Negative results are disappearing from most disciplines and countries

Daniele Fanelli

Abstract   Concerns that the growing competition for funding and citations might distort science are frequently discussed, but have not been verified directly. Of the hypothesized problems, perhaps the most worrying is a worsening of positive-outcome bias. A system that disfavours negative results not only distorts the scientific literature directly, but might also discourage high-risk projects and pressure scientists to fabricate and falsify their data. This study analysed over 4,600 papers published in all disciplines between 1990 and 2007, measuring the frequency of papers that, having declared to have "tested" a hypothesis, reported a positive support for it. The overall frequency of positive supports has grown by over 22% between 1990 and 2007, with significant differences between disciplines and countries. The increase was stronger in the social and some biomedical disciplines. The United States had published, over the years, significantly fewer positive results than Asian countries (and particularly Japan) but more than European countries (and in particular the United Kingdom). Methodological artefacts cannot explain away these patterns, which support the hypotheses that research is becoming less pioneering and/or that the objectivity with which results are produced and published is decreasing.

Keywords   Bias · Misconduct · Research evaluation · Publication ·
Publish or perish · Competition

Introduction

Competition in science is changing, and concerns that this might distort scientific knowledge are openly and commonly discussed (Young et al. 2008; Statzner and Resh 2010). The traditional race for priority of important discoveries is increasingly intertwined with a struggle for limited funding and jobs, the winners of which are determined by measures of performance and impact (Young et al. 2008; Bonitz and Scharnhorst 2001; Statzner and

D. Fanelli (✉)
ISSTI-Institute for the Study of Science, Technology and Innovation, The University of Edinburgh, Old Surgeons' Hall, Edinburgh EH1 1LZ, Scotland, UK
e-mail: dfanelli@staffmail.ed.ac.uk

Resh 2010). Individual scientists, research institutions, countries, international organizations, and scientific journals are increasingly evaluated based on the numbers of papers they publish and citations they receive (Shelton et al. 2009; Meho 2007; Nicolini and Nozza 2008; King 2004). From all these levels, therefore, come pressures on researchers to publish frequently and in high-ranking journals (Lawrence 2003). This combination of competition and bibliometric evaluation has a longer history in the United States, but is increasingly adopted across fields and countries as a way to improve productivity and the rational distribution of resources (Warner 2000; Qiu 2010; de Meis et al. 2003; Osuna et al. 2011). How well bibliometric parameters reflect actual scientific quality, however, is controversial, and the effects that this system might have on research practices need to be fully examined (De Rond and Miller 2005; Osuna et al. 2011; Young et al. 2008).

Several possible problems have been hypothesised, including: undue proliferation of publications and atomization of results (Gad-el-Hak 2004; Statzner and Resh 2010); impoverishment of research creativity, favouring "normal" science and predictable outcomes at the expense of pioneering, high-risk studies (De Rond and Miller 2005); growing journal rejection rates and bias against negative and non-significant results (because they attract fewer readers and citations) (Statzner and Resh 2010; Lortie 1999); sensationalism, inflation and over-interpretation of results (Lortie 1999; Atkin 2002; Ioannidis 2008b); increased prevalence of research bias and misconduct (Qiu 2010). Indirect empirical evidence supports at least some of these concerns. The per-capita paper output of scientists has increased, whilst their career duration has decreased over the last 35 years in the physical sciences (Fronczak et al. 2007). Rejection rates of papers have increased in the high-tier journals (Larsen and von Ins 2010; Lawrence 2003). Negative sentences such as "non-significant difference" have decreased in frequency in papers' abstracts, while catchy expressions such as "paradigm shift" have increased in the titles (Pautasso 2010; Atkin 2002). No study, however, has yet verified directly whether the scientific literature is enduring actual changes in content.

One of the most worrying distortions that scientific knowledge might endure is the loss of negative data. Results that do not confirm expectations—because they yield an effect that is either not statistically significant or just contradicts an hypothesis—are crucial to scientific progress, because this latter is only made possible by a collective self-correcting process (Browman 1999; Knight 2003). Yet, a lack of null and negative results has been noticed in innumerable fields (Song et al. 2010; Gerber and Malhotra 2008; Howard et al. 2009; Dwan et al. 2008; Jennions and Moller 2002). Their absence from the literature not only inflates effect size estimates in meta-analyses, thus exaggerating the importance of phenomena, but can also cause a waste of resources replicating research that has already failed, and might even create fields based on completely non-existent phenomena (Ioannidis 2005, 2008b; Feigenbaum and Levy 1996; Song et al. 2010). In meta-analysis, publication bias can in part be corrected by assuming that negative results are simply never written up, and are left lying in scientists' drawers (Formann 2008). However, this assumption is obviously naïve. A realistic scenario includes various forms of conscious and unconscious biases that affect all stages of research—e.g., study design, data collection and analysis, interpretation and publication—producing positive findings when there should be none, thus creating distortions that are difficult to correct a posteriori (Ioannidis 2008a; Marsh and Hanlon 2007; Jeng 2006). The problem is bound to be particularly acute in fields where theories and methods are less clearly defined, and true replication is rare or impossible (Palmer 2000; Kelly 2006; Evanschitzky et al. 2007).

This study verified whether the frequency of positive results has been increasing in the contemporary scientific literature. Papers that declared to have tested a hypothesis were

searched in over 10,800 journals listed in the ISI-Essential Science Indicators database, excluding the highest-impact multidisciplinary journals like Science, Nature or PNAS. By reading the abstracts and eventually full-text of papers sampled at random from all disciplines, it was determined whether the authors of the study had concluded to have found a "positive" (full or partial) or "negative" (null or negative) support for the tested hypothesis. Analyses on a previous sample spanning the years 2000–2007 ($N = 2,434$) found that papers were more likely to report a positive result in disciplines and methodologies believed to be "softer" (e.g., Psychology vs. Space Science, behavioural vs. chemical analyses), and when the corresponding author worked in states of the USA where academics publish more papers per capita—findings which suggest that this measure is a reliable proxy of bias (Fanelli 2010a, b). This study expanded the analysis to include papers published in the 1990s (total $N = 4,656$).

Results

The proportion of papers reporting a positive result in the sample was 70.2% in 1990–1991, peaked at 88.6% in 2005 and was 85.9% in 2007. On average, the odds or reporting a positive result have increased by around 6% every year, showing a statistically highly significant trend (Fig. 1). The size of this effect remained almost identical when controlling for differences between disciplines, country of corresponding author and papers testing multiple versus single hypotheses (Table 1). For simplicity, only countries with more than 90 papers plus all the others combined are shown in Table 1, but controlling for actual country of each paper (70 countries in total) yielded the same 6% increase rate ($B = 0.057 \pm 0.008$, Wald $= 47.554$, $P < 0.001$, OR $= 1.059(1.042–1.076)$). The effect was only very slightly reduced (to around 5% per year, a non significant difference) when controlling for all available potential confounding variables: discipline, domain, methodology, country, broad geographical area (i.e., US, EU-15, Asia, Other), and papers testing multiple versus single hypotheses ($B = 0.049 \pm 0.009$SE, Wald $= 27.641$, $P < 0.001$, $R = 1.051(1.031–1.07\ 05\%$CI). This latter maximal model explained over 8% of the variance in papers' outcomes (Nagelkerke $R^2 = 0.082$).

Positive results differed significantly between disciplines, both in the average frequency and in the rate of increase over the years. The average frequency of positive results was significantly higher when moving from the physical, to the biological to the social sciences, and in applied versus pure disciplines (Tables 1, 2), all of which confirms previous findings (Fanelli 2010b). Space science had not only the lowest frequency of positive results overall (Table 1), it was also the only discipline to show a slight decline in positive results over the years, together with Neuroscience & Behaviour (Fig. 2). A relatively stable (no-growth)

Fig. 1 Percentage of papers reporting a support for the tested hypothesis, plotted against year of publication. Logistic regression estimates are uncorrected for any covariate. A smoothed interpolation line was added to help visualize trends
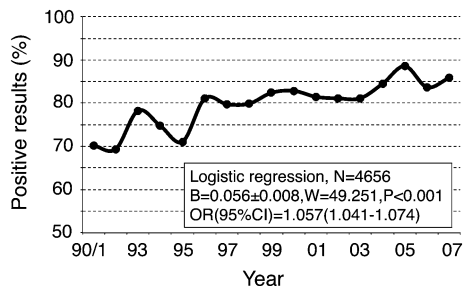

Logistic regression, N=4656
B=0.056±0.008,W=49.251,P<0.001
OR(95%CI)=1.057(1.041-1.074)

Table 1 Logistic regression slopes, standard error, Wald-test and significance, odds-ratio and 95% confidence interval predicting the likelihood of a paper to report a positive result depending on the following characteristics: year of publication, discipline of journal, national location of corresponding author (only countries with $N \geq 90$), paper testing one versus multiple hypotheses (only the first of which was included in the analysis)

Variable	B	SE	Wald	df	Sig.	OR	95% CI OR
Year	0.058	0.008	49.042	1	<0.001	1.059	1.042–1.077
Discipline (all)			53.852	19	<0.001		
Agricultural Sc.	0.510	0.238	4.605	1	0.032	1.665	1.045–2.652
Biology & Bioch.	0.836	0.244	11.742	1	<0.001	2.307	1.430–3.720
Chemistry	0.658	0.268	6.040	1	0.014	1.930	1.142–3.261
Clinical Medicine	0.866	0.250	11.962	1	0.001	2.378	1.455–3.884
Computer Sc.	0.780	0.367	4.522	1	0.033	2.181	1.063–4.474
Economics & Bus.	0.973	0.252	14.924	1	<0.001	2.654	1.615–4.333
Environment & Ec.	0.385	0.226	2.907	1	0.088	1.469	0.944–2.286
Engineering	0.756	0.296	6.535	1	0.011	2.129	1.193–3.802
Geosciences	0.362	0.232	2.435	1	0.119	1.436	0.912–2.262
Immunology	1.020	0.244	17.435	1	<0.001	2.774	1.718–4.479
Molecular Biology	0.645	0.236	7.492	1	0.006	1.906	1.201–3.024
Microbiology	0.577	0.231	6.219	1	0.013	1.781	1.131–2.803
Materials Sc.	1.152	0.344	11.198	1	0.001	3.164	1.611–6.211
Neurosci. & Beh.	0.970	0.243	15.957	1	<0.001	2.637	1.639–4.243
Plant and An. Sc.	0.383	0.220	3.021	1	0.082	1.466	0.952–2.257
Physics	0.656	0.283	5.379	1	0.020	1.927	1.107–3.353
Psyc. & Psychiatry	1.096	0.249	19.317	1	<0.001	2.992	1.935–4.877
Pharm. & Toxicol.	0.946	0.244	14.994	1	<0.001	2.576	1.596–4.160
Social Sc. General	0.763	0.236	10.404	1	0.001	2.144	1.349–3.407
Country (>100 papers)			21.426	9	0.011		
Australia	−0.160	0.199	0.646	1	0.422	0.852	0.577–1.259
Canada	−0.067	0.155	0.187	1	0.665	0.935	0.690–1.267
Germany	−0.133	0.193	0.477	1	0.490	0.875	0.599–1.278
France	−0.341	0.219	2.432	1	0.119	0.711	0.463–1.092
Italy	0.241	0.291	0.685	1	0.408	1.272	0.719–2.249
Japan	0.938	0.400	5.512	1	0.019	2.555	1.168–5.592
Netherlands	0.147	0.261	0.319	1	0.572	1.159	0.695–1.933
United Kingdom	−0.409	0.147	7.686	1	0.006	0.665	0.498–0.887
Other	0.165	0.121	1.867	1	0.172	1.179	0.931–1.493
Multiple hypotheses	−0.966	0.147	43.098	1	<0.001	0.381	0.285–0.508
Constant	−114.350	16.452	48.310	1	<0.001	<0.001	

Categorical variables were tested for overall effect in the model, then compared by indicator contrast to Space Science (for domain) and to United States (for countries). Minimum statistical power to detect a small, medium and large main effect for contrast between countries (i.e., US vs. Italy) = 0.22, 0.70, 0.95; for contrasts between disciplines (i.e., Space Sc. vs. Computer Sc.) = 0.18, 0.59, 0.87. Smoothed interpolation lines were added to help visualize trends

trend was apparent in Geosciences and Plant and Animal Sciences, but in most cases the frequency of positive results was increasing. In eight disciplines, the trend was significantly steeper than Space Science (in descending order of effect size: Economics and Business, Clinical Medicine, Psychology\Psychiatry, Pharmacology and Toxicology, Molecular Biology, Physics, Agricultural Sciences, Social Sciences General; Fig. 2). In the last year of the series, 2007, in five of these eight disciplines positive results were over 90% of the total. Controlling for other factors, the rate of growth of positive results was significantly higher in the social sciences compared to the physical sciences (Table 2; Fig. 3).

The overall frequency of positive results was significantly different between countries. The United States published significantly fewer positive results than Japan, but more than the UK (Table 1). If broader geographical regions were considered, the US published significantly fewer positive results than Asian countries but more than EU-15 (Table 2). Although no statistically significant difference in growth over the years was detected between countries or broader geographical regions (Figs. 4, 5), patterns appeared to vary, showing a marked increase in the US and UK, uniformly high frequencies in Japan, and a null or negative trend in Canada, Australia and France (Fig. 4).

Discussion

The proportion of papers that, having declared to have tested a hypothesis, reported a full or partial support has grown by more than 20% between 1990 and 2007. Underlying this overall increase were significant differences between disciplines and countries. The trend was significantly stronger in the social sciences (i.e., Psychology/Psychiatry, Economics & Business and Social Sciences, General) and in applied disciplines. Whilst a few disciplines showed a null or even a slightly declining trend (i.e., Space Science, Geosciences,

Table 2 Logistic regression slopes, standard error, Wald-test statistic and significance, odds-ratio and 95% confidence interval predicting the likelihood of a paper to report a positive result depending on the following characteristics: year of publication, scientific domain of journal, geographical location of corresponding author, journal pertaining to applied versus pure disciplines, paper testing one versus multiple hypotheses (only the first of which was included in the analysis)

Variable	B	SE	Wald	df	Sig.	OR	95% CI OR
Year	0.058	0.008	51.779	1	<0.001	1.060	1.043–1.077
Domain (overall)			11.241	2	0.004		
Biological Sc.	0.175	0.093	3.570	1	0.059	1.191	0.993–1.428
Social Sc.	0.426	0.127	11.226	1	0.001	1.531	1.193–1.963
Region (overall)			10.967	3	0.012		
Asia	0.557	0.234	5.672	1	0.017	1.746	1.104–2.761
EU-15	−0.183	0.093	3.910	1	0.048	0.833	0.694–0.998
Other	−0.006	0.106	0.004	1	0.951	0.994	0.808–1.222
Multiple vs. single Hp	−1.075	0.141	57.784	1	<0.001	0.341	0.259–0.450
Pure vs. applied	0.188	0.087	4.687	1	0.030	1.206	1.018–1.430
Intercept	−115.356	16.215	50.611	1	<0.001	<0.001	

Categorical variables were tested for overall effect in the model, then compared by indicator contrast to physical sciences (for domain) and to United States (for geographical location). Minimum statistical power to detect a small effect for contrast between domains (Physical vs. Social Sciences) = 0.95; for contrast between geographical regions (i.e., US vs. Other) = 0.95
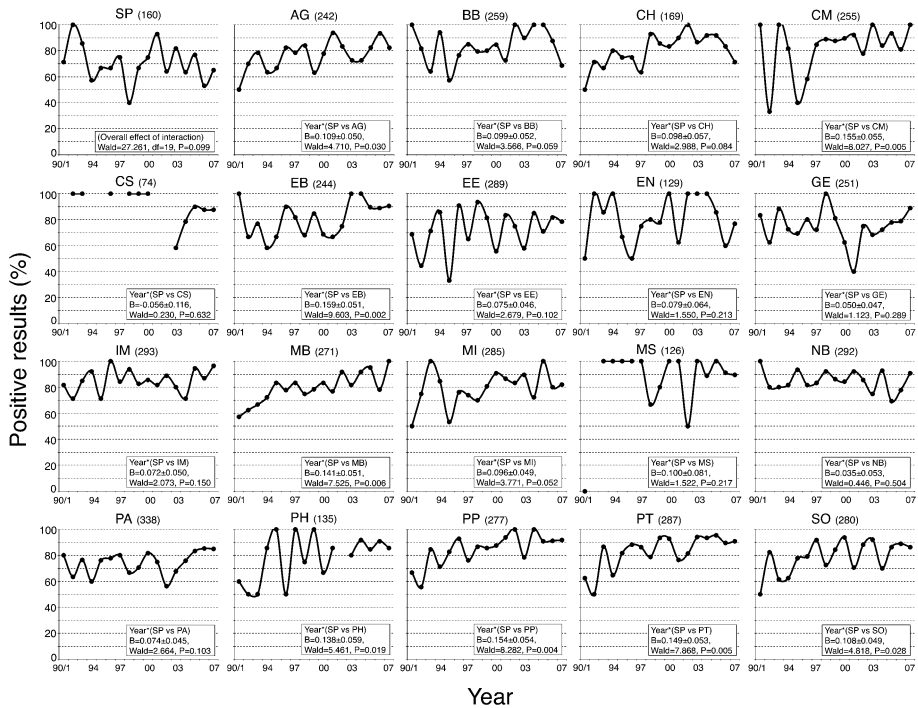
Fig. 2 Percentage of papers reporting a support for the tested hypothesis plotted against year of publication and divided by discipline of journal (*SP* Space Science, *AG* Agricultural Sciences, *BB* Biology & Biochemistry, *CH* Chemistry, *CM* Clinical Medicine, *CS* Computer Science, *EB* Economics & Business, *EE* Environment/Ecology, *EN* Engineering, *GE* Geosciences, *IM* Immunology, *MB* Molecular Biology & Genetics, *MI* Microbiology, *MS* Materials Science, *NB* Neuroscience & Behaviour, *PA* Plant and Animal Sciences, *PH* Physics, *PP* Psychiatry/Psychology, *PT* Pharmacology & Toxicology, *SO* Social Sciences, General). Regression values report the interaction effects estimated in a hierarchically well-formulated logistic regression model, controlling for: year of publication, scientific domain of journal, paper testing single versus multiple hypotheses, national location of corresponding author. The main effects of this regression model, with interaction components removed, are reported in Table 1. Numbers in brackets are sample size. Smoothed interpolation lines were added to help visualize trends

Neuroscience & Behaviour, Plant and Animal Sciences), most were undergoing a significantly positive growth (e.g., Clinical Medicine, Pharmacology and Toxicology, Molecular Biology, Agricultural Sciences). Corresponding authors based in Asian countries (and in particular Japan) reported more positive results than in the US, who in turn reported more positives than in Europe, and particularly in the UK.

Methodological artefacts cannot explain the main findings of this study. Although performed by only one author, the coding was blind to year of publication and country of corresponding author. The coding was not blind to discipline, but the effects observed are independent of discipline or domain (Tables 1, 2). The coding was not blind to decade, having been performed first for 2000–2007 and then for 1990–1999. However, if this had introduced a bias in the coding, then we would expect a discontinuity between the years 1999 and 2000. Such discontinuity was not observed (Fig. 1) and there was no significant difference in the prevalence of positive results between decades when controlling for year of publication ($B = 0.183 \pm 0.144$, Wald $= 1.607$, $df = 1$, $P = 0.205$, power to detect a small effect $= 0.996$). Indeed, positive results increased significantly within each decade
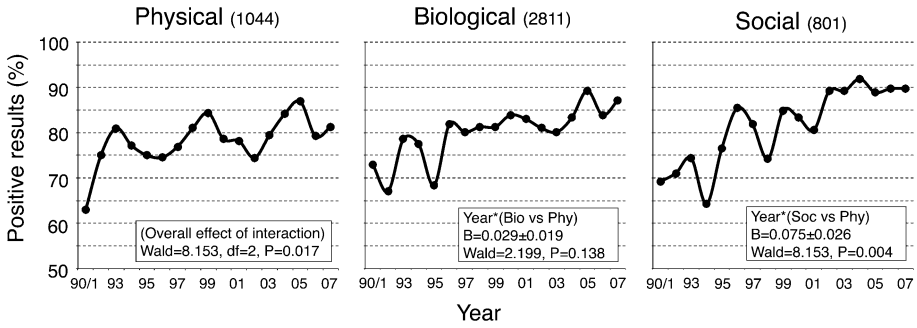
Fig. 3 Percentage of papers reporting a support for the tested hypothesis plotted against year of publication and divided by scientific domain of the journal (physical, biological and social sciences). Logistic regression estimates are interaction effects in a hierarchically well-formulated model. The main effects of this model, calculated with interaction components removed, are reported in Table 1. Numbers in brackets are sample size
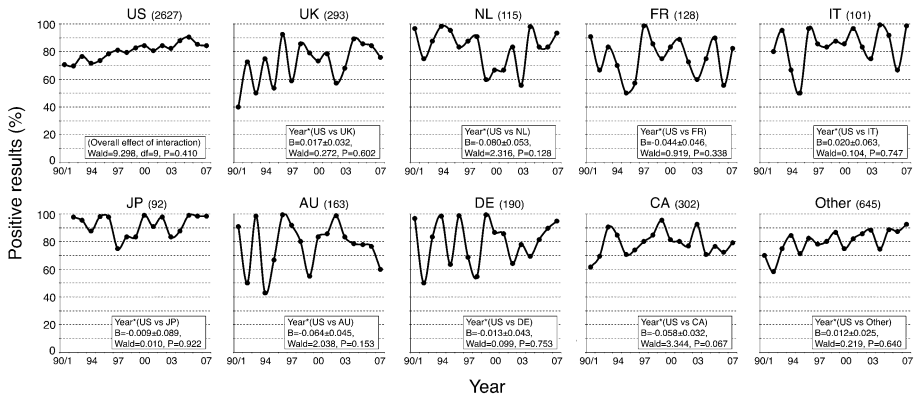

Fig. 4 Percentage of papers reporting a support for the tested hypothesis plotted against year of publication and divided by national location of corresponding author (*US* United States, *UK* United Kingdom, *NL* Netherlands, *IT* Italy, *JP* Japan, *AU* Australia, *DE* Germany, *CA* Canada). The figure shows all countries more than 90 papers in the study sample, plus all other countries combined. Logistic regression estimates are interaction effects in a hierarchically well-formulated model. The main effects of this model, calculated with interaction components removed, are reported in Table 2

(1990–1999: $B = 0.85 \pm 0.020$, Wald $= 18.978$, $P < 0.001$; 2000–2007: $B = 0.052 \pm 0.024$, Wald $= 4.584$, $P = 0.032$). This trend had not been noticed in a previous study covering the years 2000–2007, because year of publication had been treated as a purely confounding effect (i.e., tested as a categorical variable). Changing the parameterization of year in these regression models did not affect the estimation of the other parameters in any meaningful way, so previous conclusions remain valid (Fanelli 2010b).

To the best of the author's knowledge, this is the first direct evidence that papers reporting negative results have decreased in frequency across disciplines. A recent study adopting a different approach reached similar conclusions by finding a decrease in the use of the term "non-significant difference" in abstracts from various databases (i.e., Science and Social Sciences Citation index, Medline, CAB), over a period of up to 40 years (Pautasso 2010). This latter study did not examine the actual outcome of each paper, and
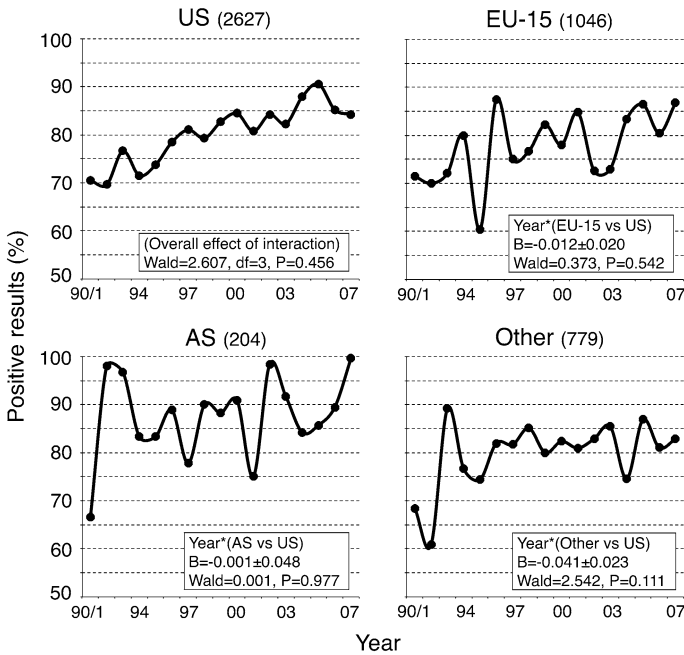
Fig. 5 Percentage of papers reporting a support for the tested hypothesis plotted against year of publication and divided by geographical location of corresponding author (US = United States; EU-15 = Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Netherlands, Portugal, Spain, Sweden, United Kingdom; AS = China, Hong Kong, India, Japan, Singapore, South Korea, Taiwan). Logistic regression estimates are interaction effects in a hierarchically well-formulated model. The main effects of this model, calculated with interaction components removed, are reported in Table 2. Numbers in brackets are sample size. Smoothed interpolation lines were added to help visualize trends

only examined the frequency of a sentence. This might have been an unreliable proxy of publication bias, as suggested by the fact that it yielded very high rates of non-significant results, contradicting ample evidence that these are the minority in all fields (Pautasso 2010). The reliability of the present study's approach, which assessed the actual conclusions of each paper, is supported by a close agreement with previous surveys that found statistically significant results to be around 95% in psychology, 91% in ecology, and between 85 and 96% in biomedicine (Sterling et al. 1995; Csada et al. 1996; Kyzas et al. 2007).

An important limitation of the present study was the use of only one journal database, a choice made to ensure coverage of all domains and unambiguous attribution of each paper to one discipline. The ESI database is a subset of the ISI-Web of Knowledge, which is currently the main source of bibliometric and citation data for research evaluation around the world. The ISI system has been criticised in the past for over-representing journals from the US (Shelton et al. 2007), and for expanding more slowly than the actual growth of the scientific literature (Larsen and von Ins 2010). Such criticisms must be taken into account when evaluating the generality of this study, but cannot undermine its conclusions. A North-American bias within the database might be supported by this study's data—in which over 50% of all papers had the corresponding author based in the US—but cannot explain away the various national patterns observed (see discussion below). The relatively

slow growth of the database would imply that it is covering a decreasing proportion of 'core' journals, amidst an expanding volume of publications (Larsen and von Ins 2010). Could negative results be increasingly published in journals not included in the ESI database? This possibility remains to be tested, but it appears unlikely, given that a similar study on abstracts in other databases (see above) reached identical conclusions (Pautasso 2010). In any case, a growing positive-outcome bias within ESI-indexed journals, which supposedly cover the most important publications and most of the citations in each discipline, would still reflect important changes occurring within the scientific system.

Excluding methodological biases, what caused the patterns observed? The likelihood for a study to publish a positive result depends essentially on three factors (Fanelli 2010b), which we will examine in turn. (1) The hypotheses tested might be increasingly likely to be true. Obviously, this would not happen because sciences are closer to the truth today than 20 years ago, but because researchers might be addressing hypotheses that are likely to be confirmed, to make sure they will get "publishable" results. (2) The average statistical power of studies might have increased (for example, if the average sample size of studies had increased), boosting the discovery rate of true relationships (Ioannidis 2005). This would be good news, suggesting an improvement of methods and quality of studies. However, it would be unlikely to explain alone all the patterns observed (e.g., differences between disciplines). Moreover, it is unsupported: statistical power appears to be very low in all fields, and there is no evidence that it has grown over the years (Delong and Lang 1992; Jennions and Moller 2003; Maddock and Rossi 2001). (3) Negative results could be submitted and accepted for publication less frequently, or somehow turned into positive results through post hoc re-interpretation, re-analysis, selection or various forms of manipulation/fabrication.

In the lightest scenario of hypothesis 3, changes would be occurring only in how results are written up: "tests" would be increasingly mentioned in the paper only when the results are positive, and negative results would be either embedded in "positive" papers or presented as positive by inverting the original hypothesis. Such scenario, which would still be the symptom of growing pressures to present a positive outcome, was not supported by the data. In almost all papers examined, the hypotheses were stated in the traditional form, with the null hypothesis representing a "no effect". There was no evidence that negative results were increasingly embedded in papers reporting positive ones: papers listing multiple hypotheses were more likely to report a negative support for the first one listed (Tables 1, 2), but their frequency has not grown significantly over the years ($B = 0.019 \pm 0.013$, Wald $= 2.058$, $P = 0.151$, power to detect a small and medium effect $= 0.543$ and 0.999). There was also no evidence that negative results are communicated in other form, such as conference proceedings: a sample of these latter was initially included in the analysis by mistake ($N = 106$), and they tended to report more positive results ($X^2 = 3.289$, $df = 1$, $P = 0.076$, power to detect a small effect $= 0.999$).

Higher frequencies of positive results from non-English speaking or non-US countries have been observed in past meta-analyses, and were usually attributed to editorial and peer-review biases, which might tend to reject papers from certain countries unless they present particularly strong or appealing results (Song et al. 2010; Yousefi-Nooraie et al. 2006). This could explain the higher rate of positive results from Asian countries, but cannot explain why the US have more positive results than the UK—an equally developed and English-speaking country. An editorial bias favouring the US would allow them to publish as many or more negative results than any other country, not fewer. Therefore, the differences observed suggest that researchers in the US have a stronger bias against negative results than in Europe. This hypothesis remains to be fully tested, but it would be

independently supported by at least two studies, one showing that the US have a higher proportion of retractions due to data manipulation (Steen 2011), and the other suggesting a higher publication bias among union-productivity studies from the US (Doucouliagos et al. 2005). The causes of these differences remain to be understood, one possible factor being higher pressures to publish imposed by the US research system.

A common argument against concerns for publication bias is that negative results are justifiably ignored per se but become interesting, and are published, when they contradict important predictions and/or previous positive evidence—ensuring self-correction of the literature in the long run (Silvertown and McConway 1997). This does indeed seem to be the case at least in some biomedical fields, where the first paper to report a finding often shows extreme effects that subsequent replications reduce or contradict entirely (Ioannidis et al. 2001; Ioannidis and Trikalinos 2005). However, even if in the long run truth will prevail, in the short term resources go wasted in pursuing exaggerated or completely false findings (Ioannidis 2006). Moreover, this self-correcting principle will not work efficiently in fields where theoretical predictions are less accurate, methodologies less codified, and true replications rare. Such conditions increase the rate of both false positives and false negatives, and a research system that suppresses the latter will suffer the most severe distortions. This latter concern was supported by the finding that positive results were more frequent and had increased more rapidly in the social and many biological sciences [where theories and methods tend to be less codified and replication is rare (Fanelli 2010b; Schmidt 2009; Evanschitzky et al. 2007; Tsang and Kwan 1999; Kelly 2006; Palmer 2000; Jones et al. 2010; Hubbard and Vetter 1996)].

In conclusion, it must be emphasised that the strongest increase in positive results was observed in disciplines—like Clinical Medicine, Pharmacology & Toxicology, Molecular Biology—where concerns for publication bias had a longer history and several initiatives to prevent and correct it have been attempted, including registration of clinical trials, enforcing guidelines for accurate reporting, and creating journals of negative results (Bian and Wu 2010; Simera et al. 2010; Kundoor and Ahmed 2010; Knight 2003). This study suggests that such initiatives have not met their objectives so far, and the problem might be worsening.

Methods

Data collection

The sentence "test* the hypothes*" was used to search all 10,837 journals available in the Essential Science Indicators database in December 2008, which classifies journals univocally in 22 disciplines (for ESI classification methodology see http://sciencewatch.com/about/met/). The discipline of mathematics, however, yielded no usable paper, while the "multidisciplinary" category, including papers such as Science or Nature, was excluded. Therefore, papers from 20 disciplines were included in the analysis. The disciplines were grouped in the following domains: Physical Sciences = Space Science, Chemistry, Computer Science, Engineering, Geosciences, Materials Science, Physics; Biological Sciences = Agricultural Sciences, Biology & Biochemistry, Clinical Medicine, Environment/Ecology, Immunology, Molecular Biology & Genetics, Microbiology, Neuroscience & Behaviour, Plant and Animal Sciences, Pharmacology & Toxicology; Social Sciences = Economics & Business, Psychiatry/Psychology, Social Sciences, General).

Papers were sampled in two phases: (1) papers published between 2000 and 2007 (already used in previous studies); (2) papers published between 1990 and 1999. In both

phases, all retrieved titles were saved on bibliographic database software, and then up to a maximum 150 papers were sampled from each discipline. When the number of titles retrieved from one discipline exceeded 150, papers were selected using a random number generator. In one discipline, Plant and Animal Sciences, an additional 50 papers from the period 2000–2007 were analysed, in order to increase the statistical power.

By examining the abstract and/or full-text, the specific hypothesis tested in each paper was identified, and it was determined whether the authors had concluded to have found a positive (full or partial) or negative (null or negative) support. If more than one hypothesis was being tested, only the first one listed in the text was considered.

Meeting abstracts were excluded from sampling, whilst sampled papers were excluded when they either did not test a hypothesis (Total $N = 546$) or when there was not sufficient information (abstract unclear, and full-text not available) to determine the outcome (Total $N = 38$). While the former have no role in the analysis, the latter are technically missing values. Since access to full-text was lower for older articles and some disciplines, these missing values were unevenly distributed between disciplines ($X^2 = 92.770$, $P < 0.001$), and were negatively associated with year ($B = -0.080 \pm 0.035$, Wald $= 5.352$, $P = 0.021$). However, we can exclude that these missing values are an important confounding factor for three reasons: (1) there is no reason to believe that these missing papers are more likely to report positive than negative results; (2) they represent a very small fraction of the sample (i.e., 0.8%); (3) their prevalence is higher until 1994 and then declines rapidly, not matching the observed increase in positive results.

All data was extracted by the author. An untrained assistant who was given basic written instructions scored papers the same way as the author in 18 out of 20 cases, and picked up exactly the same sentences for hypothesis and conclusions in all but three cases. The discrepancies were easily explained, showing that the procedure is objective and replicable.

The country of location of each paper was attributed based on the address of the corresponding author. Geographical location was defined by the following groupings: US = United States; EU-15 = Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Netherlands, Portugal, Spain, Sweden, United Kingdom; AS = China, Hong Kong, India, Japan, Singapore, South Korea, Taiwan).

Information on year of publication and country was retrieved after all papers had been coded. Therefore, the coding of papers as "positive" and "negative" was completely blind to year and country of origin.

Statistical analyses

The ability of independent variables to predict the outcome of a paper was tested with a general linear model assuming a binomial link function (i.e., logistic regression) in the form:

$$Y = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_n X_{in}$$

in which p_i is the probability of the ith paper of reporting a positive result, X_1 is the year of publication (with the years 1990 and 1991 combined due to low sample size for those years) and X_n represents the various characteristics of the ith paper that were controlled for in each model as specified in the text (e.g., dummy variables for discipline, country etc.…). Statistical significance of the effect of each variable was calculated through Wald's test. The relative fit of regression models was estimated with Nagelkerke's adjusted R^2. To

ensure robustness of results, a model simplification approach was employed, fitting a complete generalized linear model with all main effects and interactions, and backward-removing terms that did not significantly increase the deviance explained.

Post hoc statistical power for indicator contrasts in logistic regression was calculated for main effects only (not interactions) assuming a bimodal distribution and sample frequency equal to that of the categorical variable with the smallest N (each case is specified in the text), to estimate the minimum power available. Base rate variance was measured with Nagelkerke R^2 after removing the categorical variables of interest from the model. Post hoc power analysis for the effect of year assumed a standard uniform distribution of papers across years. Small, medium and large effects were assumed to equal Odds-Ratio = 1.5, 2.5 and 4.5, respectively.

All analyses were produced using the statistical packages R 2.12, SPSS 17.0 and G*Power 3.1.

References

Atkin, P. A. (2002). A paradigm shift in the medical literature. *British Medical Journal, 325*(7378), 1450–1451.

Bian, Z. X., & Wu, T. X. (2010). Legislation for trial registration and data transparency. *Trials, 11*, 64. doi: 10.1186/1745-6215-11-64.

Bonitz, M., & Scharnhorst, A. (2001). Competition in science and the Matthew core journals. *Scientometrics, 51*(1), 37–54.

Browman, H. I. (1999). The uncertain position, status and impact of negative results in marine ecology: Philosphical and practical considerations. *Marine Ecology Progress Series, 191*, 301–309.

Csada, R. D., James, P. C., & Espie, R. H. M. (1996). The "file drawer problem" of non-significant results: Does it apply to biological research? *Oikos, 76*(3), 591–593.

de Meis, L., Velloso, A., Lannes, D., Carmo, M. S., & de Meis, C. (2003). The growing competition in Brazilian science: Rites of passage, stress and burnout. *Brazilian Journal of Medical and Biological Research, 36*(9), 1135–1141.

De Rond, M., & Miller, A. N. (2005). Publish or perish—Bane or boon of academic life? *Journal of Management Inquiry, 14*(4), 321–329. doi:10.1177/1056492605276850.

Delong, J. B., & Lang, K. (1992). Are all economic hypotheses false. *Journal of Political Economy, 100*(6), 1257–1272.

Doucouliagos, H., Laroche, P., & Stanley, T. D. (2005). Publication bias in union-productivity research? *Relations Industrielles-Industrial Relations, 60*(2), 320–347.

Dwan, K., Altman, D. G., Arnaiz, J. A., Bloom, J., Chan, A.-W., Cronin, E., et al. (2008). Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLoS ONE, 3*(8), e3081. [Research Support, Non-U.S. Gov't; Review].

Evanschitzky, H., Baumgarth, C., Hubbard, R., & Armstrong, J. S. (2007). Replication research's disturbing trend. *Journal of Business Research, 60*(4), 411–415. doi:10.1016/j.jbusres.2006.12.003.

Fanelli, D. (2010a). Do pressures to publish increase scientists' bias? An empirical support from US States Data. *Plos One, 5*(4), e10271. doi:10.1371/journal.pone.0010271.

Fanelli, D. (2010b). "Positive" results increase down the hierarchy of the sciences. *Plos One, 5*(3), e10068. doi:10.1371/journal.pone.0010068.

Feigenbaum, S., & Levy, D. M. (1996). Research bias: Some preliminary findings. *Knowledge and Policy: The International Journal of Knowledge Transfer and Utilization, 9*(2 & 3), 135–142.

Formann, A. K. (2008). Estimating the proportion of studies missing for meta-analysis due to publication bias. *Contemporary Clinical Trials, 29*(5), 732–739. doi:10.1016/j.cct.2008.05.004.

Fronczak, P., Fronczak, A., & Holyst, J. A. (2007). Analysis of scientific productivity using maximum entropy principle and fluctuation-dissipation theorem. *Physical Review E, 75*(2), 026103. doi: 10.1103/PhysRevE.75.026103.

Gad-el-Hak, M. (2004). Publish or perish—An ailing enterprise? *Physics Today, 57*(3), 61–62.

Gerber, A. S., & Malhotra, N. (2008). Publication bias in empirical sociological research—Do arbitrary significance levels distort published results? *Sociological Methods & Research, 37*(1), 3–30.

Howard, G. S., Hill, T. L., Maxwell, S. E., Baptista, T. M., Farias, M. H., Coelho, C., et al. (2009). What's wrong with research literatures? And how to make them right. *Review of General Psychology, 13*(2), 146–166.

Hubbard, R., & Vetter, D. E. (1996). An empirical comparison of published replication research in accounting, economics, finance, management, and marketing. *Journal of Business Research, 35*(2), 153–164.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *Plos Medicine, 2*(8), 696–701.

Ioannidis, J. P. A. (2006). Evolution and translation of research findings: From to where? *Plos Clinical Trials, 1*, e36. doi:10.1371/journal.pctr.0010036.

Ioannidis, J. P. A. (2008a). Perfect study, poor evidence: Interpretation of biases preceding study design. *Seminars in Hematology, 45*(3), 160–166.

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

Ioannidis, J. P. A., Ntzani, E. E., Trikalinos, T. A., & Contopoulos-Ioannidis, D. G. (2001). Replication validity of genetic association studies. *Nature Genetics, 29*(3), 306–309.

Ioannidis, J. P. A., & Trikalinos, T. A. (2005). Early extreme contradictory estimates may appear in published research: The proteus phenomenon in molecular genetics research and randomized trials. *Journal of Clinical Epidemiology, 58*(6), 543–549.

Jeng, M. (2006). A selected history of expectation bias in physics. *American Journal of Physics, 74*(7), 578–583.

Jennions, M. D., & Moller, A. P. (2002). Publication bias in ecology and evolution: An empirical assessment using the 'trim and fill' method. *Biological Reviews, 77*(2), 211–222.

Jennions, M. D., & Moller, A. P. (2003). A survey of the statistical power of research in behavioral ecology and animal behavior. *Behavioral Ecology, 14*(3), 438–445.

Jones, K. S., Derby, P. L., & Schmidlin, E. A. (2010). An investigation of the prevalence of replication research in human factors. *Human Factors, 52*(5), 586–595. doi:10.1177/0018720810384394.

Kelly, C. D. (2006). Replicating empirical research in behavioral ecology: How and why it should be done but rarely ever is. *Quarterly Review of Biology, 81*(3), 221–236.

King, D. A. (2004). The scientific impact of nations. *Nature, 430*(6997), 311–316. doi:10.1038/430311a.

Knight, J. (2003). Negative results: Null and void. *Nature, 422*(6932), 554–555.

Kundoor, V., & Ahmed, M. K. K. (2010). Uncovering negative results: Introducing an open access journal "Journal of Pharmaceutical Negative Results". *Pharmacognosy Magazine, 6*(24), 345–347. doi:10.4103/0973-1296.71783.

Kyzas, P. A., Denaxa-Kyza, D., & Ioannidis, J. P. A. (2007). Almost all articles on cancer prognostic markers report statistically significant results. *European Journal of Cancer, 43*(17), 2559–2579.

Larsen, P. O., & von Ins, M. (2010). The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics, 84*(3), 575–603. doi:10.1007/s11192-010-0202-z.

Lawrence, P. A. (2003). The politics of publication—Authors, reviewers and editors must act to protect the quality of research. *Nature, 422*(6929), 259–261. doi:10.1038/422259a.

Lortie, C. J. (1999). Over-interpretation: Avoiding the stigma of non-significant results. *Oikos, 87*(1), 183–184.

Maddock, J. E., & Rossi, J. S. (2001). Statistical power of articles published in three health psychology-related journals. *Health Psychology, 20*(1), 76–78.

Marsh, D. M., & Hanlon, T. J. (2007). Seeing what we want to see: Confirmation bias in animal behavior research. *Ethology, 113*(11), 1089–1098.

Meho, L. I. (2007). The rise and rise of citation analysis. *Physics World, 20*(1), 32–36.

Nicolini, C., & Nozza, F. (2008). Objective assessment of scientific performances world-wide. *Scientometrics, 76*(3), 527–541. doi:10.1007/s11192-007-1786-9.

Osuna, C., Crux-Castro, L., & Sanz-Menedez, L. (2011). Overturning some assumptions about the effects of evaluation systems on publication performance. *Scientometrics, 86*, 575–592.

Palmer, A. R. (2000). Quasireplication and the contract of error: Lessons from sex ratios, heritabilities and fluctuating asymmetry. *Annual Review of Ecology and Systematics, 31*, 441–480.

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.

Qiu, J. (2010). Publish or perish in China. *Nature, 463*(7278), 142–143. doi:10.1038/463142a.

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology, 13*(2), 90–100. doi:10.1037/a0015108.

Shelton, R. D., Foland, P., & Gorelskyy, R. (2007). Do new SCI journals have a different national bias? *Proceedings of ISSI 2007: 11th international conference of the international society for scientometrics and informetrics, Vols I and II* (pp. 708–717).

Shelton, R. D., Foland, P., & Gorelskyy, R. (2009). Do new SCI journals have a different national bias? *Scientometrics, 79*(2), 351–363. doi:10.1007/s11192-009-0423-1.

Silvertown, J., & McConway, K. J. (1997). Does "publication bias" lead to biased science? *Oikos, 79*(1), 167–168.

Simera, I., Moher, D., Hirst, A., Hoey, J., Schulz, K. F., & Altman, D. G. (2010). Transparent and accurate reporting increases reliability, utility, and impact of your research: Reporting guidelines and the EQUATOR Network. *Bmc Medicine, 8*, 24. doi:10.1186/1741-7015-8-24.

Song, F., Parekh, S., Hooper, L., Loke, Y. K., Ryder, J., Sutton, A. J., et al. (2010). Dissemination and publication of research findings: An updated review of related biases. *Health Technology Assessment, 14*(8), 1–193. doi:10.3310/hta14080.

Statzner, B., & Resh, V. H. (2010). Negative changes in the scientific publication process in ecology: Potential causes and consequences. *Freshwater Biology, 55*(12), 2639–2653. doi:10.1111/j.1365-2427.2010.02484.x.

Steen, R. G. (2011). Retractions in the scientific literature: Do authors deliberately commit research fraud? *Journal of Medical Ethics, 37*(2), 113–117.

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited—The effect of the outcome of statistical tests on the decision to publish and vice versa. *American Statistician, 49*(1), 108–112.

Tsang, E. W. K., & Kwan, K. M. (1999). Replication and theory development in organizational science: A critical realist perspective. *Academy of Management Review, 24*(4), 759–780.

Warner, J. (2000). A critical review of the application of citation studies to the Research Assessment Exercises. *Journal of Information Science, 26*(6), 453–459.

Young, N. S., Ioannidis, J. P. A., & Al-Ubaydi, O. (2008). Why current publication practices may distort science. *Plos Medicine, 5*(10), 1418–1422. doi:10.1371/journal.pmed.0050201.

Yousefi-Nooraie, R., Shakiba, B., & Mortaz-Hejri, S. (2006). Country development and manuscript selection bias: A review of published studies. *BMC Medical Research Methodology, 6*, 37.