TITLE:

**Reply to de Winter and Dodou (2014): Growing bias and the hierarchy are actually supported, despite different design, errors, and disconfirmation-biases.**

AUTHOR: Daniele Fanelli, Université de Montréal.

I appreciate the efforts that de Winter and Dodou (2014) have put into replicating and challenging claims made by Fanelli (2010, 2012), as well as those of Pautasso (2010). This is how all sciences should make progress, and it is therefore both a duty and an honour to respond to this challenge.

The results presented are largely in agreement with claims by Fanelli (2012 and 2010), but this fact is obfuscated by a somewhat selective interpretation of findings, reinforced by differences in study design, and major flaws in the sampling and analytical design.

FLAWS IN INTERPRETATION:

1) Fanelli (2012) claimed that negative results are disappearing in percentage, which is exactly what is found here. Even de Winter and Dodou (2014) quote Fanelli (2012) as using percentage figures, so I am quite baffled as to why they consider their results at odds with mine. For the record, the absolute number of negative results in Fanelli (2012) did not show a decline, and it was never claimed in the paper that it did.

   An absolute increase in the number of both positives and negatives is generally to be expected, since the annual number of records added to databases has regularly increased. In the discussion de Winter and Dodou perceive a 13.9-fold increase in positives as "equally staggering" as a 4.3-fold increase in negatives. This is a rather surprising lack of enthusiasm for a result that is possibly more extreme than what both Fanelli (2012) and Pautasso (2010) had reported.

2) The Hierarchy of the Sciences seems also remarkably supported by de Winter and Dodou, despite their claims of the contrary. By their own admission, the rate of increase in positive results has been fastest in the social sciences and, as reported in Table 1 of their work, actually shows a hierarchy-like pattern (i.e. physical-biological-social) with all proxies.

   Instead of conceding this point, de Winter and Dodou (2014) claim that the Hierarchy theory is refuted because the ratio of non-significant results is similar across all domains when based on p-values (Figure 6), and higher in the physical sciences when based on textual reporting (Figure 7, but note how the Social-Biological difference is strong, and exactly in the direction predicted).

   In reality, all that these findings show is that the *overall* frequency of positive and negatives is highly sensitive to the particular proxy used. Temporal changes for the same proxies, however, should logically be considered more consistent, since different practices between disciplines or countries are controlled for. This point was already made by Fanelli (2012). Even de Winter and Dodou (2014) discuss at length this issue, yet they overlook it when it comes to discussing their evidence for the Hierarchy of the Sciences.

   In Figure 1 de Winter and Dodou (2014) took care in reporting data already presented by Fanelli (2012). Since Fanelli 2012 had also presented these data in

graph form, the need for Figure 1 is rather unclear. In any case, Figure 1 illustrates how, even with Fanelli (2012)'s proxy, the three domains overlap, yet differ in absolute magnitude as well as steepness when controlling for other confounders. So, once again, it is unclear why the authors would consider their findings, which are visually identical to, and in some respects are actually stronger than Fanelli (2012), a refutation of this latter's claims.
It should also be remarked that, unlike Fanelli (2012), Fanelli (2010), which made the original claim for a correlation between the Hierarchy of the Sciences correlation and reporting biases, had shown how this was only true for pure disciplines, whereas the applied disciplines showed uniformly high frequencies. Aggregating all disciplines, as de Winter and Dodou (2014) do, would risk erasing differences between domains.

3) Finally, even though de Winter and Dodou (2014) compared trends between geo-economic regions just as as Fanelli (2012) did, they spend little time comparing the two independent results. This is quite a pity, since such results are remarkably in agreement: Asian countries show an overall stronger increase than both US and EU, as Figures 9 and 10, and table 2 report.

Part of the author's claims to have refuted previous evidence rests on the lack of statistical significance in some of their analyses. But this is where weaknesses in their study design become an important source of confusion.

FLAWS IN STUDY DESIGN

1) De Winter and Dodou (2014) use linear regression on proportion data. This is a statistical mistake, which violates the critical assumptions of normality and homoscedasticity for linear regression. Moreover, judging by the data sets they have posted online, de Winter and Dodou (2014) have applied linear regression to each proportion by year, irrespective of sample size or any other factor. Such analysis is invalid unless each data point is weighted by sample size. Furthermore, the statistical power is hugely limited, since the effective sample size corresponds to the number of years considered.
All these mistakes were avoided by Fanelli (2012), in which each data point corresponds to one paper from the sample and is analysed for multiple characteristics through logistic regression, a much more robust and powerful analysis, which models binary outcomes in the only correct way.

2) De Winter and Dodou (2014) present individual, linear regression slopes in tables, and do not even attempt to conduct multivariable analyses, for no apparent reason other than, I take the liberty to presume, a lack of familiarity with such techniques. De Winter and Dodou (2014) acknowledge that their analyses do not correct for obvious confounders, in particular the different frequencies with which countries appear in different fields, and correctly consider this a major limitation. Unfortunately, they fail to mention that Fanelli (2012) (as well as Fanelli (2010) and all other studies I conducted on these issues) avoided such limitations, precisely by using a multiple regression approach.

3) Disciplinary classification: de Winter and Dodou (2014) used the Scopus classification system. This is another major flaw, avoided by Fanelli (2012). Like most classification systems in other databases, Scopus disciplinary categories are overlapping, which means that any given paper in de Winter and Dodou's sample is likely to fall in multiple domains, for example both in the physical and in the social sciences. This alone should make any comparison between domains severely flawed. Fanelli (2012) (as well as Fanelli (2010) and all of my studies) used the Essential Science Indicators classification system, which is mutually exclusive (no overlap between disciplines).

4) A similar mistake to the above was made by de Winter and Dodou (2014) when classifying countries. De Winter and Dodou (2014) have aggregated all the countries of all addresses, which means that each paper could appear simultaneously in the US, EU and Asian samples. This time, the mistake is rather unjustified, since Fanelli (2012) had explicitly limited the analysis to the country of the corresponding author, an attribution that is usually easy to make for any paper.

5) Finally, as de Winter and Dodou (2014) correctly discuss, their proxy differs in major ways from that used by Fanelli (2012), and partially from what Pautasso (2010) measured, too.
I deem it unnecessary to discuss this matter at length, although it alone should send a warning against any claims to refutation. Claims of support should be hedged too, of course, but there is an important asymmetry, which I discuss in the last paragraph.

In conclusion, despite substantial differences in the proxy used and study design, and despite major flaws in the sampling and analytical strategy, results presented by de Winter and Dodou (2014) are in remarkable agreement with claims made by Fanelli (2010, 2012). De Winter and Dodou (2014) are adamant of the contrary, and in several passages seem to betray a "disconfirmation bias" against my findings.

Hostile replications are a positive force in science, so by no means I wish to criticize de Winter and Dodou (2014)'s scepticism towards my results, or discourage other researchers from attempting similar replications.
A more balanced interpretation would make scientific self-correction more efficient. However, the fact that their expectations are made explicit, together with their meticulous reporting of methods and results, is an example of how scientific disputes should be conducted in all fields.

When independent studies, using different methods and performed by researchers who are sceptical of previous claims, find completely different results, the lack of agreement is easily explained away as an effect of biased methodological choices on one or both sides. Conversely, however, when under the same conditions studies find patterns that are, to any extent, in agreement, this is a strong suggestion that the underlying phenomena are

real, because they are measurable despite all a-priori biases and methodological degrees of freedom.

I therefore thank de Winter and Dodou (2014) for offering results that will help the scientific community get to the bottom of important and controversial problems.

REFERENCES

de Winter and Dodou (2014). A surge of p-values between 0.040 and 0.049 in recent decades (but negative results are increasing rapidly too). PeerJ PrePrints doi:10.7287/peerj.preprints.447v1

Fanelli D (2010). "Positive" results increase down the Hierarchy of the Sciences. PLoS ONE - DOI:10.1371/journal.pone.0010068

Fanelli D (2012) Negative results are disappearing from most disciplines and countries. Scientometrics - DOI:10.1007/s11192-011-0494-7

Pautasso, M. (2010). Worsening file-drawer problem in the abstracts of natural, medical and social science databases. Scientometrics, 85(1), 193–202. doi:10.1007/s11192-010-0233-5.

For further literature and information please go to: danielefanelli.com