Lies, Damned Lies and Statistics
The Adverse Incentive Effects of the Publication Bias

Michał Krawczyk
CREED, University of Amsterdam

08.08.2008

PRELIMINARY. PLEASE DO NOT QUOTE WITHOUT PERMISSION
Comments welcome: m.w.krawczyk@va.nl

Abstract. In this project I investigate the use and misuse of *p* values in papers published in five (high-ranked) journals in experimental psychology. I use a data set of over 135 thousand *p* values from more than five thousand papers. I inspect (1) the way in which the *p* values are reported and (2) their distribution. The main findings are following: first, the authors choose the mode of reporting their results in an arbitrary and strategic way, that is, to make them seem more statistically significant than they really are (which is well known to improve the chances for publication). Specifically, they frequently report *p* values "just above" significance thresholds directly, whereas other values are reported by means of inequalities (e.g. "*p*<.1"), they round the *p* values down more eagerly than up and choose between the significance thresholds and between one- and two-sided tests after seeing the data. Further, about 8% of reported *p* values are inconsistent with their underlying statistics (e.g. F or t) and it appears that there are "too many" "just significant" values, suggesting use of data or model manipulation techniques to bring the *p* value to the right side of the threshold.

1. Introduction

Statistics is the corner stone of nearly all empirical science. It provides the essential tools to distinguish between the systematic and the incidental. However, the power of statistics brings with it the temptation to abuse it. Most authors are well aware of the fact that the statistics in their papers (notably: the statistical significance of the effects they are reporting) may substantially affect their publishability ("publication bias") and also their impact. If so, we can expect them to employ what we can call strategic reporting: to include those results and in such a way that favors their theories. And, at the end of continuum of such practices, we may wonder how many scholars are willing to totally misrepresent or fabricate the data to improve their chances for publication.

For the purposes of this study a large data set of statistics (*p* values and otherwise) reported in about 5 thousand papers has been collected and investigated. I find evidence of various types of decisions (e.g. whether or not to round given number and whether to use a one- or a two-sided test) being made strategically, that is, with the aim of presenting the results as "more significant". I also find that in about 8% of the cases the reported *p* value is inconsistent with the underlying statistic (e.g., the F statistic). Finally, there seems to be "too many" "just significant" *p* values, which might have resulted from the use of smallest data manipulations sufficient to bring the value below the threshold.

The remainder of this paper is structured as follows. In section 2 I analyze the publication bias and the perverse incentive it creates for the researchers. I also try to answer the question whether they are likely to give in and if so, whether data manipulation is actually feasible. In section 3 I describe the design of my empirical study, the data set collected, empirical predictions and results. Section 4 contains a brief discussion and some conclusions.

2. Publication bias and its incentive effects

One of the ways in which statistics turns out to generate rather undesirable effects has been called the publication bias (PB) and documented in dozens of studies (see e.g. Gerber et al. 2001, Stanley 2005). We speak of PB when statistical significance of an effect reported in a paper (in other words: whether or not the null hypothesis can be rejected) strongly affects the chances for publication: "negative" results are much more difficult to sell, controlling for other qualities of the study. Of course, failure to reject the null may result from deficiencies of the design or data set. For instance, if the theory from which the alternative hypothesis has been derived is highly implausible in the first place, then the "null result" is not very interesting. Similarly, it may be due to researcher's inability to collect enough data or the data being too noisy (particularly when proxies have to be used instead of the variables of interest). To the extent that it is difficult to establish the exact reasons of the null result in every single case, it is a reasonable strategy to be prejudiced against them. Further, the alternative hypothesis is often more "attractive", such as e.g. in the case of research on functional differentiation of the brain hemispheres.

However, the publication bias has long been recognized as highly undesirable, as it creates a blurred picture of the strength of an effect.[1] In an extreme case, if out of 20 studies investigating some non-existing difference, only, say, one that happens to deliver significance at 5% level will be published and read, we obtain a totally distorted view. This may, for example, lead to costly policies being based on unwarranted finding of effectiveness of some measures. Further, if negative results are not made public, other researchers may waste time and money challenging the same, correct, null hypotheses over and over again.

The publication bias is said to be strengthened by the file-drawer problem (Iyengar and Greenhouse, 1988). Indeed, if researchers realize that their negative results have poor chances to be published, they have little incentive to ever describe them properly and submit to a journal.

I argue, however, that the incentive effects of the publication bias may be more perverse and more severe. With increasing pressure to publish, the researchers may start considering fraudulent behavior (Wolf 1986), including manipulation of their data to assure statistical significance of the findings. The two questions we need to ask are, first, whether they would be willing to do so and, second, whether it is feasible. These are addressed in the next subsections.

2.1 Can we trust the scientists?

Reliability is a crucial if not defining feature of all research activities. We therefore should expect the researchers to be extremely careful in the way they design, conduct and report their studies. However, we have some evidence that this is not always so (Steneck, 2006). To be sure, proving violations of research standards is often difficult (White, 2005). Therefore, we may suspect that a huge majority of such cases go

[1] Many techniques have been developed to correct for the publication bias when aggregating past results, particularly by means of meta-analyzes. See for example Stanley (2005).

unnoticed or at least unpunished. It is therefore even more startling that we do regularly observe cases of large-scale scams being discovered. Some of the most spectacular and well-known recent incidents include the case of physicist J. Hendrik Schön from Bell Laboratories faking experimental data on nanomaterials (see NY Times, June 15, 2004) and biologist Hwang Woo Suk from the Seoul National University who fabricated evidence on the human stem cells (NY Times, Jan 10, 2006). A number of older scams is described by Fox (1994).

To investigate the frequency of such transgressions more systematically, we can turn to the use of surveys. Again, results should be treated with great caution; because admitting violations of the proper standards of scientific conduct involves shame and potentially also serious risk to career prospect and reputation, we may better think of the figures emerging from such studies as a lower bound on the actual frequency.

In a large-scale study of scientists funded by the National Institutes of Health, Martinson et al (2005) find that 0.2-0.3% admit to falsifying research data. Some 11-12% admit "withholding details of methodology or results" and about 15% confess "dropping observations (...) based on a gut feeling (...)"

In the domain of social science, List et al. (2001) present results of a survey of economists present at the January 1998 meeting of the American Economic Association on in Chicago. The study was run in two modes, one of which (the “Randomized Response”) created a possibility for the responders to hide the "improper" answer. The results show that perplexing 4-5% admit to falsifying research data and 7-10% admit to any of four "minor infractions". The responders estimated the proportion of studies in top 30 journals contaminated by these practices at 5-7% and 13-17% respectively.

2.2 How to manipulate your data

Manipulating statistical data[2] may at first sight seem a risky venture, as potential consequences are severe. However, as mentioned before, the odds of actually having to suffer them are low. Suppose now that you run a study testing some H0 against

[2] As the short review provided below shows, I am interested in a broad range of behaviors, some of which are relatively benign and widely tolerated; certainly, some of them may be followed without any "criminal" intentions. "Manipulation" may therefore be too strong a word. Actually, on critical introspection, most readers, and the author himself, may end up admitting being involved in one of these practices in the past. For lack of a better term encompassing all the relevant practices I shall however stick to the term "manipulation".

some H1 and, regrettably, the *p* value of what you considered the most informative test turns out to be, say, .05351. (thus above the standard significance threshold of 5%). What options are left open if you want to sell it as a significant result?

Starting from the most crude ways, you may simply make up a whole new data set. This option is however ethically rather appalling and quite risky if other researchers (including your collaborators) have access to the original data.[3] A somewhat better option seems thus to modify just some entries, e.g. changing the observations that speak against H1, sufficiently many of them to push the *p* value below .05. You may also simply misreport the relevant statistic (or just the *p* value associated with it).

Further, it is possible to modify the hypothesis after seeing the data. If originally the effect was expected to be there for all subjects (but it turns out not to be the case), after reconsideration you may come to a conclusion the really appropriate test of the theory is whether it shows for, say, females, risk-averters or individuals who score highly on the neuroticism scale (depending on the kind of data you have and in which group the effect turned out to be greatest).

Econometrics provides a myriad of ways: run a tobit or a probit, include or exclude outliers, include more or less control variables etc, always choosing the option that yields more significant results. In Ronald Coase's words, “if you torture the data long enough, nature will confess”. This problem has long been recognized, of course, and readers have learnt to take the authors' assertions with a pinch of salt. As Leamer (1983) puts it, "(...) in fact, all the concepts of traditional theory, utterly lose their meaning by the time an applied researcher pulls from the bramble of computer output the one thorn of a model he likes best, the one he chooses to portray as a rose. The consuming public is hardly fooled by this chicanery".

When field data is used, it is often necessary to use proxies for the variable of interest. The author may then report the one that "works best" where two or more were initially considered.

The possibilities seem to be even richer for experimentalists. There is anecdotal evidence of researchers reporting only some of the experimental sessions, all of which, it would seem, provide relevant data. More fundamentally, especially in experimental economics it is often not well-defined what constitutes an "experiment"

[3] It appears thus that "cooking" the data may be a workable strategy in the case of experimental data and less so when publicly available field data is used.

(Roth 1994). Similarly, it seems that many (if not most) researchers decide whether or not to run additional sessions after seeing the data. While it may be a reasonable strategy (e.g. it seems to make little sense to continue spending money if early sessions already provide very strong results), the researchers should carefully spell it out and calculate what the effect for the reported *p* values is.

Many of the more subtle manipulations mentioned above will typically result in *p* value being just below the significance threshold. Such "minimal" changes are probably more tempting, less ethically appalling, easier to perform and more difficult to detect. It appears therefore promising to focus on the "just significant" results, the main hypothesis being that we will observe “too many” of them.

Even if the case the author is not willing to employ any of these options, there are still choices to be made about the mode in which the value is reported. First, she chooses whether to give the value to the reader directly (“*p*=…”) or by means of an inequality (comparison with some conventional significance threshold). I will call it the choice of sign.

Next, if the author had decided to report directly, she has to choose how many decimals to include. For example she can report her *p* value as equal to .05351, .0535, . 054 or simply .05. I shall refer to this as the choice of precision. It is clear in the example above that reporting just two digits is most appealing, and reporting four seems better than three, as the latter involves the greatest, thus “least-significant” number.

If the author chooses to report the value by means of comparison (inequality), some specific threshold needs to be selected and the sign reported accordingly. In our example, she can for example report “*p*<.1” or “*p*>.05”. This will be called the choice of threshold. Here, the first option appears most attractive.

Related to this, for some tests (e.g. t-tests), it is possible to choose between a one-sided and a two-sided test. This is another way of choosing a suitable threshold, as significance in one-sided test at 5% corresponds to significance at 10% in a two-sided test.

To sum up, the ways to manipulate the data in order to achieve statistical significance are abundant and many of them appear to be relatively safe and perhaps not totally unacceptable from ethical viewpoint. In view of the evidence mentioned in the previous subsection we cannot be sure that the researchers resist the temptation.

3. The study

3.1 Data set

The data set consists of over 135 thousand records. The data have been collected by means of computer-based search from all volumes of the five Journals of Experimental Psychology[4] in the period Jan. 1996-March 2008. This choice of the data set requires some justification.

I need to stress that I do not think that psychological science is somehow particularly prone to practices of data manipulation. There is certainly no reason to purport that psychologists are less reliable or honest than other researchers (even though they do routinely deceive their subjects, some of which later go on to become academic psychologists themselves). However, as suggested in the previous subsection, it appears that experimental investigations provide greater opportunities for the researchers to manipulate the data without having anyone notice it and psychology is inherently an experimental science. Further, such practices may generally be safer in behavioral sciences, where unsuccessful replication is less indicative of the quality of the original study (e.g. may always result from the subject pool specificity). Further, comparing e.g. to medicine, the bad conscience resulting from manipulating the outcomes may be less of a burden in social sciences, where direct and grave consequences of implementation are unusual.

There is also an advantage over the closely-related field of economics. There, the publications are full of multivariate regressions, whereas psychologists tend to focus on (non-parametric) tests of treatment effects. We may expect the practice to manipulate the data to be applied to "central" rather than control variables. If we thus wanted to use the data from economics papers, we would either have to distinguish between these two types of variables in thousands of papers or end up with a lot of noise. Besides, the common habit (actually: requirement put forward by the publishers) in psychology to actually report the *p* values makes the data collection easier. All in all, experimental psychology may give the speculated (subtle) effects the best chance to show up.

[4] These are: Journal of Experimental Psychology: Animal Behavior Processes, Journal of Experimental Psychology: Human Perception and Performance, Journal of Experimental Psychology: General, Journal of Experimental Psychology: Applied and Journal of Experimental Psychology: Learning, Memory, & Cognition.

Regarding the ranking of the journals, it is possible that there is less (detectable) data manipulation going on in the leading ones, as referees and editors are likely to be doing a better job. On the other hand, however, it may take some tricks to make it to a really good journal, whereas the “raw”, negative findings would be good enough for a mediocre outlet. While it may be difficult to compare the prevalence, there is clearly a difference in the implications: it is important to establish whether we can at least trust the statistics reported in top journals.

The data set consists of the following variables: title of the journal, volume, identification number of the paper, a categorical variable showing which sign is used ("=", ">" or "<") and, of course, the *p* value (with precision from .1 to .00001) itself. Further, for some 27 thousand entries I also have the underlying statistic (F, t or χ^2) i.e. its name, degrees of freedom and value – these are only available if they directly precede or follow the *p* value (“(*F*(*x*,*y*)=*z*, *p*=*w*”)[5] For these cases it is possible to re-calculate the *p* value and compare it to the one being reported. More precisely, because the underlying statistics are rounded, I can only calculate the upper and lower end of the range in which the true *p* value lies. Typically, however, this gives a *p* value up to three or more decimals, i.e. more precisely than the authors report themselves (even if the *p* value is reported directly and not by means of “*p*<…”)

The reliability of the search has been verified manually in a number of randomly selected papers. It has been confirmed that, indeed with incidental exceptions, all *p* values and *p* values only (along with underlying statistics) are being collected.

It should be stressed that we do not have any further information on the type of study, sample being used, the content of the hypotheses etc. (E.g. it could even be that some of the entries refer to quoted findings from other papers, though no such case has been found in the process of manual verification.) It is also clear that many statistical results have been overlooked, e.g. those reported in words. This is in stark contrast with earlier studies on issues related to publication bias, where a much smaller sample of publications related to one selected issue would be much more carefully investigated.

[5] Needless to say, there are also many other types of statistics. Each of them, however, appears relatively infrequently.

3.2 Overview – distribution of p *values*

Reported values

We start with presenting the distribution of reported *p* values. We find that the *p* values are rarely reported directly (e.g. "*p*=.391"), even though this is the recommendation of the APA, publisher of the journals in our sample. In fact, only about 16.5% of the entries in our sample involve an equality sign. The rest is reported by comparison with a threshold, usually .001, .01 or .05. Interestingly, in 72.8% of the cases, the obtained *p* value is lower than the threshold ("*p*<.05") and only in 10.5% higher, suggesting that either distribution of the originally obtained *p* values is highly skewed (with sharply decreasing density, i.e. most values located very close to 0) or publication bias favors low *p* values.

Figure 1 presents the distribution of directly reported *p* values.[6] Obviously, the high spikes correspond to the fact that many *p* values are rounded to .01, .02 etc. We see that the density is generally decreasing; however, contrary to what might be expected given the prevalence of "*p*<…", there is no sign of publication bias in the form of a sharp drop at conventional significance thresholds of .01 or .05. On the contrary, an interesting structure around .05 is clearly visible – there are "extra" values just below and especially just above this conventional threshold.

[6] Only values between .001 and .15 are depicted in this figure. Greater values are infrequent and smaller – very frequent. Omitting these values makes the most interesting data patterns more clearly visisble.

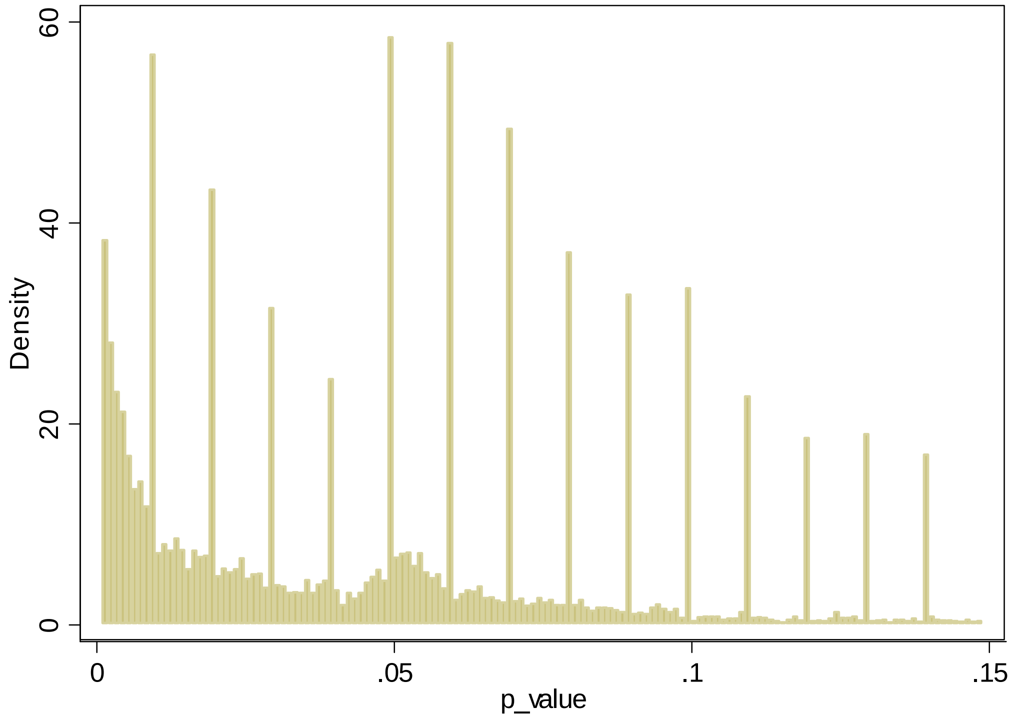

Figure 1. Distribution of directly reported *p* values (restricted to .001-.15)

This bi-modality is confirmed by means of kernel smoothing and "bump-hunting" methodology (Cox, 1966), both available from the author. Multi-modality is always a perplexing feature of the data that calls for an explanation (Good and Gaskins, 1980). Often it results from a mixture of two different distributions. In fact, I have mentioned the distinction between the "central" and the "control" variables, which is difficult to made without carefully reading the papers. However, we shall expect the "control" variables to have a roughly uniform distribution, certainly not to have a mode exactly at the arbitrary significance threshold of .05. Besides, splitting the sample based on a guess that there are more "central" variables early in the paper delivered no insights.

It could also be that there are clusters of types of studies. For example, if experiments on animals involve a lower number of observations (as, contrary to psychology students, animal subjects are costly) we could expect that they typically result in higher *p* values (e.g. with a mode close to .05), whereas other experiments have lower *p* values (often below .01). If this was the case, however, we should observe different patterns in different journals (e.g. the Journal of Experimental Psychology: Animal Behavior Processes should show *p*-values shifted to the right). In fact, the distributions are quite similar across journals, all of them showing the perplexing bi-modality.

The left part of the structure (below the 5% threshold) is consistent with the hypothesis that researchers employ "minimal" data manipulations, ending up in "just significant" results.

Actual values

Another way of obtaining the general picture of the data is to consider the distribution of "true" *p* values re-calculated from the reported underlying statistics (Figure 2, restricted to range .001-.15). It is of interest how this distribution compares to the one depicted in Figure 1, as this difference stems from the (strategic) ways to report the data.

The general decreasing pattern is now more clearly visible, as well as a clear publication bias at the 5% threshold. The artificial spikes have disappeared, as expected, as well as the "bump" around 5%, apparently created by the modes of reporting. Interestingly, however, while density is monotonically decreasing, the decrease is less pronounced just below the 5% threshold. This is confirmed by the kernel density estimation (Figure 3).

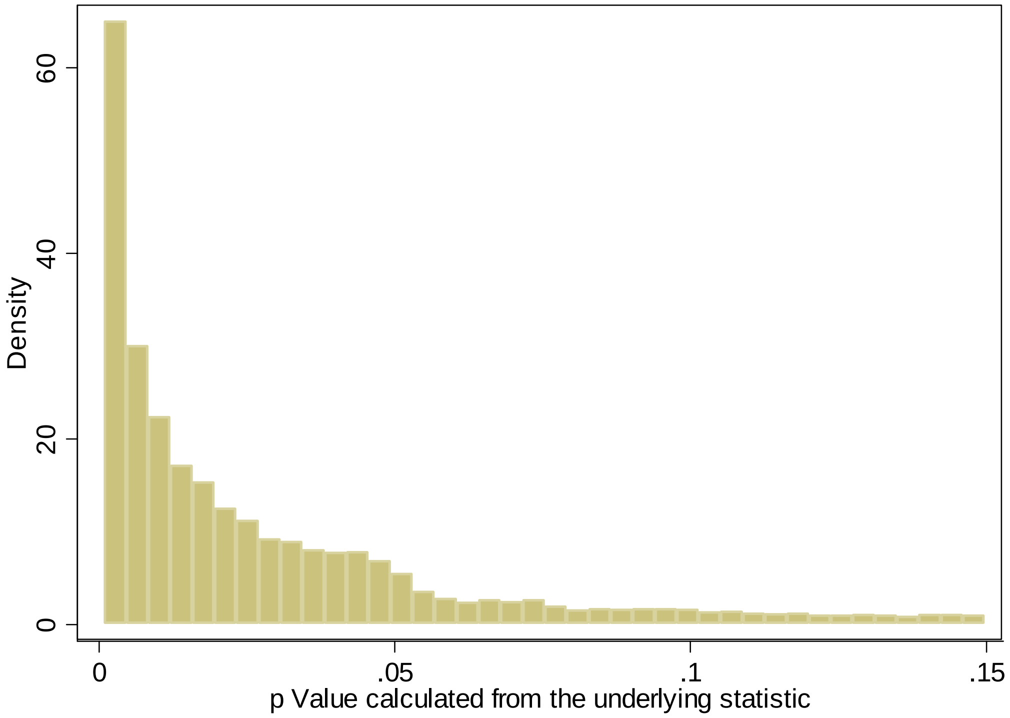

Figure 2. Distribution of re-calculated actual *p* values (restricted to .001-.15)

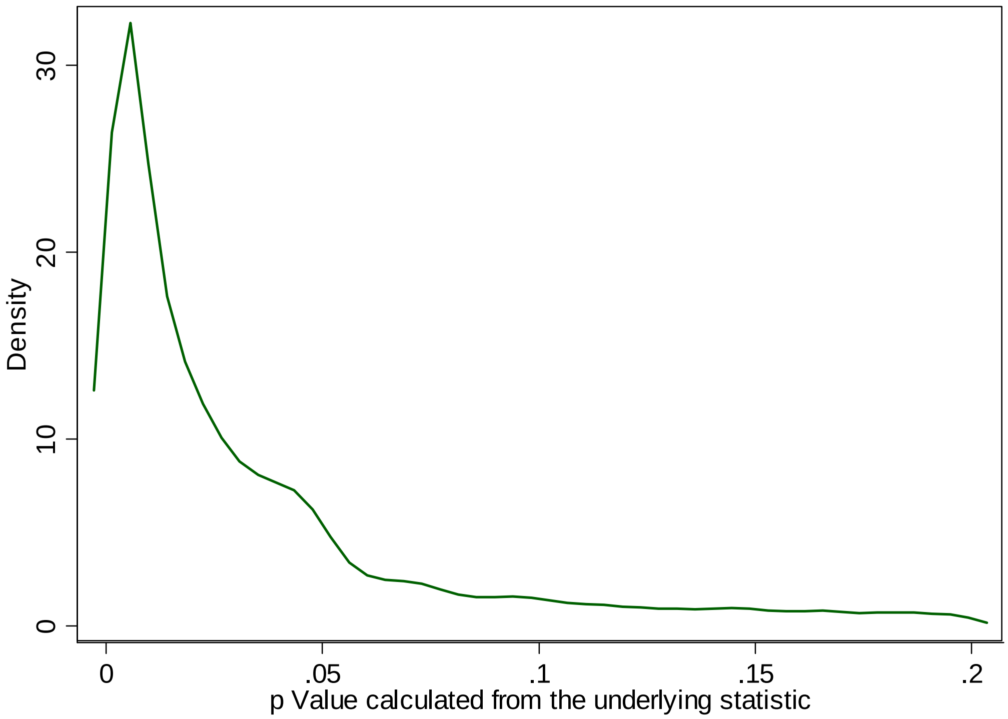

Figure 3. Kernel density smoothing of re-calculated actual *p* values (restricted to .001-.2)

Similar pattern can be observed when we consider different journals or different types of statistics separately. Of course, there is little theory to determine what distribution of *p* values should be expected. It is fully possible that the distribution depicted in the figures emerges naturally; however, the fact that "additional values" appear just below the threshold is also consistent with the notion of "minimal manipulations".

In the next subsections I will try to find out, basing on the notions of strategic data reporting, where these data patterns, particularly bi-modality of the distribution of reported values and the differences between reported and calculated *p* values come from.

3.3 The mode of reporting

The choice of sign

Suppose the author is trying to convince the reader that the finding being reported is significant. As was suggested in the subsection on the methods to manipulate data, if the *p* value is below some significance threshold, say, 5%, it suffices to report "*p*<.05". If it is above, then seeing "*p*>.05" the reader thinks it can be anywhere between .05 and 1. Thus if the value is close to the lower end of this interval, say .0687, it seems a better strategy to report it directly.

Hypothesis *(Strategic choice of sign) (1) Papers will combine the use of equality (with exact p value) and inequality (with threshold value) signs. (2) "Almost significant" values will be most likely to be reported directly (3) Values below the threshold will be least likely to be reported directly.*

Result (1) Different reporting modes are used within the same paper: nearly half the papers report both some equalities and some "smaller than". Importantly, it is not about very low values (it is natural to write "*p*<.001" rather than the actual value). Some 43% of papers report both exact values and inequalities involving numbers greater than .01.

Result (2) and (3). These points of the hypothesis can be verified by inspection of Figure 4, presenting the frequency of choosing the equality sign, depending on the actual *p* value (rounded up to the nearest percent, only values up to.2 considered for the sake of clarity of the picture).

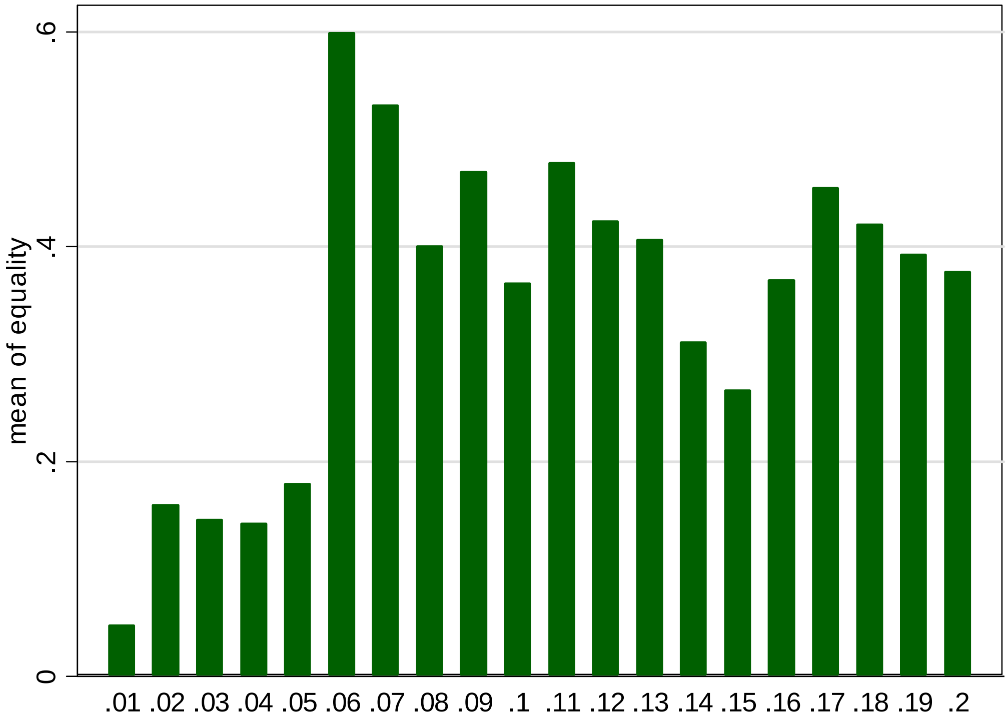

Figure 4. Frequency of reporting directly over intervals of actual *p* value.

The hypothesis is fully confirmed. The values that are truly significant (at 5%) are typically reported indirectly. This gives a hint why the "<" sign appears so much more often than ">" in our sample and why the publication bias is not visible in Figure 1.

By contrast, "almost significant" values typically are given directly, with the frequency decreasing slowly as we proceed to higher *p* values. This impression is supported by LOWESS approximation on .05-.25 (Figure 5)

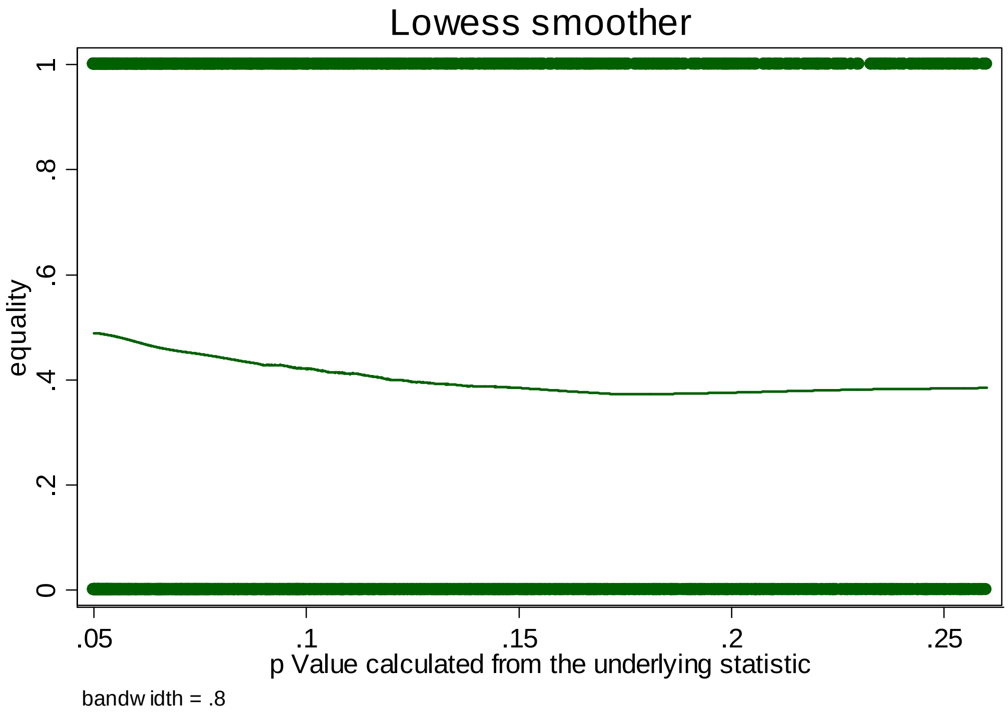

Figure 5. Lowess approximation of the frequency of direct reporting.

This finding contributes to the explanation of the right part of the observed structure around the threshold of .05 in Figure 1 – the values just above the threshold are simply most likely to be reported directly.

One important point needs to be made here, which will, to varying degree, apply to other findings: it is difficult to distinguish two effects: authors systematically reporting their "almost significant" *p* values directly more often than other values and journals accepting more readily papers in which this happens to be the case (publication bias – selection on reported *p* value). In other words, the prevalence of directly reported "almost significant" values means that either authors use a little trick or that such a trick works (or, most likely, both). On a general note, however, I should say that, first, such a selection based on publication bias should not be expected to be very strong, given the fact that there are on average about 30 *p* values per paper in our sample (among those papers that have at least one *p* value). The impact of each individual value on the publication probability is therefore limited. Second, the referees are likely to look at the underlying statistics as well – this would make the selection based on the rounded value still weaker.

The choice of threshold

Conditional on choosing not to report the *p* value directly, but rather in a comparison, which of the available thresholds should be chosen? According to the textbook hypothesis testing methodology, the critical region of statistic leading to rejection of H0 (thus equivalently: significance threshold) should be determined *ex ante*, i.e. before the data was collected, basing on the strength of hypotheses, number of observations, relative cost of errors of type I vs. type II, etc.. If, however, the goal is to persuade the referees, editors and readers that findings are significant, we should expect the author to choose possibly low threshold for which the data admits rejection of the null hypothesis. To see whether this is taking place we can consider distributions of *p* values separately for each threshold being used. Figure 6 presents histograms of *p* values for two conventional thresholds, .05 and .1 (and of course only for the sign "<" being used).

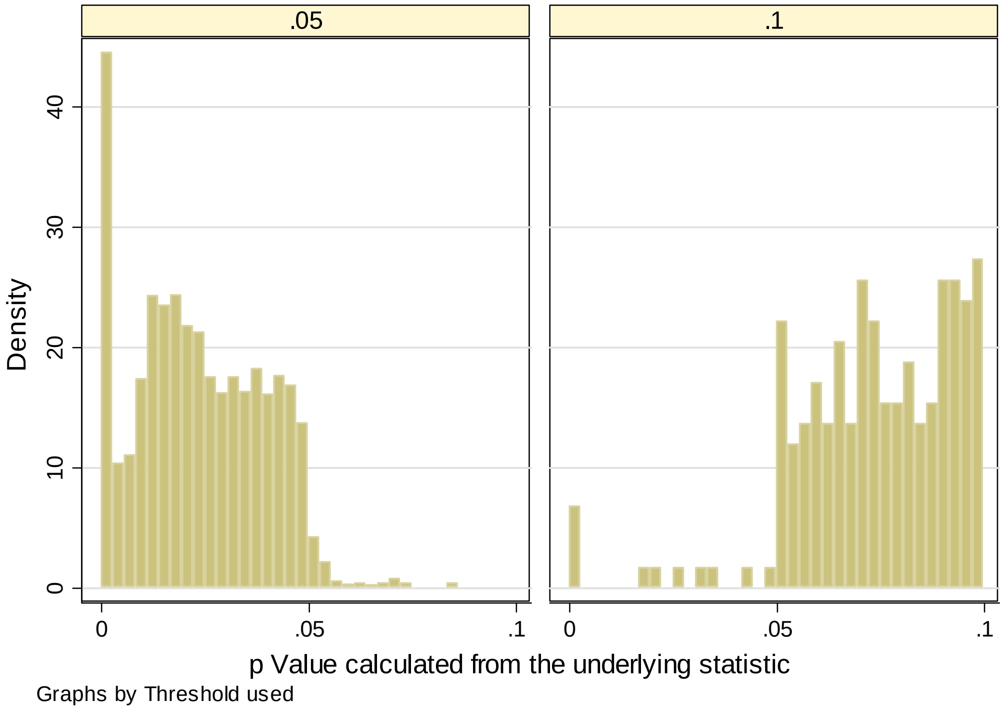

Figure 6. Actual *p* values for “*p*<.05” and “*p*<.1” being reported.

Visual inspection confirms that the thresholds are nearly always chosen ex post, in such a way to make the effect look as significant as possible the threshold of .1 is chosen only if the finding is not significant at .05. This is another major source of the prevalence of the “<” sign over the “>” sign. Obviously, the dramatic effect shown on Figure 4 cannot be accounted for by publication bias.

The choice of one-sided vs. two-sided tests

Some statistical tests, notably t-tests, require to choose between a one-sided and two-sided critical region. According to standard hypothesis testing methodology, the choice of a one-sided or two-sided test should precede data acquisition. However, as this cannot be verified, researchers may in fact use the one-sided test when it gives significance where there would be none with a one-sided test and a two-sided test otherwise.

Hypothesis. *One sided tests will prevail on .0005-.001, .005-.01 and .025-.05. Two-sided tests will prevail elsewhere.*

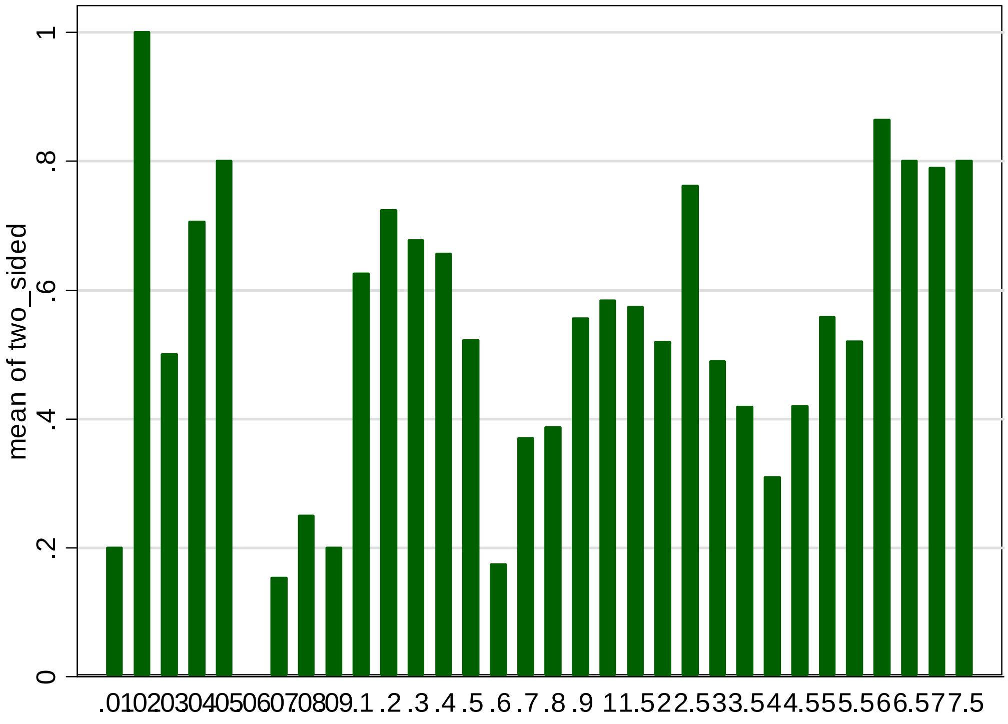

Figure 7: Frequency of use of two-sided tests over actual one-sided *p* values (rounded up, percent sign omitted).

Figure 7 shows frequency of use of two-sided tests for different actual *p* values (if one sided test was used), only for the cases when the underlying t-statistic is given and it was possible to establish, whether one- or two-sided test was used. For values below .001, intervals of length .01% were used, for values between .001 and .01: intervals of length .1%, and between .01 and .075: of length .5%. (The range is restricted for the sake of clarity of the picture, while mixed scale makes the exhibition of the hypothesized effects possible. It also corresponds to the fact that the density of the empirical distribution of *p* values is decreasing – there would not be enough entries to make inference about very short intervals further away from 0.)

Result. The data displays all the hypothesized patterns: Indeed, we see that authors tend to use two-sided tests (overall, of all 1376 cases, two-sided test was used 64.2% of the time), except for the range of .0005-.0009, .005-.0.008 and .025-.045. These intervals overlap almost perfectly with the intervals on which a one sided test is significant at .1%, 1% and 5% respectively, but a two-sided test is not and thus the use of one-sided test is predicted.

Again, while this is a combined effect of authors' strategizing and publication bias, the latter is unlikely to account for the large part of it.

The choice of precision

Suppose now that the researcher chooses to report the *p* value directly. The data shows that she typically (about 90% of the cases) reports either 2 or 3 decimals. Can we find evidence that the choice which of the two to use is made strategically, i.e. in order to make the findings appear more significant.?

First of all, it would imply that different entries within one paper have different precision. This, of course would not yet prove that the choice of precision is being made strategically. To investigate that issue, we need to focus on these entries where the underlying statistic is given and thus the actual *p* value can be recalculated. We can only consider those cases, where it can be established whether rounding from three to two decimals raised or lowered the number. For example, if we know that the true *p* value falls in the range of (.0795,.08), it will be, when rounded to two decimals (thus to .08), increased. We call such entries "roundable up". Similarly if the true value is known to be between .13 and .135, it is "roundable down" (to .13). We can now speculate that values that are “roundable up” will indeed be rounded less often than those “roundable down”.

Hypothesis *(Strategic rounding). (1) Precision (number of decimals) will vary within papers. (2) Roundable up values will be rounded less often than those roundable down, (3) especially at the thresholds.*

Indeed, we note that researchers are not consistent in their choices of precision within one paper. Suppose for example that three decimals are initially obtained (typically from a statistical package). If these are reported with maximum precision, we expect that about 90% of values reported directly in the paper will have three decimals – the remaining 10% will have 0 as the last digit, thus being likely to be abbreviated to just two decimals. About 9% will have two decimals (0 as the last but not second to last digit) and 1% just one decimal.

In fact, only about 61.4% of entries, rather than 90%, have maximum precision. Again, this is not about very low values only. If we discard all *p* values lower or equal to .01 and redefine the paper-specific maximum precision, we find that still just 66% of entries achieve it.

Regarding (2), we can see in Table 1 that the option to round is used 60.6% of the time if it lowers the original value (roundable down) and only 52.4% of the time if it raises it (roundable up).[7] The difference is significant, *p*=.001 (two sided t-test).

Table 1: Actual rounding in roundable downs and roundable ups

```
      -----------------------------------------------
         Group |     Obs    Mean  Std. Err. Std. Dev.
      ---------+-------------------------------------
Roundable down |     817    .606     .017    .489
  Roundable up |     694    .524     .019    .500
      ---------+-------------------------------------
H0: mean(0) - mean(1) = diff = 0
Ha: diff != 0
t =   3.192
P > |t| =   0.001
```

[7] It may be more natural to round (up) values like .008 than to round (down to 0) values such as .003. Therefore all the actual *p* values below .01 have been discarded for this calculation. Values above .99 were left out for the same reason.

Further analysis revealed that the effect is driven by *p* values around the threshold. Figure 8 shows how often values that are roundable down (green bars) and up (red bars) are in fact rounded, for each interval of length .01 (for clarity of the picture only first 20 intervals have been considered). The green bars are usually greater than neighboring red bars, except for the first interval (where rounding down means generating an unnatural statement "*p*=0") and for high *p* values (where presumably e.g. writing .15 rather than .153 would not alter the chances for publication). It turns out that the discrepancy is greatest for the values close to the 5% threshold. In particular, values just above it (in the sixth interval) are rounded down (if possible) more than 60% of the time and up (if possible) less than 40% of the time. Interestingly, values just below the interval, i.e. values in .005-.01 and .0045-.05 (roundable up in the first and fifth interval) have lowest overall probabilities of being actually rounded up. This partly accounts for the left part of the structure around 5% visible in Figure 1. For example, values from the (.40,.45) interval are more often rounded than the values from the (.45,.50) interval.

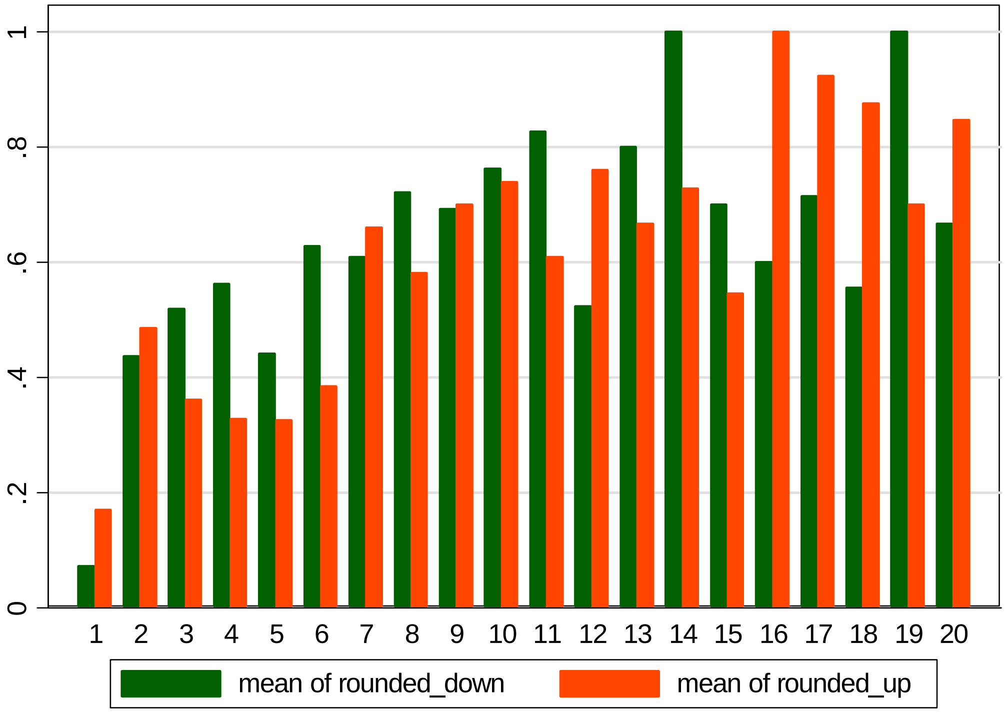

Figure 8. Frequency of rounding up and down among roundable ups and downs respectively, for each "round" value (in percent).

Again, it is not impossible that the fact that we observe more opportunities to round down than up being taken results simply from the publication bias – selection on reported *p* values. However, this would imply it is very strong. Consider, for example, typical entries from the fourth interval, i.e. *p* values of .0325 and .0375. If the publication bias leads to the probability of publication of the first one being s times higher than the second one, the rounded value of .03 must have a probability about 1.5s times higher than the rounded value of .04 to account for the relation of frequencies of rounding up and down on this interval (.058 to .038 or about 1.5)[8]. If widening the gap by .5% raises the severity of selection by the factor of 1.5, raising one *p* value from 3% to 4% lowers the probability of publication by 2.25. This would be a very severe publication bias indeed.

Strategic mistakes?

In this subsection we investigate the cases in which the recalculated *p* values are inconsistent with the reported ones. To begin with, these instances are fairly common. The reported *p* value is different from what can be recalculated about 8% of the time.

To address the issue of whether some of these "mistakes" can in fact be deliberate, we investigate how often the reported value is higher than the actual value, for each interval of length .01 separately[9]. Inspection of Figure 9 yields two observations: First, "mistakes" typically lead to lower, not higher values (the mean of mistake_increases is way below .05 for virtually all intervals). This is perplexing, as we initially had reasons to expect that most mistakes should lead to reported values being higher than true values, at least in the case of low true values. First, there is more "space" for mistakes leading to higher reported value. For example if incorrectly reported values were distributed uniformly, there should be, e.g. just a chance of .054 for the incorrectly reported *p* value to be lower than the original value of .054. Second, a "typical" mistake we find in the data relatively frequently, i.e. omission of the initial "0" (thus turning .067 into .67 etc.) obviously leads to the reported *p* value being higher than the true one.

[8] We make a conservative simplifying assumption here that the underlying true *p* value does not affect the probability of publication.

[9] For clarity of the picture only first 25 intervals are considered.

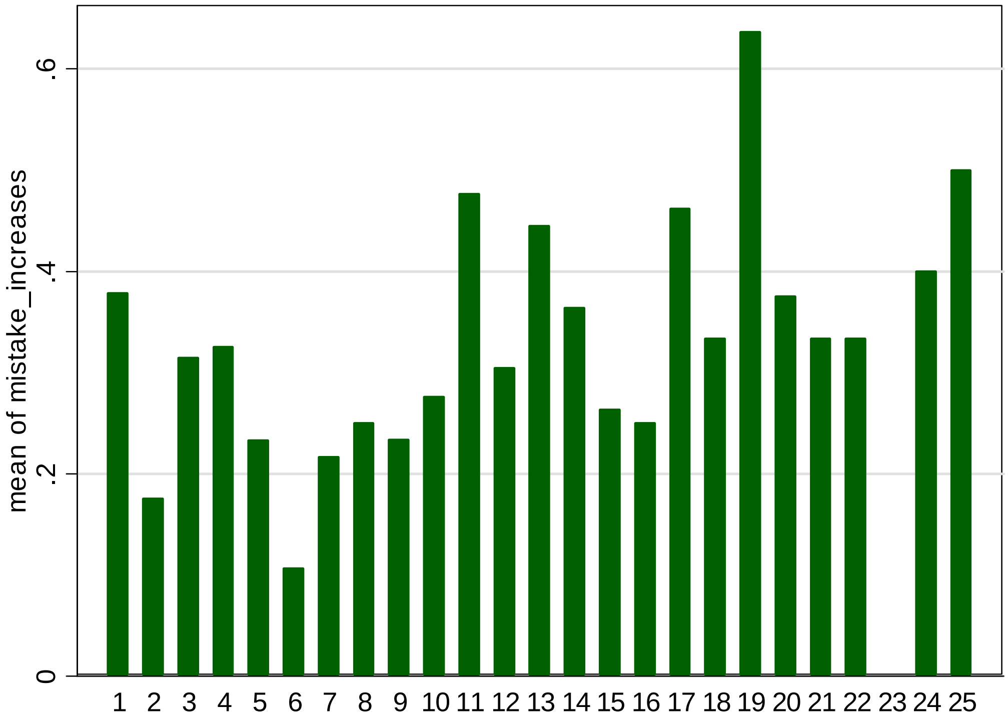

Figure 9. Fraction of mistakes leading to the reported *p* value being lower than the actual *p* value, over the calculated *p* value (rounded up, in percent).

Most mistakes do not lead to great misrepresentation of the data. For example, the median *p* value among the entries wrongly reported as "*p*<5%" is 0.066. Still, for about 25% of these cases the true *p* value exceeds .1, thus is two times higher than the reported threshold; the mean is equal to .144. Given that mistakes are generally infrequent and that they sometimes lead to substantial changes in the *p* value, the fact that most of them lead to lower, not higher *p* values, may have well resulted chiefly from selection (publication bias). In other words, we cannot prove that authors deliberately report values lower than actually obtained, though they apparently could benefit from doing so.[10]

5. Discussion and conclusion

Upon analyzing the *p* values and underlying F, t and χ^2 reported in top journals in experimental psychology we find some perplexing patterns, indicating that researchers use a number of little tricks to make their findings appear more significant than they really are, including strategic choice of sign and threshold, strategic choice between one-sided and two-sided tests and strategic rounding. It also appears that

[10] Of course, we do not know how many of the mistakes are discovered by the referees (and, most probably, lower chances for publication).

more significant infractions may be going on. Reported *p* values are sometimes inconsistent with the underlying (reported) statistic and these mistakes tend to lower the *p* value.

Without obtaining the original data and re-running all the estimations (which of course could only be done with a much smaller sample) we cannot identify other mistakes or data manipulations. We do however observe an intriguing bi-modality in the distribution of *p* values: next to the mode situated close to 0 we also have a bump centered near the significance threshold of .05. The left part of the bump – “extra” values just below the threshold, is not fully explained by any of the identified patters in the mode of reporting and thus survives (in the form of flatter density curve) also in the distribution of the actual *p* values, re-calculated from the underlying statistics. One possible explanation is based on the general assumption that researchers recognize and strategically react to the publication bias and it means we are witnessing a serious transgression against the integrity of science.

It has to be stressed that the analysis presented above is rather novel and exploratory and perhaps poses more questions than it can answer. For example, we are at risk of underestimating the frequency of data manipulation in a sense that I have no means to distinguish between the "crucial" and the "control" variables. Further, on some rare occasions, the null result can actually be desirable (e.g. when a control treatment is compared to past experiments to rule out the possibility that the current subject pool is highly unusual) which results in the opposite tendencies. The analysis may also be expanded in many interesting directions. For example, we could try to identify the "strategic reporters" to verify whether they are consistently inconsistent in different papers.

In any case, if my findings prove to be robust, they have significant implications. They generally call for more discipline in research, more robustness checks and more replications.[11] They also suggest the referees should look more carefully at the reported statistics, ask and inspect the raw data.

Perhaps we should also substantially change the way we do statistics. Such calls are frequent in the literature. In his paper "The earth is round (*p*<.05)" published in 1994, Cohen complained that "After four decades of severe criticism, the ritual of null hypothesis significance testing – mechanical dichotomous decisions around a sacred

[11] Dewalt et al. (1986) claim that "(...) errors in published articles are a commonplace rather than a rare occurrence".

.05 criterion – still persists.". And apparently little has changed since, making the researchers call anew for abandoning the "*p*-value fallacy" (Dixon, 2003). Others suggest reporting exact *p* values instead of just the inequality (Sterne and Smith, 2001) and forgetting about the conventional significance thresholds altogether, perhaps switching to the Bayesian approach (Goodman, 1999). My findings about the adverse incentive effects of the standard use of the significance thresholds indicate that these suggestions should be taken very seriously.

REFERENCES

Cohen, J. (1994). The Earth Is Round ($p < .05$). American Psychologist, 49(12).

Cox, D.R. (1966). Notes on the analysis of mixed frequency distributions. British Journal of Mathematical Statistical Psychology 19, 39-47.

Dixon, P. (2003). The *p* value Fallacy and How to Avoid It. Canadian Journal of Experimental Psychology, 2003, 57:3, 189-202.

Fox, M.F. (1994). Scientific Misconduct and Editorial and Peer Review Processes. The Journal of Higher Education 65(3): 298-309.

Goodman, S.N. (1999). Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy. Annals of Internal Medicine 130: 995-1004.

Leamer, E.E. (1983) Let's take the con out of econometrics, American Economic Review 73(1): 31.

List, J., C. Bailey, P. J. Euzent, T.L.Martin (2001). Academic Economists Behaving Badly? A Survey on Three Areas of Unethical Behavior. Economic Inquiry 39(1): 162-170.

Roth, A. (1994). Let’s Keep the Con out of Experimental Economics. A Methodological Note. Empirical Economics 19: 279-289.

Stanley, T.D. (2005) Beyond Publication Bias. Journal of Economic Surveys 19(3).

Steneck, N. (2006). Fostering Integrity in Research: Definitions, Current Knowledge, and Future Directions. Science and Engineering Ethics (2006) 12, 53-74

Sterne, J.A.C and Smith, G.D. (2001). Sifting the evidence—what's wrong with significance tests? BMJ 322:226--31.

White, C. (2005). Suspected research fraud: difficulties of getting at the truth. BMJ 331: 281-288.

Wolf, P.K. (1986). Pressure to publish and fraud in science. Annals of Internal Medicine. 104(2):254-6.