later on.

The decision strategies denoted in Proposition 1 can be written in terms of a as well:

COROLLARY 1: In equilibrium, before introducing the p-curve, the actions of agents who choose to be a researcher are determined by the following decision strategy:

1) Agents with research ability $a \geq \bar{a}_e$ always report their results honestly.
2) Agents with research ability $\underline{a}_e \leq a < \bar{a}_e$ report significant results honestly, p-hack to mask marginally insignificant results and report highly insignificant results honestly.

3) Agents with research ability $a < \underline{a}_e$ report significant results honestly, p-hack to mask marginally insignificant results and data fabricate to mask highly insignificant results.

$\underline{a}_e$ is defined as $(p-\pi)^{-1}\left[\left(c_f S + \Xi_I\right)/(J-W)\right]$, and
$\bar{a}_e$ is defined as $(p-\pi)^{-1}[(c_h S + \Xi_M)/(J-W)]$, where $(p-\pi)^{-1}$ denotes the inverse of $(p-\pi)$. The three mentioned strategies correspond to those in Table 1.

The payoffs of the strategies are again illustrated using a graph. In Figure 2b, e is kept fixed at e_0. Turning (9) and (10) around as well, one finds that $\underline{a}_e$ and $\bar{a}_e$ decrease in e:

$$\frac{d\underline{a}_e}{de} = \frac{d\Xi_I}{de}\frac{1}{(J-W)} * \frac{d(p-\pi)^{-1}}{d\left[\left(c_f S + \Xi_I\right)/(J-W)\right]} < 0, \tag{11}$$

$$\frac{d\bar{a}_e}{de} = \frac{d\Xi_M}{da}\frac{1}{(J-W)} * \frac{d(p-\pi)^{-1}}{d[(c_h S + \Xi_M)/(J-W)]} < 0 \tag{12}$$

An increase in e can be represented in Figure 2b by a downward shift of the line of strategy 3 and, to a lesser extent, that of strategy 2. The payoff of Strategy 1 is unaffected by the change.

Figure 2a

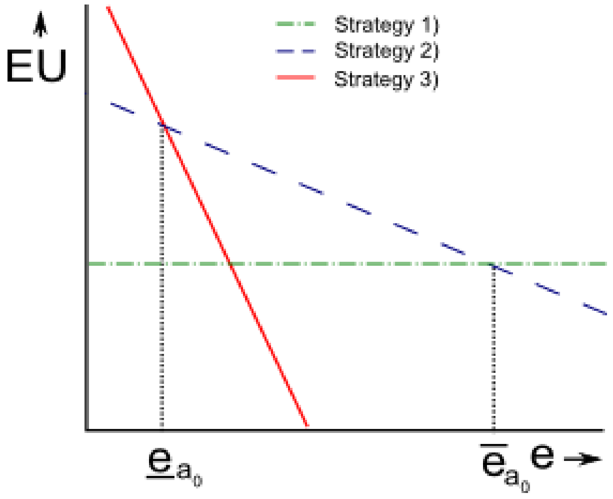
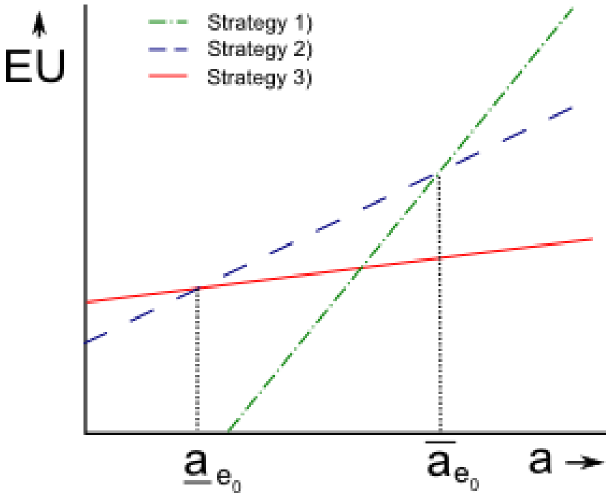


Figure 2b



The participation decision (Node 1)

The equilibrium put forward in Proposition 1 describes the behavior of agents conditional on their participation constraint being satisfied. This section reinforces the equilibrium by endogenizing that constraint.

Agents decide whether to participate if and only if their inside utility exceeds their outside utility. The inside utility equals the weighted average utility at the three bottom nodes, plus the (fixed) wage w. This is denoted by equation (13), where X denotes the chosen strategy in equilibrium:

$$EU = w + s * U(\text{significant}, X) + m * U(\text{marginally insignificant}, X) + (1 - s - m) * U(\text{highly insignificant}, X) \tag{13}$$

The solid line in Figure 3a illustrates equation (13) for the case that agents follow the decision strategy as discussed in Proposition 1. Note that the line is a copy of Figure 2a showing only the payoffs of the optimal strategy. When $e \geq \bar{e}_a$, the expected utility is "independent" of e. This is because these agents never commit fraud, and thus they are not harmed intrinsically by their actions, no matter the level of their integrity. If $\underline{e}_a \leq e < \bar{e}_a$ holds, one chooses Strategy 2. In that case, expected utility does depend (negatively) on e. Particularly, $\partial U(\text{marginally insignificant},\ X = 2)/\partial e < 0$, because Strategy 2 involves committing fraud only when the results show up marginally significant. The higher e, the more costly this strategy is, which explains why the slope of the solid line in Figure 3a is negative between $\underline{e}_a$ and $\bar{e}_a$. For the lowest integrity levels, Strategy 3 is optimal. In that case, both $U(\text{marginally insignificant}, X = 3)$ and $U(\text{highly insignificant}\ X = 3)$ depend negatively on e. This is the reason why the solid line is the steepest in the left section of Figure 3a.

Figure 3a

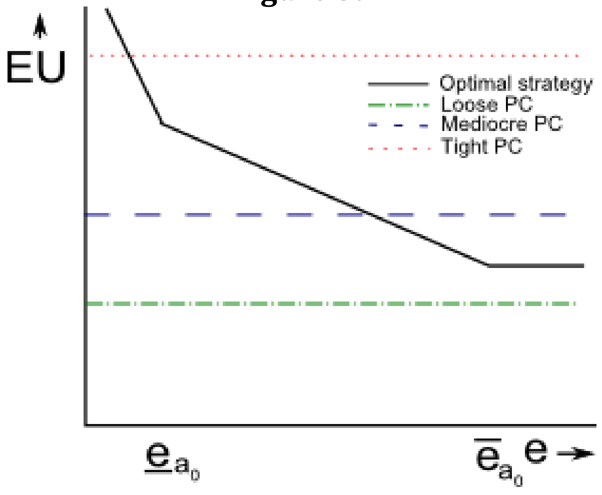
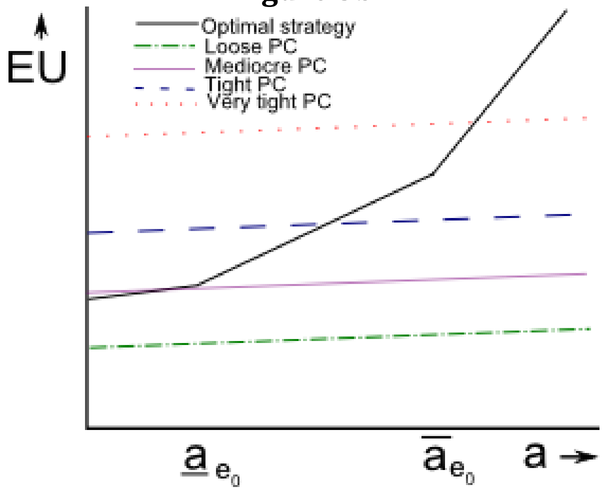


Figure 3b



When an agent chooses to take his outside option, he obtains a utility of $\overline{U} = \overline{w} + \kappa(a)$.[31] $\overline{w}$ depicts the outside wage earned by an agents with ability level 0. Assume $\kappa(0) = 0$ and $d\kappa/da > 0$. Notice that the outside option does not depend on e, which arises partly because

[31] This specification inescapably entails some loss of generality.

of the assumption that agents are only harmed by their own fraud and not by that of others[32]. Since the "inside utility" does depend (weakly negatively) on e, participation (weakly) decreases in e. Then, the participation constraint of an agent can be represented by:

$$EU \geq \bar{w} + \kappa(a) \qquad \text{(PC14)}$$

In Figure 3a, the outside option is represented by one of the three[33] dotted lines, depending the parameters of the model. When $\overline{U}$ takes on a relatively low value, the participation constraint is considered loose: all agents participate. When the outside option is somewhat more attractive (mediocre PC), the agents with the highest integrity will not participate. These agents gain relatively too little from publishing papers, or have too little chance of an academic publication, and rather opt out of science. The agents with less integrity do choose to be a researcher, because committing fraud gives them extra utility. If the participation constraint is tight, only the agents with very little integrity will find it optimal to participate. Whatever participation constraint is relevant, Proposition 2 always holds:

> **PROPOSITION 2:** In equilibrium, only agents with integrity level $e \leq e_a^{*}$[34] choose to participate, where e_a^* is defined as the unique level of e where an agent of ability a is indifferent between participating and taking the outside option. If there is no such indifference level, $e_a^* = 1$.

The solid line in Figure 3b represents the optimal strategy of agents when ability is put on the horizontal axis rather than integrity. At both $\underline{a}_e$ and $\bar{a}_e$, the slope of the line becomes steeper. The reason is that, when $a < \underline{a}_e$, the agent's paper will always display significant results to the journal. This means that the publication chance is always equal to p. An increase in a then leads to an increase in publication probability by dp/da. When $\underline{a}_e \leq a < \bar{a}_e$, the publication probability is equal to π when the results turned out highly insignificant, because these agents decide to report those results honestly. In any other case, p is still the relevant publication probability. In expected terms, an increase in research ability will in this case lead to an

[32] If agents are harmed intrinsically by fraud of others, the outside option does depend on e. If an agent who opts out is replaced by a new agent, the former is harmed if the replacement decides to p-hack or data fabricate. In this case, for agents with $e \geq \bar{e}_a$, participation increases in e. The reason is that these agents themselves would never commit fraud. However, replacements do have a non-zero probability of committing fraud. If e increases, "letting a replacement do the research" thus becomes less attractive. For agents with $e < \underline{e}_a$, participation unambiguously decreases in e. The reason is that replacements have a less than one probability of committing fraud, in contrast to what these agents would like. For the remaining agents, an increase in e has ambiguous effects on participation, depending on the (conditional) distribution of ability and integrity of replacements.

[33] One other case is not depicted: no agents participate (outside option above the solid line).

[34] Assuming that one participates when indifferent.

increase in the chance of publication by $(s+m)*dp/da+(1-s-m)*d\pi/da$. Since $d\pi/da > dp/da$, this increase is larger than before. This explains why the slope of the solid line is larger here. If $a \geq \bar{a}_e$, the publication probability is also π when the results are just insignificant. Thus, the slope of the expected utility curve also becomes steeper at $\bar{a}_e$.

Research ability can be put to fruitful practice in other professions as well. Therefore, $d\bar{U}/da > 0$. But since research ability is used most when one is a researcher, it is reasonable to assume that inside utility increases faster in a than does outside utility: $d\bar{U}/da < (J-W)*dp/da$. This assumption ensures that Proposition 3 always holds:

> **PROPOSITION 3:** In equilibrium, only agents with research ability $a \geq a_e^*$ choose to participate, where a_e^* is defined as the unique level of a where an agent of ability e is indifferent between participating and taking the outside option.

This is illustrated in Figure 3b, where the four possible cases of the outside option are represented, where $\bar{U}$ in this case is linear in a. I assume that $\bar{w}$ is not so large that no agent will participate. Combining the insights from Propositions 2 and 3 and the narratives leading to them yields:

> **LEMMA 1:** $de_a^*/da \geq 0,\ da_e^*/de \geq 0$

Since inside utility relative to outside option utility increases in a and (weakly) decreases in e, the region in which it is profitable to participate ($0 \leq e \leq e_a^*$ or $a_e^* \leq a \leq 1$) increases in a and decreases in e as well. de_a^*/da is weakly positive, because e_a^* cannot rise above 1. da_e^*/de is weakly positive because in the region $e \geq \bar{e}_a$, an increase in e does not change inside or outside utility. This insight from Lemma 1 can be used to draw inferences about the expected level of integrity for different levels of ability. In particular:

> **PROPOSITION 4:** In equilibrium, the expected level of integrity is higher in a group of scientists with higher research ability. Equivalently, the expected level of ability is lower amongst scientists with little ethical integrity.

The proof of Proposition 4 can be found in Appendix A. This result can be used strategically by agents when applying for a job as researcher. When a and e are not known by a prospective academic employer yet, showing off good ethical values can be a signal of ability.

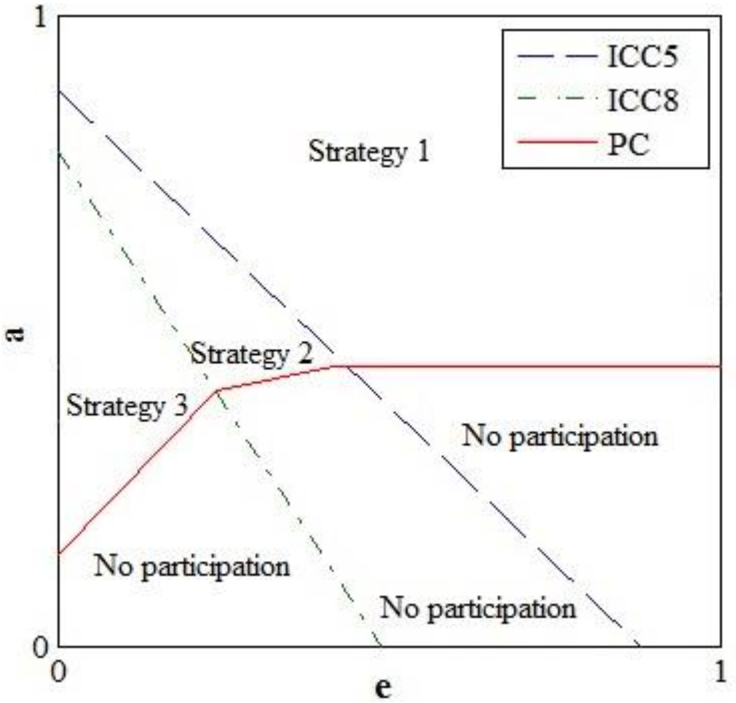
A way to do this is, for instance, participating in charity activities. This can become a signal because it is costly to replicate the signal for agents with low e[35]:

> **<u>PROPOSITION 5:</u>** In labor markets where a and e are private knowledge, revealing a high ethical integrity can be a signal of high research ability.

Equilibrium

The complete equilibrium is illustrated in Figure 4. For simplicity, Ξ_M and Ξ_I are presented as linear in e and $\overline{U}$ is linear in a. The exact parameter inputs can be found in Appendix B.

Figure 4



In Figure 4, the intermittent lines describe the two relevant incentive compatibility constraints. The ICCs are satisfied for those agents with a and e to the northeast of those lines. This confirms Proposition 1 and Corollary 1, which describe that agents with higher research ability or integrity behave more ethically. The most able agents will always behave according to Strategy 1, and under realistic parameter cases[36], the most ethical ones will do likewise. Agents with low scores for either of the attributes adhere to unethical Strategy 3 and the others choose Strategy 2. Obviously, agents can only choose to do this when they decide to become a researcher. The participation decision is described by the solid line, which

[35] A necessary condition for this signal to be costly to replicate is that e is positively associated to more general ethical integrity. In that case, agents with low e find it more costly to show off good moral values. For a formal model on signaling in job markets with productivity uncertainty, see Spence (1972).

[36] A realistic parameter case would be when Ξ_m and Ξ_I are such that all agents with the highest levels of e will find it prohibitively costly to commit any type of fraud.

depicts e_a^* or a_e^* from Propositions 2 and 3 respectively. Notice that in "Strategy area" 3, the slope of the PC is steeper than in area 2. This means that in the former area, a marginal increase in e must be compensated by a larger increase in a in order to still make it optimal for the agent to participate, in line with Figure 3a. In area 1, the agent would never commit fraud, so his utility is independent of e, which explains why the PC line is straight there. The participation constraint also confirms that expected research ability conditional on participation weakly increases in e, since the distance between the PC and the graph ceiling weakly decreases in e. Similarly, one can see that the conditional average integrity level weakly increases in a.

6. Introducing the P-curve

After characterizing the equilibrium, it is time to introduce the p-curve. I model this as an increase in the probability that one is caught p-hacking: $c_h' > c_h$.[37] Two cases can be distinguished. The first is when the effect of the p-curve is modest, i.e. that p-hacking still dominates data fabrication in the case that the results turn out marginally significant. The second case is when the opposite holds true. Recall that p-hacking cannot easily be proven if one incidentally p-hacks (Simonsohn *et al.* 2013)[38]. With this in mind, I assume that the action that researchers take in the model is representative for a series of their publications, not just the one under scope[39]. Furthermore, it is assumed that the power of the agents' research is low enough to ensure that p-hacking is indeed detectable with non-zero probability.

The effect of the p-curve is modest

Incentive effects

As a start, reconsider condition (7). Suppose that the effect of introducing the p-curve is small. In that case, condition (7) remains satisfied: $c_f \geq c_h' > c_h$ and data fabrication is still

[37] One could argue that the p-curve will also increase the probability of getting caught data fabricating c_f, since investigators are looking at the density plot of significant p-values, which could contain evidence for data fabrication as well. However, as long as the increase in c_f is smaller than the increase in c_h, the qualitative results of the paper do not change. A sufficient condition for this to hold is that the correlation between initial and after-fraud p-value is higher for p-hacking than for data fabricating. This is reasonable to assume, since data fabrication is not constrained by the available "researcher degrees of freedom". Therefore, data fabricators may select basically any desired p-value, while p-hackers can only arrive at marginally significant p-values.

[38] Allowing for mixed strategies, where agents incidentally p-hack to benefit from the fact that sporadically p-hacking is hard to prove using the p-curve, will not change the qualitative results of the paper. It does, obviously, reduce the effectiveness of the p-curve at correcting fraudulent behavior.

[39] This is in line with Ehrlich (1973), who posed that choosing to commit crime is some sort of an occupational choice rather than an incidental decision.

strictly dominated when the results are initially marginally insignificant. However, in that case, the utility of p-hacking does decrease by $(c_h' - c_h)S$. This implies a parallel shift downwards of the expected utility curves of Strategy 2 and 3. This is represented in Figure 5a and 5b.

Figure 5a
Before p-curve (Figure 2a)

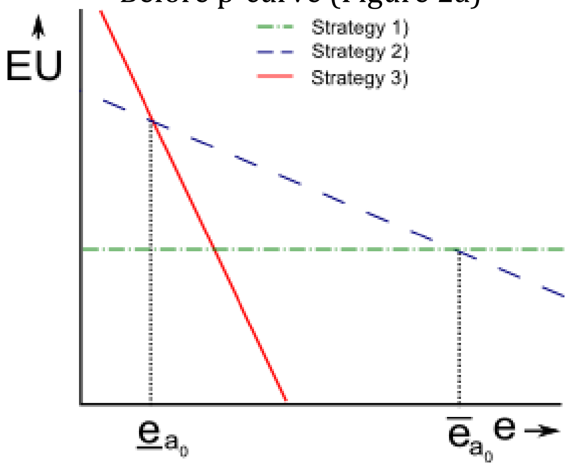


Figure 5b
After p-curve

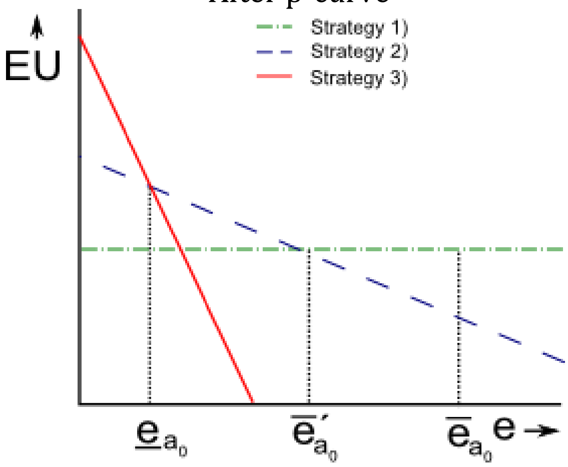


Figure 5c
Before p-curve (Figure 2b)

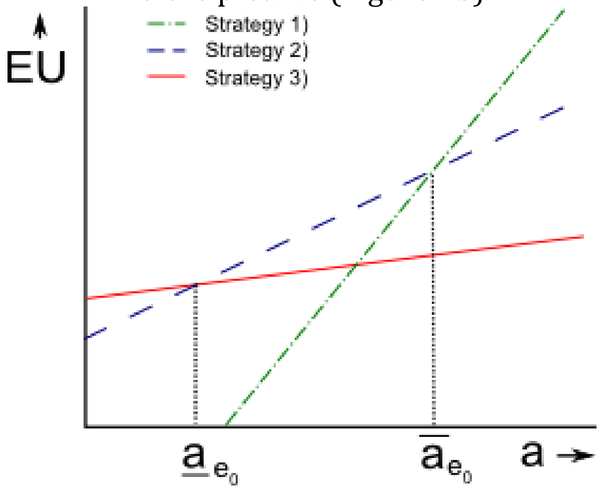
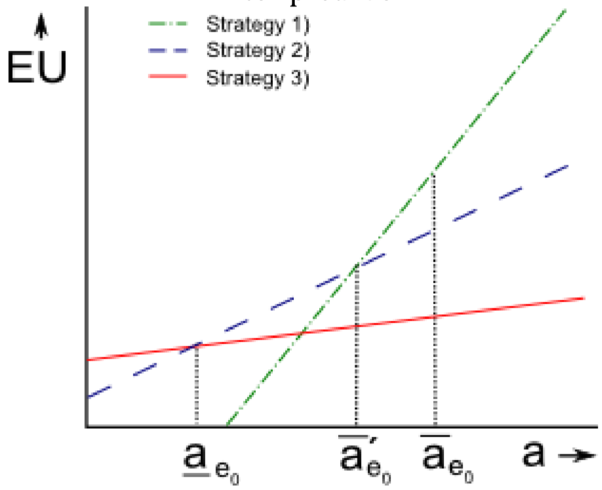


Figure 5d
After p-curve



The shift implies that the maximum level of integrity at which one chooses Strategy 2 $\bar{e}_a'$, is smaller than before: $\underline{e}_a \leq \bar{e}_a' < \bar{e}_a$. Intuitively, when the chance of getting caught p-hacking increases, less people will be willing to commit this type of fraud. The threshold between Strategy 2 and 3, $\underline{e}_a$, is not affected, since both strategies involves p-hacking with the same probability. Therefore, the utility curves of both strategies shift downwards by the same amount. The "incentive effect" of introducing the p-curve is thus that agents in the range $[\bar{e}_a', \bar{e}_a)$ will now always decide to report their results honestly:

PROPOSITION 6: Introduction of the p-curve has the following incentive effects, provided that the effect of the p-curve on the probability of p-hacking detection is modest and that the participation constraint is satisfied:

1) The decision of agents with integrity level $e \geq \bar{e}_a$ is not affected.
2) Agents with integrity level $\bar{e}'_a \leq e < \bar{e}_a$will no longer decide to p-hack marginally insignificant p-values, but report their results honestly instead.
3) The decision of agents with integrity level $e < \bar{e}'_a$ is not affected.
$\bar{e}_a$ is defined as in Proposition 1, $\bar{e}'_a = \Xi_M^{-1}[(J - W)(p - \pi) - c'_h S]$

Figure 5c and 5d represent the same result in terms of ability. It shows that:

COROLLARY 2: Introduction of the p-curve has the following incentive effects, provided that the effect of the p-curve on the probability of p-hacking detection is modest and that the participation constraint is satisfied:

1) The decision of agents with research ability $a \geq \bar{a}_e$ is not affected.
2) Agents with research ability $\bar{a}'_e \leq a < \bar{a}_e$ will no longer decide to p-hack marginally insignificant p-values, but report their results honestly instead.
3) The decision of agents with research ability $a < \bar{a}'_e$ is not affected.
$\bar{a}_e$ is defined as in Corollary 1, $\bar{a}'_e = (p - \pi)^{-1}[(c'_h S + \Xi_M)/(J - W)]$

The p-curve thus induces some agents (with mediocre integrity or ability) to report honestly rather than to p-hack. The agents who are highly unethical or very unable will still commit fraud, because for them the extra cost of p-hacking is insufficient to compensate for respectively the low integrity costs and the strong increase in publication probability that committing fraud yields them.

Selection effects

Introducing the p-curve affects the participation constraint of agents who p-hacked before the introduction of the p-curve. This is shown in Figure 6a, where the solid line reflects the optimal strategy both before and after the introduction of the p-curve. Both arrows in the curve represent a downward shift of $m(c'_h - c_h)S$ utils[40]. One can see that the people who still commit fraud after introduction of the p-curve (those with $e < \bar{e}'_a$) lose exactly this amount.

[40] This value is obtained by multiplying the decrease in the utility of p-hacking, $(c'_h - c_h)S$, by the probability that one obtains a marginally insignificant or "p-hackable" p-value (m).

Agents that were p-hacking before, but now choose to be honest, lose less than this. In fact, the closer the agent's e is to $\bar{e}_a$, the less utility is lost from the introduction of the p-curve. This is because the preference for p-hacking over reporting honestly decreases in e. Agents close to $\bar{e}_a$ lose little when they switch from p-hacking to reporting honestly, as compared to agents who are close to $\bar{e}'_a$. Agents with integrity below $\bar{e}'_a$ would lose even more, which is exactly the reason why they do not switch to reporting honestly at all.

Figure 6a

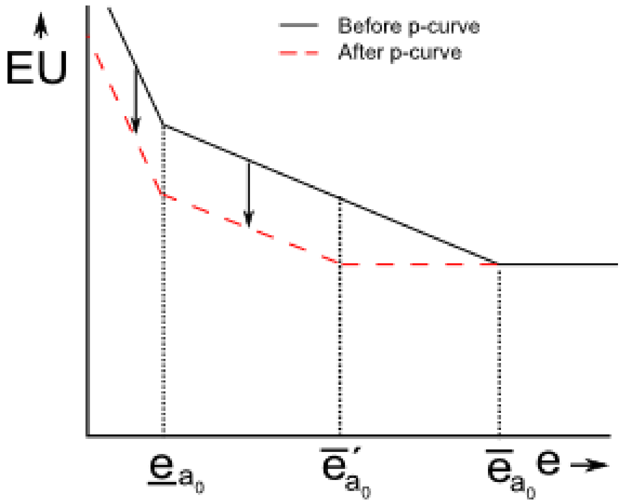
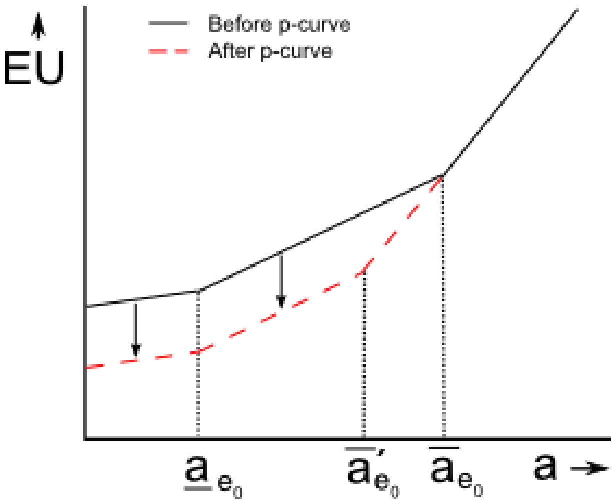


Figure 6b



Given that the outside option is a horizontal line, participation will decrease after the introduction of the p-curve, unless even the agents with integrity level $e \geq \bar{e}_a$ participate. In that case, participation is not affected by the p-curve:

<u>**PROPOSITION 7:**</u> Introduction of the p-curve has the following selection effects:

1) If $e_a^* = 1$, there are no selection effects: all agents still participate: $e'^*_a = 1$
2) If $e_a^* < \bar{e}_a$[41], agents with $e'^*_a < e \leq e_a^*$ no longer participate.

e_a^* is defined as in Proposition 2, e'^*_a is defined as the unique level of e where an agent of ability a is indifferent between participating and taking the outside option after introduction of the p-curve. If all agents participate, $e'^*_a = 1$.

The result of Proposition 7 is somewhat paradoxical: the introduction of the p-curve leads to (weakly) less participation by agents of relatively high integrity. This result holds, however, only when agents who would report honestly do not participate ($e_a^* < \bar{e}_a$). In this case, the participating agents with the highest e will want to commit fraud, and thus incur higher costs

[41] $e_a^* < 1$ would be a correct representation as well, since e_a^* by definition does not lie in the range $[\bar{e}_a, 1)$.

if the p-curve is introduced. For these agents, the extra cost can be high enough to induce them to quit science, because they already face quite some costs in terms of integrity. On the other hand, for the highly unethical agents, the p-curve does not produce a high enough cost to make it optimal to pick the outside option, because they face little integrity costs when participating.

The downward shift in utility arising from the p-curve is also depicted for each level of ability, see Figure 6b. Also here, the downward shift wears of between $\bar{a}'_e$ and $\bar{a}_e$, for the same reason as described before. If, before the p-curve, only agents with ability $a > \bar{a}_e$ participated, the p-curve will shift only the utility curves of those who did not participate anyway. In that case, the p-curve will have no selection (nor incentive) effects at all. Otherwise, the p-curve will induce some (p-hacking) agents to quit the scientific world.

PROPOSITION 8: Introduction of the p-curve has the following selection effects:

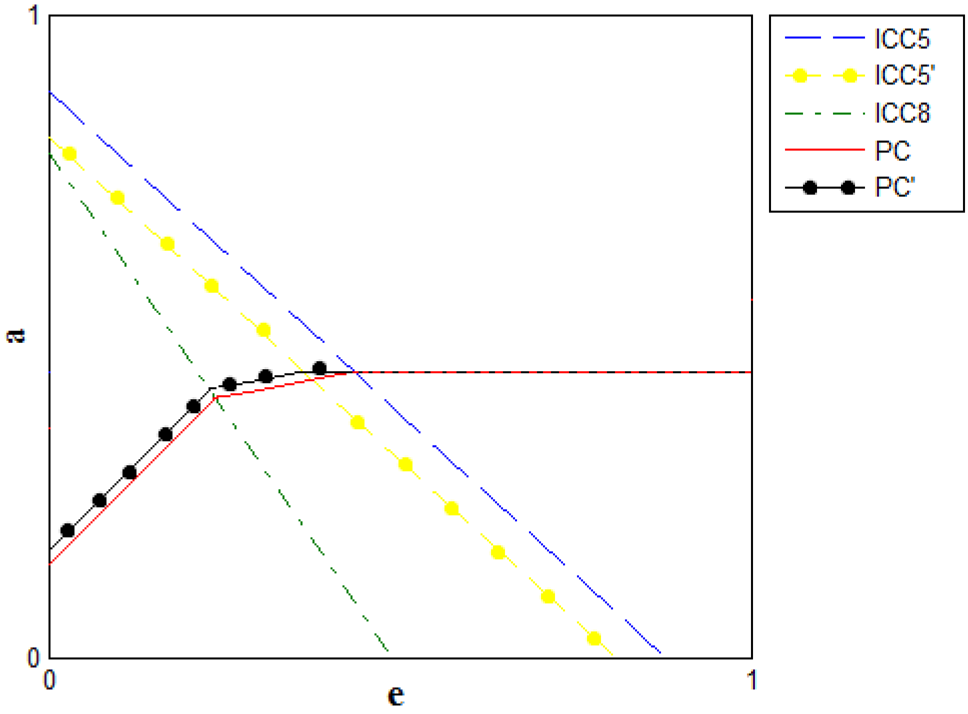
1) If $a_e^* < \bar{a}_e$, agents with $a_e^* \leq a < a'^*_e$ no longer participate.
2) If $a_e^* \geq \bar{a}_e$, there are no selection effects: $a'^*_e = a_e^*$.

a_e^* is defined as in Proposition 3, a'^*_e is defined as the unique level of a where an agent of integrity level e is indifferent between participating and taking the outside option after introduction of the p-curve.

Figure 7 shows the new equilibrium, fitted in the same figure as the initial equilibrium. The shift to the southwest of the intermittent line representing ICC5 illustrates the incentive effect of the p-curve. In line with Proposition 6 and Corollary 2, the fraudsters with the highest ability or the highest integrity now choose to report their results honestly. Propositions 7 and 8 are illustrated by the shift to the northwest of the participation constraint in Strategy areas 2 and 3. This shift induces the formerly participating agents with the lowest ability and the highest integrity in these areas to leave the world of science. Notice that the participation constraint shifts up by a smaller amount in the region between ICC5 and ICC5', in line with Figures 6a and 6b. It is easy to see in Figure 7 that the p-curve weakens the link between e and a. In particular, the participation constraint becomes "more horizontal". This implies that the p-curve diminishes the opportunity for agents to effectively signal their ability through a revelation of high integrity. Therefore:

PROPOSITION 9: Introduction of the p-curve results in an equilibrium where less agents signal their research ability to employers through revelation of high integrity.

Figure 7



It is possible that society ends up in a pooling equilibrium after introduction of the p-curve, because it is no longer beneficial for anyone to signal ability. This is problematic if there are large benefits attached to a good match between academic employer and researcher. It can be beneficial as well, if signaling is merely privately optimal.

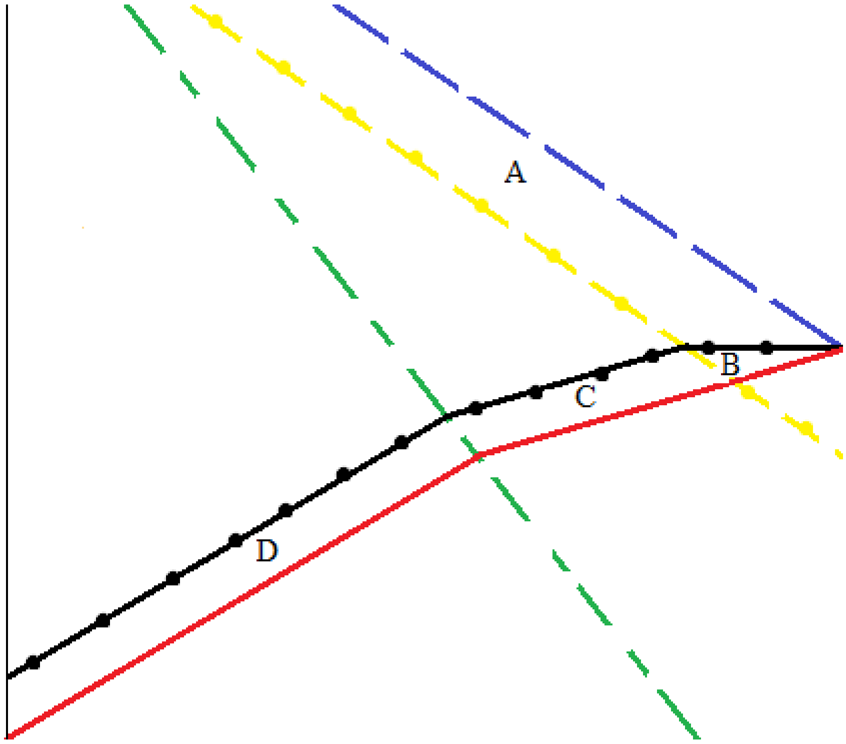
Welfare analysis[42]

The result that the p-curve corrects "bad" behavior and induces some fraudsters to leave science, does not automatically imply that the net effect of the p-curve is beneficial to society. To see why, let's take a look at Figure 8, a zoomed-in version of Figure 7. The incentive effect is depicted by area A, the selection effect of the p-curve is represented by areas B, C and D. Suppose simply that society gains utility B_J when a result is reported in a journal honestly. An honest result is worth $B_W < B_J$ utils when it ends up in working paper format. Society loses utility $C_{k,q}$ when a result is misreported. This costs depends on the format of publication $k \in \{J, W\}$ and the insignificance of results $q \in \{m, I\}$. The loss to society is larger when the misreported result ends up in an academic journal ($C_{J,q} > C_{W,q}$) and when the misreported result is more insignificant: $C_{k,I} > C_{k,m}$. It depends on society's preferences whether a highly insignificant result misreported in a working paper is more damaging than a

[42] This welfare analysis is also representative for the case that the effect of the p-curve is profound.

marginally insignificant result that is published as significant in an academic journal: $C_{J,m} \gtrless C_{W,I}$.

Figure 8



R reflects the cost to society of re-matching agents with employers because some agents opt out of science[43]. Another factor incorporated in R is the possibility that the exodus of researchers necessitates in increase in w in order to attract replacements[44]. I will refer to it as adjustment costs. Keep in mind that R also incorporates the degree to which an agent and his replacement are productive or destructive to society in their best alternative occupations. For instance, a fraudster might be scared away from science because of the p-curve. Instead, he joins the financial world, where he might be much more destructive than in science. Using the above, the incentive effect has the following welfare[45] consequences:

$$\text{Incentive effect} = \text{m}\int_A \left(\pi B_J + (1-\pi)B_w + \left[pC_{J,m} + (1-p)C_{W,m}\right]\right) f(a)da > 0 \quad (15)$$

[43] R may be negative as well, indicating that society achieves a better match between employers and agents.

[44] Allowing R, B_w, B_J and $C_{k,q}$ to vary in ability and ethical integrity is left to future research. These variations are not essential to arrive at the main conclusion of this section.

[45] Note that society does not gain from reading significant results as compared to insignificant results, since the “preference” for significance arises from a bias.

Here, $f(a)$ denotes the density function of a and A indicates the domain to be integrated: area A of Figure 8. Since the actions before and after the p-curve only change for the case that the results are marginally insignificant, equation (15) only denotes the change in welfare for that particular case (probability m). $\pi B_J + (1 - \pi)\, B_w$ equals the welfare that results from the action of agents in area A after the p-curve is introduced. This number is positive, because these agents report marginally insignificant results honestly when the p-curve is introduced. $[pC_{J,m} + (1 - p)C_{W,m}]$ reflects the absolute value of the loss that society incurred before the p-curve was introduced, because agents used to p-hack marginally insignificant results. The p-curve prevents this loss, which makes the second term positive as well. All in all, the incentive effect is beneficial to society, because agents shift from harmful fraud towards welfare improving honesty. This effect is larger if $f(a)$ is larger in area A.

Areas B and C in Figure 8 depict the first selection effect of the p-curve: agents that used to adhere to Strategy 2 now leave science. The welfare consequences of this first selection is denoted by equation (16):

$$\text{Selection effect 1} =$$

$$-\int_{B+C} \langle \begin{pmatrix} s[pB_J + (1-p)B_w] + (1-s-m)[\pi B_J + (1-\pi)B_w] - \\ m[pC_{J,m} + (1-p)C_{W,m}] \end{pmatrix} + R \rangle f(a)da \qquad (16)$$

The part of equation (16) between the large brackets denotes the value to society of an agent in area B or C if he would participate. Society in that case would get some utility because agents in areas B and C would report significant and highly insignificant results honestly. On the other hand, these agents do report marginally insignificant results dishonestly, which would harm social welfare. The (unweighted) opportunity cost of having these agents leave science is the contribution to social welfare if they would have stayed, plus adjustment costs, as denoted by the R. This opportunity cost cannot be unambiguously signed without future research regarding the signs and magnitudes of the parameters in equation (16). To obtain the weighted magnitude of the selection effect, one also needs to estimate $f(a)$, which describes how densely areas B and C are populated.

Area D represents the second selection effect, which dictates that some agents who used to behave according to Strategy 3 leave science after introduction of the p-curve. The change in welfare because of this exodus is represented in equation (17):

$$\text{Selection effect 2} = -\int_D \left\langle \begin{pmatrix} s\left[pB_J + (1-p)B_w\right] - m\left[pC_{J,m} + (1-p)C_{W,m}\right] - \\ (1-s-m)\left[pC_{J,I} + (1-p)C_{W,I}\right] \end{pmatrix} + R \right\rangle f(a)da \tag{17}$$

Also this second selection effect has ambiguous sign and magnitude. The interpretation of it is similar to that of the first selection effect. There is one striking thing that is worth mentioning. The agents in area D, being on average less able then those in areas B and C, have a smaller chance to get their research published, depicted by a lower value for p (and π). Therefore, their research will on average have less influence on social welfare. On the one hand, this harms society, because significant results will benefit society less. On the other hand, society is harmed less by insignificant results being published as significant. The above tells us that agents in area D not necessarily destroy more social welfare than those in area B or C, even though they commit fraud more often. Suppose for instance that agents in area D have such a small publication probability that their results are seldom published. In that case, they can hardly damage society, in contrast to agents in area B and C, who might have a publication probability that is quite high. This implies that the part of equation (17) between large brackets need not be smaller than that of equation (16).

Adding up equations (15) to (17) brings us the overall welfare effect of the p-curve. If the incentive effect is large enough, the p-curve will be beneficial to society. This occurs, for example, when there are many agents in area A or when the costs of fraud to society are large. One should, however, not leave the possibly adverse selection effects out of the discussion. For instance, one should keep in mind that the fraud committed by leavers might not burden society too much anyway ($C_{k,q}$ is low). Incurring the adjustment costs to correct this might therefore not be worthwhile.

PROPOSITION 10: The overall welfare effect of the p-curve can be negative if the following factors hold in combination:

1) R is high
2) $C_{k,q}$ is low
3) Few agents are located in area A of Figure 8.
4) Many agents are located in area B, C and/or D of Figure 8.

One should also take into account the parameters with ambiguous effects on welfare. For instance, m affects the size of the four areas, and this parameter increases the (unweighted) incentive effect and first selection effect. However, it might decrease the second selection effect if $C_{k,I}$ is large enough relative to $C_{k,m}$. Therefore, the overall effect of an increase in m is ambiguous. This is a counterintuitive result, as one would expect that the p-curve will be more attractive if the chance that it can effectively be applied increases. Increases in B_j and B_W increase the incentive effect, but decrease both selection effects. And obviously, one should not forget the parameters upon which agents base their decision, because these determine the size of the four areas.

The effect of the p-curve is profound

Now consider the case that the p-curve increases the probability of getting caught p-hacking by such a margin that condition (7) is no longer satisfied: $c_h'' > c_f \geq c_h$. In this case, p-hacking is strictly dominated by data fabrication. Therefore, the agents' possible decision strategies should be re-characterized, since some of the ones described in Table 1 are strictly dominated when $c_h'' > c_f$. Also, condition (5) is no longer relevant. Incentive compatibility constraint (6) should be looked at instead. Table 2 describes the new possible strategies:

Table 2

Strategy	Condition	Actions
Strategy 1	ICC6 and ICC8 hold	Always reports the results honestly.
Strategy 2b	ICC6 does not hold, ICC8 holds	Data fabricate when the results are marginally insignificant. In any other case, report the results honestly.
Strategy 3b	ICC6 and ICC8 do not hold	Reports significant results honestly and data fabricate when the results are insignificant.

Incentive effects

Besides the change in strategies, the analysis is comparable to that of the previous case. I focus the discussion of the p-curve on the case where the results are marginally significant, as the p-curve has incentive effects only then. Known is that agents with integrity level $e \geq \bar{e}_a$ would never commit fraud before the p-curve, so when the p-curve is introduced, their preference for reporting honestly is even stronger. Therefore, these agents will still choose Strategy 1. The p-curve discourages fraud by somewhat less ethical agents (agents just below $\bar{e}_a$). For them, the preferred action after observing a marginally insignificant result initially was p-hacking, followed by reporting honestly. But now, when p-hacking is no longer part of an optimal strategy, these agents will always report their results honestly and switch to Strategy 1. In other words: the threshold until which one commits fraud after observing a marginally insignificant result now becomes $\bar{e}_a'' < \bar{e}_a$. Since $\bar{e}_a'' < \bar{e}_a'$, the incentive effect is larger when the effect of the p-curve is profound, which is intuitive. For agents with $e < \bar{e}_a''$, the preferred action when p-hacking is no longer optimal is data fabrication. Thus, for these agents, the p-curve has no incentive effects; it only induces them switch from p-hacking to data fabrication if the results are marginally insignificant. Which strategy these agents choose depends on what they want to do when their results turn out highly insignificant. Agents who are honest in that case, choose Strategy 2b. Others opt for Strategy 3b:

PROPOSITION 11: Introduction of the p-curve has the following incentive effects, provided that the effect of the p-curve on the probability of p-hacking detection is profound and that the participation constraint is satisfied:

1) The decision of agents with integrity level $e \geq \bar{e}_a$ is not affected.
2) Agents with integrity level $\bar{e}_a'' \leq e < \bar{e}_a$will no longer decide to p-hack marginally insignificant p-values, but report all their results honestly instead.
3) Agents with integrity level $\underline{e}_a \leq e < \bar{e}_a''$will no longer decide to p-hack marginally insignificant p-values, but data fabricate these p-values instead. The other p-values are still reported honestly.
4) Agents with integrity level $e < \underline{e}_a$ will no longer decide to p-hack marginally insignificant p-values, but data fabricate these p-values instead, just like they data fabricate highly insignificant p-values.

$\bar{e}_a$ and $\underline{e}_a$ are defined as in Proposition 1, $\bar{e}_a'' = \Xi_M^{-1}\left[(J - W)(p - \pi) - c_f S\right]$

Analogously:

COROLLARY 3: Introduction of the p-curve has the following incentive effects, provided that the effect of the p-curve on the probability of p-hacking detection is profound and that the participation constraint is satisfied:

1) The decision of agents with research ability $a \geq \bar{a}_e$ is not affected.

2) Agents with research ability $\bar{a}_e^{''} \leq a < \bar{a}_e$ will no longer decide to p-hack marginally insignificant p-values, but report all their results honestly instead.
3) Agents with research ability $\underline{a}_e \leq e < \bar{a}_e^{''}$will no longer decide to p-hack marginally insignificant p-values, but data fabricate these p-values instead. The other p-values are still reported honestly.
4) Agents with research ability $a < \underline{a}_e$ will no longer decide to p-hack marginally insignificant p-values, but data fabricate these p-values instead, just like they data fabricate highly insignificant p-values.

   $\bar{a}_e$ and $\underline{a}_e$ are defined as in Proposition 1, $\bar{a}_e^{''} = (p - \pi)^{-1}\left[\left(c_f S + \Xi_M\right)/(J - W)\right]$

Figures 5a to 5d are representative for Proposition 11 and Corollary 3 as well, if one substitutes $\bar{a}_e'$ for $\bar{a}_e^{''}$ and reads Strategy 2b (3b) instead of Strategy 2 (3). Keep in mind that the parallel shift of the curves are now larger than when the effect of the p-curve is modest.

Selection effects

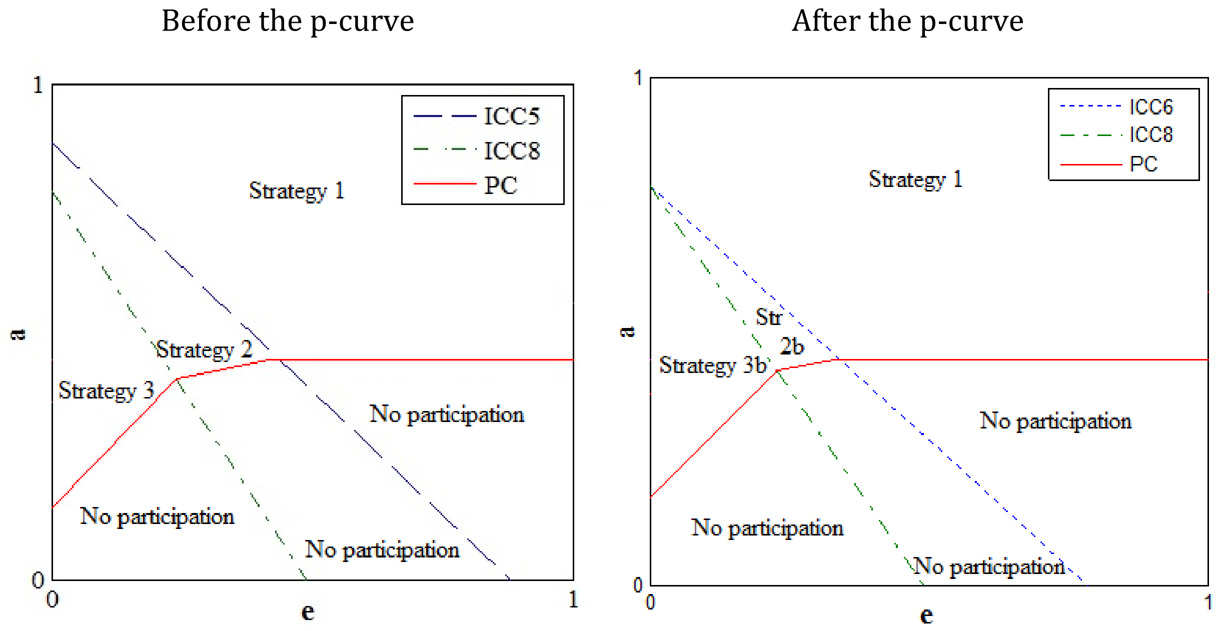
Also the selection effects are similar to those described in Propositions 7 and 8, but this time, the participation constraints of fraudulent agents tighten more, as the parallel shift in expected utility for them is now larger: $m\left(c_f - c_h\right)S > m(c_h' - c_h)S$, which could be depicted in Figures 6a and 6b by a larger shift downwards of the utility of the optimal strategy. Because the shift is larger, the possible changes in e_a^* and a_e^* are larger than in the case that the p-curve's effect is modest. Figure 9 illustrates the equilibrium when the p-curve's effect is strong. Notice the similarity with Figure 7, except that the lines shift a bit further and the optimal strategies change.[46]

Notice that any increase in c_h further than c_f does not change the behavior of agents anymore, since p-hacking is not part of the optimal strategy described in Table 3 anymore. Also, the participation of agents is no longer affected[47]. Therefore, further efforts to combat fraud can in this case best be pointed at eliminating data fabrication rather than p-hacking.

[46] Readers interested in viewing the shift in one figure can refer back to Figure 7, while keeping in mind that the shift of the lines in this case is larger.

[47] This relies on the assumption that the p-curve does not affect c_f. If it does, this result is somewhat weaker.

Figure 9


Before the p-curve

After the p-curve

Notice that ICC6 and ICC8 now coincide when $e = 0$. This is because these agents do not care intrinsically about committing fraud, they only care about its extrinsic consequences. Since the chance to get caught data fabricating is assumed not to vary in the insignificance of results, the extrinsic consequences of data fabricating marginally insignificant results are equal to that of data fabricating highly insignificant results. Therefore, agents with $e = 0$ either choose Strategy 1 or Strategy 3b. "Mediocre Strategy" 2b is never chosen. Therefore, the p-curve induces polarization in the decision whether or not to commit fraud for these agents. This polarization is predicted for agents with $e > 0$ in one specific case. Particularly, if agents are *insensitive to scope*[48] regarding fraud ($\Xi_m = \Xi_I$), ICC6 and ICC8 coincide for all agents, given that the effect of the p-curve is profound. In that case, agents only care about whether or not they commit fraud, not about the severity of it. Then, if they want to misreport marginally insignificant results, they also want to misreport highly insignificant results. In Figure 9, insensitivity to scope is illustrated by ICC5/ICC6 and ICC8 being parallel to one another. The steeper ICC8 is compared to ICC5/ICC6, the more agents care for the scope of fraud. This paragraph is summarized in Proposition 12:

> **<u>PROPOSITION 12:</u>** The introduction of the p-curve has the following "polarization" effects, given that the effect of the p-curve on the probability to detect p-hacking is profound and given that the participation constraint is satisfied:

[48] Stigler (1970), for instance, finds that there is a belief that "moral guilt does not vary closely with the size of the offense" (pp. 534).

1) Agents with integrity level $e = 0$ choose from two strategies: one that dictates to report all results honestly and one that dictates to misreport all insignificant results.
2) Prediction 1) also counts for agents with integrity level $e > 0$ if and only if they are insensitive to scope: $\Xi_m = \Xi_I$.

   This Proposition only holds if c_f is independent of the p-value of the initial result.

Overall result

All in all, introducing the p-curve leads to beneficial incentive effects, i.e. less agents commit fraud. Particularly, the ones that have mediocre integrity and ability start to report all their results honestly, also the marginally insignificant results which they used to p-hack. A potentially beneficial selection effect of the p-curve is fraudsters leave the world of science. The fraudsters that leave are the ones with the highest level of integrity and lowest level of ability. The question is, however, whether this is attractive in terms of social welfare, since these agents might be more destructive in other occupations, and having replacements fill in the vacancy of researcher might be costly in terms of adjustment. If the effect of the p-curve on the probability to get caught p-hacking is profound, the incentive and selection effects are larger. However, the size of these effects is bounded, since agents have the opportunity to switch to other types of fraud if the p-curve results in a large increase in the probability to detect p-hacking.

7. Model extensions

A. *Optimal investment in fraud combating technologies*

Suppose that the government has to decide how much to invest in two fraud combating technologies. ρ is the selected amount of funds allocated to the p-curve, which increases c_h, and φ the amount to be invested in some technology that increases c_f. Every dollar invested in the p-curve increases c_h by a smaller amount:$dc_h/d\rho > 0$, $d^2{c_h}^2/d\rho < 0$. The same counts for investments in the other technology: $dc_f/d\varphi > 0$, $d^2{c_f}^2/d\varphi < 0$. Increases in c_h benefits society by factor P.[49] However, it was shown before that the marginal social benefit of increasing c_h is zero when $c_h \geq c_f$, since all fraudsters will have switched from p-hacking towards data fabricating, making any effort to combat p-hacking pointless. On the other hand, increasing c_f beyond the level of c_h does benefit society (by factor f), since there might be

[49] P, f and F may vary in c_h and c_f, but assuming that they do not can be done without loss of generality.

some agents that data fabricate highly insignificant p-values, no matter how c_h relates to c_f. Still, the benefits of increasing c_f are higher (factor F) when there are agents that also data fabricate marginally insignificant results, which occurs when $c_h \geq c_f$. This is illustrated in Figure 10, where c_f is kept fixed at c_f^1 in graph a and c_h at c_h^1 in graph b. The dotted line in graph c shows the case when one invests more in c_f, up to $c_f^2 > c_f^1$. Graph d shows a similar case for a larger investment in c_h.

Figure 10a

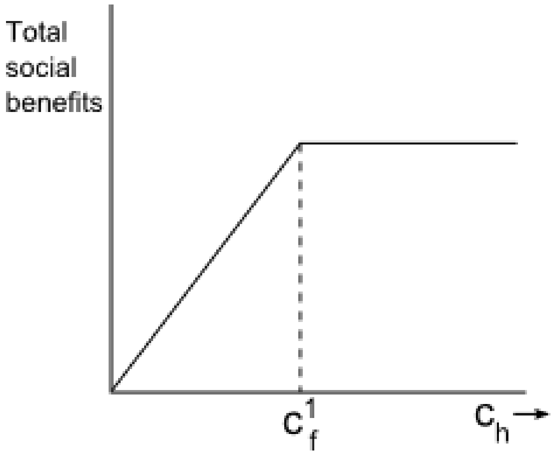


Figure 10b

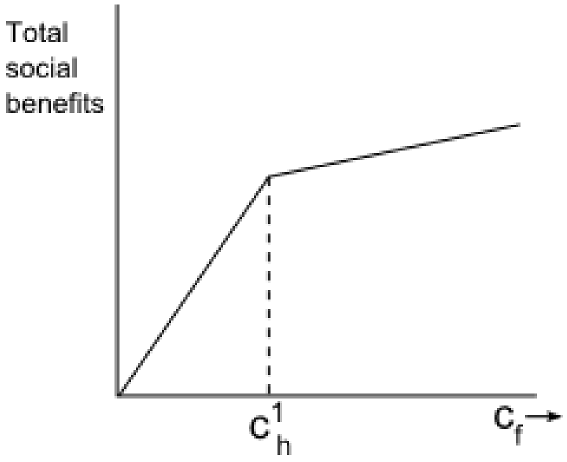


Figure 10c

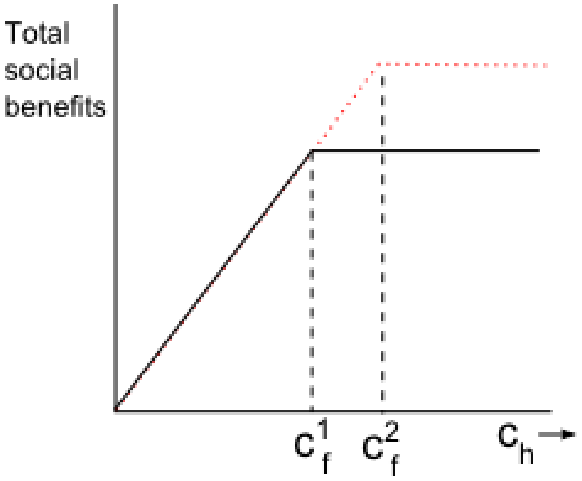
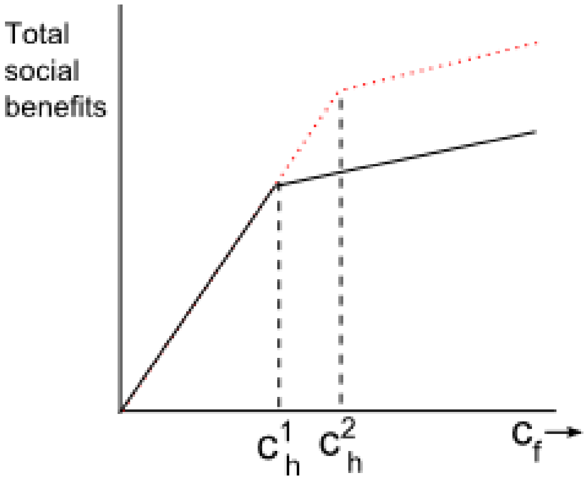


Figure 10d



The graphs show that investments in c_h and c_f are complementary. An increase in $c_f(c_h)$ does not only result in societal benefits directly, it also yields benefits to society indirectly by making investment in $c_h(c_f)$ possibly more profitable. This is also shown when the social welfare function SW is maximized:[50]

$$SW = (P + F) \cdot \min\{c_h, c_f\} + f \cdot \left(\max\{c_f - c_h, 0\}\right) - \rho - \varphi \qquad (18)$$

[50] Inserting a budget constraint into the analysis yields qualitatively similar results.

The first-order derivatives of (23) are presented below:

$$\frac{\partial SW}{\partial \rho} = \begin{cases} (P+F)\cdot\frac{dc_h}{d\rho} - f\cdot\frac{dc_h}{d\rho} - 1 & \text{if } c_h < c_f \\ -1 & \text{if } c_h \geq c_f \end{cases} \tag{19}$$

$$\frac{\partial SW}{\partial \varphi} = \begin{cases} (P+F)\cdot\frac{dc_f}{d\varphi} - 1 & \text{if } c_f < c_h \\ f\cdot\frac{dc_f}{d\varphi} - 1 & \text{if } c_f \geq c_h \end{cases} \tag{20}$$

Investing in the development of the p-curve can only be profitable if $c_h < c_f$. Therefore, in the optimum denoted by stars, $c_h^* \leq c_f^*$ must hold. If investment in technology to combat data fabrication is costly ($dc_f/d\varphi$ is low) or not very beneficial to society (f is low), $c_h^* = c_f^*$. Whether or not that is the case, it is clear that developing the p-curve cannot be seen in isolation to developing instruments to combat data fabrication.

B. Adjusting the remuneration structure

As discussed before, at the heart of the fraud issue lies the remuneration structure in the academic world that is built around journal publications. A quick peak back at ICC5, ICC6 and ICC8 reveals that $(J - W)$, the difference in utility between academic and working paper publications, is an important determinant in the decision whether or not to commit fraud also in the model of this paper. A policy maker could argue that reducing the difference between J and W is a good idea to combat the incidence of fraud in the scientific world. W is considered not influenceable for policy makers, as no payment can credibly be tied to working papers. J, on the other hand, can be decreased if desired, for instance by reducing the degree to which promotions are tied to academic journal publications. In that case, the incentive compatibility constraints will loosen, resulting in less agents opting for fraud.

The downside is that the participation constraint will tighten, leading some agents to opt out of science. Particularly, agents with low ability and high integrity will leave, provided that $d\bar{U}/da < (J - W) * dp/da$ will continue to hold. But, if J is lowered too much, this condition, which ensures that participation increases in research ability, no longer holds. If the decrease in J is large, it might even be that participation will unambiguously decrease in research ability[51]. This implies that it is not the least able agents, but the highly able agents who are most likely to opt out of science. Either way, in order to prevent an exodus of

[51] This occurs if $d\bar{U}/da > (J - W)[s * dp/da + (1 - s) * d\pi/da]$ holds.

researchers into other occupations, the base wage w will have to be raised and other adjustment costs will have to be incurred.

Perhaps more importantly in this case, a high level of J can be a good incentive for researchers to exert a high amount of effort. To see why, suppose that researchers have to exert costly effort η before they observe the significance of their results. Exerting a higher level of effort results in a more creative research topic (effort can be exerted by consulting more literature, spending more time discussing with other academics, etcetera). Therefore, p and π increase not only in ability, but also in research effort. In particular, assume that effort yields larger research creativity improvements especially for high ability individuals, which is not unrealistic, since more able people can more easily understand other people's ideas and transfer those ideas into new inquiries. A simple representative case of the narrative above could be:

$$p = p(0) + ba + ga\eta \tag{21}$$

$$\pi = \beta a + \gamma a\eta \tag{22}$$

All the parameters in (21) and (22) are nonnegative and defined in such a way that $0 \leq \pi \leq p \leq 1$ holds. Recall the assumption that for more creative researches, p is closer to π. This implies that $\gamma > g$ must hold. Equation (13), with effort costs equal to $\eta^2/2$ added, is the function to be maximized. Taking simple first-order conditions for each of the three optimal strategies X in equation (13) yields optimal level of effort η^*:[52]

$$\eta^* = \begin{cases} sga(J-W) + (1-s)\gamma a(J-W) & \text{if } X = 1 \\ (s+m)ga(J-W) + (1-s-m)\gamma a(J-W) & \text{if } X = 2 \\ ga(J-W) & \text{if } X = 3 \end{cases} \tag{23}$$

Lowering J affects the optimal level of effort as described in equation (22):

$$-\frac{\partial\eta^*}{\partial J} = \begin{cases} -sga - (1-s)\gamma a & \text{if } X = 1 \\ -(s+m)ga - (1-s-m)\gamma a & \text{if } X = 2 \\ -ga & \text{if } X = 3 \end{cases} \tag{24}$$

Thus, lowering the promotions and payments tied to academic publications will result in lower effort, no matter which optimal strategy agents opt for. This holds particularly for high ability agents: $\partial^2\eta^*/\partial J\partial a > 0$. Because effort for them yields a relatively high increase in the

[52] How effort affects the fraud and participation decisions is abstracted from. Also, the effect of the p-curve is not analyzed for the case that effort is included in the model.

chance to get a publication, a decrease in the reward for a publication harms them more. Less able agents would not get the reward anyway, so lowering it does not change their incentives much. A second reason why high ability agents lower their effort more is because they choose strategy $X = 1$ (honest) more often. Notice that $\partial\eta^*/\partial J$ is the largest for strategy $X = 1$. For this strategy, effort (resulting in creative research) is important in securing publication, while other strategies rely less on effort and more on (artificial) significance to convince journals to publish results. This implies that the optimal effort level is the highest and most responsive to changes in the reward for publication when choosing strategy $X = 1$.

The result of a decrease of J is thus an overall decrease in research quality, particularly at the high end of the ability distribution. In combination with possibly undesirable selection effects and adjustment costs, it therefore remains to be seen whether altering the payment structure is a good idea to combat fraud. However, it is important to keep in mind that *increasing* J, in combination with a decrease in w, might in certain cases be a useful strategy to combat fraud as well. This new remuneration structure could force agents with little ability out of science, who are the ones that are most likely to commit fraud. This selection effect must be added to the incentive effects that dictate that the agents who participate exert a higher amount of effort, but are also more likely to commit fraud.

C. *Ability required to p-hack*

In the main text, it was implicitly assumed that the act of p-hacking can be executed by everyone without any extra effort. It is, however, conceivable that p-hacking requires a certain econometric knowledge to be effectively applied. In this section, I will model this possibility by assuming that "p-hacking costs" equal $\vartheta/a > 0$.[53] It is easy to see that these costs decrease in research ability, which is a proxy for the degree of econometric knowledge.[54] I will only discuss the decision that agents take when the results are marginally insignificant, since the decisions taken at other levels of significance are not affected.

To infer the behavior of agents in this extension, one should first figure out whether agents prefer to p-hack or to data fabricate. In the main text model, agents always preferred p-hacking before the p-curve was introduced, according to condition (7). But now, this no

[53] $a = 0$ is now defined as the ability level where $\pi = 1$ *and* where the p-hacking cost is infinite. This loss of generality is incurred deliberately for illustrative purposes.

[54] The ability to data fabricate, which involves making up numbers, could be related to ability as well. But since this technique is less "sophisticated" than p-hacking, it can be assumed without generality that data fabrication is costless.

longer holds. The new condition that indicates which agents prefer p-hacking over data fabrication is as follows:

$$a \geq \hat{a} = \frac{c_h}{c_f} + \frac{\vartheta}{c_f S} \tag{25}$$

So, only agents with a sufficiently high ability level prefer to p-hack. This is intuitive, since p-hacking is not so costly for the highly able. The first term on the right indicates that more agents prefer p-hacking when p-hacking is more covert than data fabrication. The rightmost term states that less people p-hack when the p-hacking costs increase, which is also logical. When condition (25) does not hold, one should compare data fabricating to reporting honestly, using ICC6. Recall from the analysis section that agents with higher ability choose to data fabricate less often. This criterion still holds here. So, for the agents for whom $a \geq \hat{a}$ does not hold, fraud unambiguously decreases in a. If (25) does hold, agents either p-hack or report honestly. ICC26, the new version of ICC5, is the relevant criterion to assess this. It states when agents choose to report honestly, given that (25) holds:

$$(J - W)(p - \pi) \leq \ c_h S + \Xi_M + \frac{\vartheta}{a} \tag{ICC26}$$

In the analysis section, p-hacking loses attractiveness if a increases. However, ICC26 indicates that the increase can also result in more p-hacking, because not only the benefits of p-hacking (left-hand side of ICC26), but also the costs of p-hacking decrease in a. The net effect of a will depend on how ability relates to the publication probabilities ($d\pi/da$ and dp/da) and on ϑ.

Recall that a is defined in such a way that $p(a = 1) = \pi(a = 1)$. Therefore, the prediction that agents with $a = 1$ never commit fraud still holds, even though p-hacking is cheapest for them. What happens to the other agents is less clear. Known is that agents just below $\hat{a}$ are less likely to commit fraud (data fabrication) than agents with zero ability. But it is unknown whether agents just above $\hat{a}$ are more likely to commit fraud (p-hack) than those close to $a = 1$. For instance, when p-hacking is quite costly, a wave shaped function can arise: agents with little ability data fabricate, those with mediocre ability report honestly, those with quite high ability p-hack, and the most able choose to report honestly. If data fabrication was prohibitively costly, an inverted U-shape may arise: the low and high ability types p-hack little, while the mediocre types p-hack more often. Either way, a peak in fraud at medium or high levels of ability cannot be explained by the model used in the main text, and could point at a relation between research ability and the ability to p-hack. The incentive and selection effects of the p-curve will be largest around that peak.

8. Conclusion

This paper examined the benefits and limitations of the p-curve in combating fraud in science. In order to do so, an economic-theoretical model was constructed in which agents had the option to report their statistically insignificant results honestly, or to commit fraud either by p-hacking or by data fabrication. Committing fraud here means reporting statistically insignificant results as significant, in order to increase the odds of academic publication. P-hacking is the type of fraud the p-curve is meant to unmask.

The model predicts that the incidence of fraud decreases in researcher ability and ethical integrity. Mild fraudsters, meaning those with mediocre able and integrity, are the ones that are willing to commit fraud only when the results are marginally insignificant. The p-curve is predicted to correct the incentives of a part of this group, so that they will start to prefer honest reporting. In a special case, the incentives of this whole group will be corrected. Particularly, this happens when the p-curve is very effective at tracking p-hackers and agents are insensitive to the scope of fraud. Whatever the case, the incentives of heavy fraudsters, whose integrity and ability are so low that they are willing to misreport any insignificant result, cannot be corrected by the p-curve. The selection effect of the p-curve does affect all fraudsters. Particularly, those fraudsters with the lowest ability and highest integrity will leave science if the p-curve is introduced.

Taking these considerations into account, one might be tempted to believe that the p-curve is beneficial to society. However, a preliminary welfare analysis shows that the p-curve may destroy welfare. In particular, if fraud is considered not so costly to society and if the adjustment costs that need to be incurred to re-match agents with employers are large, society might be better off without the p-curve. In fact, one cannot also say with certainty that the p-curve is effective at reducing scientific fraud, since the replacements of leaving scientists might be more inclined to commit fraud than those they came to replace. Therefore, this paper cannot definitively conclude in favor or against the introduction of the p-curve. The value of the analyses lies in revealing its benefits and limitations, thereby facilitating the discussion. It is up to future research to conclude whichever of the two is largest.

Next to the above analyses, it was found that the p-curve must not be developed in isolation to methods that combat other forms of fraud. The reason is that fraudsters have the option to switch from p-hacking towards data fabrication if the p-curve is very effective at unmasking p-hackers. If they do switch, developing the p-curve further will be ineffective. Instead, the attention should then be focused at ways to combat data fabrication. But developing only those techniques will induce researchers to switch back to p-hacking again. Therefore, developing the p-curve is a complement, not a substitute, to developing other fraud combating techniques. Also without the complementarity, developing other techniques might be a good idea, since data fabrication likely is more destructive to society than p-hacking.

The analyses in this paper revealed some other interesting propositions as well. For example, the p-curve was found to reduce the possibility of highly able researchers to signal their ability to potential employers through showing good ethics. This is because the positive correlation between expected ability and integrity conditional on participation was stronger before than after the p-curve. The final contribution of this paper's model is a discussion on the possible implications of altering the remuneration structure in science.

One has to keep in mind that the results presented above are all contingent on the validity of the made assumptions. For instance, as shown in the extensions, if the ability to p-hack varies in research ability, the most basic predictions of the model already change. Other simplifying assumptions probably limited the generality of the results. For instance, agents were assumed to be harmed only when they themselves commit fraud and ability was confined to "researcher ability" only. Second, it was assumed that there are no general equilibrium effects induced by the p-curve. However, it is conceivable that a journal publication is worth more after the p-curve, since, in expected terms, the results presented in it are more accurate. Similarly, the publication probabilities may change, knowing that significant results are more often real than before. A third limitation is that it could not be assessed whether the outside option of researchers in- or decreases in ethical integrity, since there are reasons to argue for both. The implicit assumption was that these opposite effects cancel each other out, but reality might prove different. Lastly, the possibility that p-hacking is path dependent is ignored. If a researcher's p-curve has many p-hacked p-values included, adding one new, honest p-value might not alter the conclusion that the researcher has p-hacked. Then, the researcher might as well continue p-hacking, since the marginal costs of it are zero anyway.

This paper has already pointed at the necessity of future research in assessing the welfare effects of the p-curve. Other fruitful avenues for further research could be to endogenize the choice of statistical power, since the p-curve is less effective at proving p-hacking for studies with higher power. Additionally, file drawering is an option for researchers when their results turn out insignificant, so taking this possibility into consideration is advocated. As already touched upon in the extensions section of this paper, effort choice by researchers could also be included in further analyses, since the payoffs to committing fraud interact with it. Then, we hopefully will be able to assess whether *the p-curve: a key to the file drawer* was indeed the key we were looking for.

9. References

Abma, R. (2013) *De publicatiefabriek*: *over de betekenis van de affaire-Stapel*. Nijmegen: Uitgeverij Vantilt.

Aghion, P. and P. Howitt (1990) A model of growth through creative destruction. *NBER working paper*, 3223.

Al-Marzouki, S., S. Evans, T. Marshall and I. Roberts (2005) Are these data real? Statistical methods for the detection of data fabrication in clinical trials. *Bmj:* 331, 7511, pp. 267–270.

Anselin, L., A. Varga and Z. Acs (1997) Local geographic spillovers between university research and high technology innovations. *Journal of urban economics:* 42, pp. 422–448 1997.

Angell, M. (1989) Negative studies. *New England journal of medicine:* 321, 7, pp. 464–466.

Ashenfelter, O., C. Harmon and H. Oosterbeek (1999) A review of estimates of the schooling/earnings relationship, with tests for publication bias. *Labour economics:* 6, 4, pp. 453–470.

Bakan, D. (1966) The test of significance in psychological research. *Psychological bulletin:* 66, 6, pp. 423–437.

Becker, G.S. (1968) Crime and punishment: an economic approach. *Journal of political economy:* 76, 2, pp. 169–217.

Begg, C.B. and M. Mazumdar (1994) Operating characteristics of a rank correlation test for publication bias. *Biometrics:* 50, 4, pp. 1088–1101.

Bossuyt, P.M., J.B. Reitsma, D.E. Bruns, C.A. Gatsonis, P.P. Glasziou, L.M. Irwig, ... and H.C. de Vet (2003). Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Clinical chemistry and laboratory medicine:* 41, 1, pp. 68–73.

Brodeur, A.,M. Lé, M. Sangnier and Y. Zylberberg (2013) Star wars: the empirics strike back. *IZA discussion paper*, 7268.

Cameron, S. (1988) The economics of crime deterrence: a survey of theory and evidence. *Kyklos*: 41, 2, pp. 301–323.

Carver, R.P. (1978) The case of significance testing. *Harvard educational review:* 48, 3, pp. 378–399.

Cohen, J. (1994) The earth is round (p < .05). *American psychologist:* 49, 12, pp. 997–1003.

Dasgupta, P. and P.A. David (1987). Information disclosure and the economics of science and technology. In: Feiwel, G.R. (Ed.) *Arrow and the ascent of modern economic theory*, pp. 519–542. New York: NYU press.

Dasgupta, P. (1989) The economics of parallel research. In: Frank, H. (Ed.) *The economics of missing markets, information, and games,* pp. 129–148. Oxford: Clarendon press.

Dickersin, K. (1990) The existence of publication bias and risk factors for its occurrence. *JAMA:* 263, 10, pp. 1385–1389.

Dickersin, K. (2005) Publication bias: recognizing the problem, understanding its origins and scope, and preventing harm. In: Rothstein, H.R., A.J. Sutton and M. Borenstein (Ed.) *Publication bias in meta-analysis: prevention, assessment and adjustments*, pp. 11–34. West Sussex: John Wiley and sons ltd.

Dills, A.K., G.A. Miron and G. Summers (2008) What do economists know about crime? *NBER working paper*, 13759.

Duval, S. and R. Tweedie (2000) Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics:* 56, pp. 455–463.

Dwan, K., D.G. Altman, J.A. Arnaiz, J. Bloom, A-W. Chan, E. Cronin, … and R. Williamson (2008) Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLOS one:* 3, 8, e3081.

Easterbrook, P.J., J.A. Berlin, R. Gopalan and D.R. Matthews (1991) Publication bias in clinical trials. *Lancet:* 377, 8746, pp. 867–872.

Egger, M., G.D. Smith and C. Minder (1997) Bias in meta-analysis detected by a simple, graphical test. *BMJ:* 315, 7109, pp. 629–634.

Ehrlich, I. (1973) Participation in illegitimate activities: a theoretical and empirical investigation. *Journal of political economy:* 81, 3, pp. 521–565.

Elm, E. von, D.G. Altman, M. Egger, S.J. Pocock, P.C. Gøtzsche, and J.P. Vandenbroucke (2007) The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *Preventive medicine*: 45, 4, pp. 247–251.

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

Feigenbaum, S. and D.M. Levy (1993) The market for (ir)reproducible econometrics. *Social epistemology:* 7, 3, pp. 215–232.

Freeman, R.B. (1999) The economics of crime. In: Ashenfelter, O.C. and D. Card (Ed.) *Handbook of labour economics*, pp. 3530–3571. Amsterdam: North Holland.

Freiman, J.A., T.C. Chalmers, H. Smith jr., R.R. Kuebler (1978) The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. Survey of 71 "negative" trials. *New England journal of medicine:* 299, 13, pp. 690–694.

Garoupa, N. (1997) The theory of optimal law enforcement. *Journal of economic surveys:* 11, 3, pp. 267–295.

Garoupa, N. (2003) Behavioral economic analysis of crime: a critical review. *European journal of law and economics:* 15, pp. 5–15.

Gladbury, G.L. and D.B. Allison (2012) Inappropriate fiddling with statistical analyses to obtain a desirable p-value: tests to detect its presence in published literature. *PLOS one:* 7, 10, e46363.

Gottfredson, M.R. and T. Hirschi (1990). *A general theory of crime*. Stanford: Stanford university press.

Graves, P.E., J.M. Marchand and R. Thompson (1982) Economics departmental rankings: research incentives, constraints, and efficiency. *American economic review:* 72, 5, pp. 1131–1141.

Heckman, J.J. (1979) Sample selection bias as a specification error. *Econometrica:* 47, 1, pp. 153–161.

Hedges, L.V. (1992) Modeling publication selection effects in meta-analysis. *Statistical science:* 7, 2, pp. 246–255.

Henry, E. (2009) Strategic disclosure of research results: the cost of proving your honesty. *Economic journal:* 119, pp. 1036–1064.

Ioannidis, J.P.A., S.J, Evans, P.C. Gøtzsche, R.T. O'neill, D.G. Altman, K. Schulz and D. Moher (2004) Better reporting of harms in randomized trials: an extension of the CONSORT statement. *Annals of internal medicine:* 141, 10, pp. 781–788.

Ioannidis, J.P.A. (2005) Why most published research findings are false. *PLOS medicine:* 2, 8, e124.

Ioannidis, J.P.A. and T.A. Trikalinos (2007) An exploratory test for an excess of significant findings. *Clinical trials:* 4, 3, pp. 245–253.

Ioannidis, J.P.A. (2008) Why most discovered true associations are inflated. *Epidemiology:* 19, 5, pp. 640–648.

Ioannidis, J.P.A. (2012) Why science is not necessarily self-correcting. *Perspectives on psychological science:* 7, 6, pp. 645–654.

John, L.K., G. Loewenstein, D. Prelec (2012) Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science:* 23, 5, pp. 524–532.

Jovanovic, B. (1982) Truthful disclosure of information. *Bell journal of economics:* 13, 1, pp. 36–44.

Keulemans, M. (2012, July 2) Ontslagen hoogleraar zoog wel cijfers uit zijn duim. *Volkskrant*. Accessed 6 September 2013 from *http://www.volkskrant.nl*.

Kohn, A. (1986) *False prophets: fraud and error in science and medicine*. New York: Basil Blackwell publishers inc.

Lazear E. P. and M. Gibbs (2009) *Personnel economics in practice* (2nd ed.). Hoboken: John Wiley & sons inc.

Levitt, S.D. and T.J. Miles (2006) Economic contributions to the understanding of crime. *Annual review of law and social science:* 2, pp. 147–164.

Lykken, D.T. (1968) Statistical significance in psychological research. *Psychological bulletin:* 70, 3, pp. 151–159.

Mahoney, M.J. (1977) Publication prejudices: an experimental study of confirmatory bias in the peer review system. *Cognitive therapy and research:* 1, 2, pp. 161–175.

Angell, M. (1986) Publish or perish: a proposal. *Annals of internal medicine:* 104, 2, pp. 261–262.

McGrail, M.R., C.M. Rickard and R. Jones (2006) Publish or perish: a systematic review of interventions to increase academic publication rates. *Higher education research and development:* 25, 1, pp. 19–35.

Merton, R.K. (1942) Science and technology in a democratic order. *Journal of legal and political sociology:* 1, 1–2, pp. 115–126.

Nickerson, R.S. (1998) Confirmation bias: a ubiquitous phenomenon in many guises. *Review of general psychology:* 2, 2, pp. 175–220.

Nickerson, R.S. (2000) Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological methods:* 5, 2, pp. 241–301.

Nosek, B.A., J.R. Spies and M. Motyl (2012) Scientific utopia: ii. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science:* 7, 6, pp. 615–631.

Nutley, S.M., H.T.O. Davies and P.C. Smith (Ed.) (2000) *What works? Evidence-based policy and practice in public services*. Bristol: The policy press.

Orwin, R.G. (1983) A fail-safe n for effect size in meta-analysis. *Journal of educational statistics:* 8, 2, pp. 157–159.

Polinsky, A.M. and S. Shavell (1999) The economic theory of public enforcement of law. *NBER working paper*, 6993.

RePEc (2013) Top 10% authors, as of August 2013. Accessed 6 September 2013 from *http://ideas.repec.org*.

Roediger, H.L. III (2012) Psychology's Woes and a Partial Cure: The Value of Replication. *Observer:* 25, 2, pp. 27–29.

Rond, M. de and A.N. Miller (2005) Publish or perish: bane or boon of academic life? *Journal of management inquiry:* 14, 4, pp. 321–329.

Rosenthal, R. and J. Gaito (1963) The interpretation of levels of significance by psychological researchers. *Journal of psychology:* 55, pp. 33–38.

Rosenthal, R. (1979) The "file drawer problem" and tolerance for null results. *Psychological Bulletin:* 86, 3, pp. 638–641.

Rothstein, H.R., A.J. Sutton and M. Borenstein (Ed.) (2005) *Publication bias in meta-analysis: prevention, assessment and adjustments*. West Sussex: John Wiley and sons ltd.

Rozeboom, W.A. (1960) The fallacy of the null-hypothesis significance test. *Psychological Bulletin:* 57, 5, pp. 416–428.

Scargle, J.D. (2000) Publication bias: the "file-drawer" problem in scientific inference. *Journal of scientific exploration:* 14, 1, pp. 91–106.

Schumpeter, J.A. (1950) *Capitalism, socialism, and democracy* (3rd ed.). New York: Harper and brothers.

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

Simonsohn, U. (2013) *Just post it: the lesson from two cases of fabricated data detected by statistics alone.* Accessed 16 September 2013 from http://papers.ssrn.com.

Simonsohn, U., L.D. Nelson and J.P. Simmons (2013) *Supplementary materials – P-curve: a key to the file drawer.* Accessed 10 September 2013 from http://www.p-curve.com.

Simonsohn, U., L.D. Nelson and J.P. Simmons (forthcoming) P-curve: a key to the file drawer. *Journal of experimental psychology: general.*

Spence, M. (1973). Job market signaling. *The quarterly journal of economics:* 87, 3, pp. 355–374.

Stephan, P.E. (1996) The economics of science. *Journal of economic literature*: 34, 3, pp. 1199–1235.

Sterling, T.D. (1959) Publication decisions and their possible effects on inferences drawn from tests of significance – Or vice versa. *Journal of the American statistical association:* 54, 285, pp. 30–34.

Stigler, G.J. (1970) The optimum enforcement of laws. *Journal of political economy:* 78, 3, pp. 526–536.

Tabelleni, G. (2008) The scope of cooperation: values and incentives. *Quarterly journal of economics:* 123, 3, pp. 905–950.

Thornton, A. and P. Lee (2000) Publication bias in meta-analysis: its causes and consequences. *Journal of clinical epidemiology:* 53, pp. 207–216.

Toedter, K-H. (2009) Benford's law as an indicator of fraud in economics. *German economic review:* 10, 3, pp. 339–351.

Waldman, M. (1990) Up-or-out contracts: a signaling perspective. *Journal of labor economics:* 8, 2, pp. 230–250.

Wallis, W.A. (1942). Compounding probabilities from independent significance tests. *Econometrica:* 10, 3–4, pp. 229–248.

Weiss, B. and M. Wagner (2011) The identification and prevention of publication bias in the social sciences and economics. *Journal of economics and statistics:* 231, 5–6, pp. 661–684.

Young N.S., J.P.A. Ioannidis, O. Al-Ubaydli (2008) Why current publication practices may distort science. *PLoS Medicine:* 5, 10, e201.

10. Appendix

Appendix A: Proof of Proposition 4

Proposition to be proven: The expected level of integrity is higher in a group of scientists with higher research ability.

Suppose e_x^* is the "participation threshold" for $a = x$ and e_y^* that for $a = y = x{+}\in$. The expected integrity levels associated with these ability levels are, conditional on participation, equal to $E(e|a = x, e \leq e_x^*)$ and $E\left(e\middle|a = y, e \leq e_y^*\right)$, respectively. Thus, the Proposition in mathematical terms is:

$$E\left(e\middle|a = y, e \leq e_y^*\right) \geq E(e|a = x, e \leq e_x^*) \tag{A1}$$

First, let us rewrite (A1):

$$\tau E(e|a = y, e \leq e_x^*) + (1 - \tau)E\left(e\middle|a = y, e_x^* \leq e \leq e_y^*\right) \geq E(e|a = x, e \leq e_x^*)$$

Here, τ is defined as $\Pr(e \leq e_x^*|a = y)/\Pr(e \leq e_y^*|a = y)$.

Now, recall the assumption that the distributions of a and e are independent: $E(e|a = x) = E(e|a = y) = E(e)$. This implies that $E(e|a = x, e \leq e_x^*) = E(e|a = y, e \leq e_x^*) = E(e|\ e \leq e_x^*)$: the expected level of ethical integrity conditional on $e \leq e_x^*$ being true does not depend on ability level. The same reasoning applies for $E\left(e\middle|a = y, e_x^* \leq e \leq e_y^*\right)$, which is equal to $E\left(e\middle|e_x^* \leq e \leq e_y^*\right)$. Then, (A1) becomes:

$$\tau E(e|e \leq e_x^*) + (1 - \tau)E\left(e\middle|e_x^* \leq e \leq e_y^*\right) \geq E(e|e \leq e_x^*)$$

$E\left(e\middle|e_x^* \leq e \leq e_y^*\right)$ is greater than $E(e|e \leq e_x^*)$, so (A1) always holds.■

Proposition to be proven: The expected level of ability is lower amongst scientists with little ethical integrity.

By the same reasoning, a_s^* is the participation threshold for $e = s$ and a_t^* that for $e = t = s{+}\in$. Then, the Proposition to be proven is:

$$E(a|e = s, a \leq a_s^*) \leq E(a|e = t, a \leq a_t^*) \tag{A2}$$

Rewriting yields:

$$\omega E(a|e=t, a \leq a_s^*) + (1-\omega)E(a|e=t, a_s^* \leq a \leq a_t^*) \geq E(a|e=s, a \leq a_s^*)$$

ω equals $\Pr(a \leq a_s^* | a=t)/\Pr(a \leq a_t^* | a=t)$.

$E(a|e=s) = E(a|e=t) = E(a)$ holds, so $E(a|e=s, a \leq a_s^*) = E(a|e=t, a \leq a_s^*) = E(a|a \leq a_s^*)$ and $E(a|e=t, a_s^* \leq a \leq a_t^*) = E(a|a_s^* \leq a \leq a_t^*)$ also hold. (A2) becomes:

$$\omega E(a|a \leq a_s^*) + (1-\omega)E(a|a_s^* \leq a \leq a_t^*) \geq E(a|a \leq a_s^*)$$

$E(a|a_s^* \leq a \leq a_t^*) \geq E(a|a \leq a_s^*)$ by definition, so (A2) always holds.■

Appendix B: Parameter cases for Figures 4, 7 and 9

For all the Figures that describe the equilibrium strategies depending on a and e, I used the following (linear) model specification:

$$p(a) = p(0) + \Delta a \tag{A3}$$

$$\pi(a) = [p(0) + \Delta]a \tag{A4}$$

$$\Xi_m(e) = \xi_m e \tag{A5}$$

$$\Xi_I(e) = \xi_I e, \text{ where } \xi_I \geq \xi_m \tag{A6}$$

$$\bar{U} = \bar{w} + ka \tag{A7}$$

Using this specification, incentive compatibility constraint (ICC5) becomes:[55]

$$a \geq 1 - \frac{c_h S + \xi_m e}{p(0)(J-W)} \tag{A8}$$

And ICC8 can be represented by:

$$a \geq 1 - \frac{c_f S + \xi_I e}{p(0)(J-W)} \tag{A9}$$

The participation constraint (PC14) consists of three different equations, one for each of the optimal Strategies. For Strategy 1, PC14 equals:

$$a \geq \frac{(w-\bar{w}) + s[p(0)J + \{1-p(0)\}W] + (1-s)W}{k - s\Delta(J-W) - (1-s)[p(0)+\Delta](J-W)} \tag{A10}$$

For Strategy 2, the participation constraint equals:

[55] Choosing to represent the constraints in terms of e yields the same results.

$$a \geq \frac{(w-\bar{w}) + (s+m)[p(0)J + \{1-p(0)\}W] + (1-s-m)W - m(c_h S + \xi_m e)}{k - (s+m)\Delta(J-W) - (1-s-m)[p(0)+\Delta](J-W)} \tag{A11}$$

For Strategy 3, one participates if and only if:

$$a \geq \frac{(w-\bar{w}) + p(0)J + [1-p(0)]W - m(c_h S + \xi_m e) - (1-s-m)\left(c_f S + \xi_I e\right)}{k - \Delta(J-W)} \tag{A12}$$

Figure 4 displays the output that arises when one uses the values described in Table A1 to fill in equations (A8) to (A12):

Table A1

Parameter	Number	Parameter	Number
J	8	$p(0)$	0.1
W	3	Δ	0.5
c_h	0.05	$(w-\bar{w})$	-3.5
c_f	0.09	s	0.2
S	1.2	m	0.3
ξ_m	0.5	k	2
ξ_I	0.9		

The numbers for Figure 7 equals those in Table A1, with the exception that $c'_h = 0.08$.

For Figure 9, one should use the values indicated in Table A1 as well. Figure 9, however, makes use of ICC6 rather than ICC5.

ICC6 equals:

$$a \geq 1 - \frac{c_f S + \xi_m e}{p(0)(J-W)} \tag{A13}$$

For Strategy 2b, PC14 equals:

$$a \geq \frac{(w-\bar{w}) + (s+m)[p(0)J + \{1-p(0)\}W] + (1-s-m)W - m(c_f S + \xi_m e)}{k - (s+m)\Delta(J-W) - (1-s-m)[p(0)+\Delta](J-W)} \tag{A14}$$

For Strategy 3b, one participates if and only if:

$$a \geq \frac{(w-\bar{w}) + p(0)J + [1-p(0)]W - m\left(c_f S + \xi_m e\right) - (1-s-m)\left(c_f S + \xi_I e\right)}{k - \Delta(J-W)} \tag{A15}$$