A surge of *p*-values between 0.040 and 0.049 in recent decades (but negative results are increasing rapidly too)

J. C. F. de Winter, D. Dodou

Department of BioMechanical Engineering, Delft University of Technology, Delft, The Netherlands

Abstract

It is known that statistically significant results are more likely to be published than results that are not statistically significant. However, it is unclear whether negative results are disappearing from papers, and whether there exists a 'hierarchy of sciences' with the social sciences publishing more positive results than the physical sciences. Using Scopus, we conducted a search in the abstracts of papers published between 1990 and 2014, and calculated the percentage of papers reporting marginally positive results (i.e., *p*-values between 0.040 and 0.049) versus the percentage of papers reporting marginally negative results (i.e., *p*-values between 0.051 and 0.060). The results indicate that negative results are not disappearing, but have actually become 4.3 times more prevalent since 1990. Positive results, on the other hand, have become 13.9 times more prevalent since 1990. We found no consistent support for a 'hierarchy of sciences'. However, we did find large differences in reporting practices between disciplines, with the reporting of *p*-values being 60.6 times more frequent in the biological sciences than in the physical sciences. We argue that the observed longitudinal trends may be caused by negative factors, such as an increase of questionable research practices, but also by positive factors, such as an increasingly quantitative research focus.

1. Introduction

In the last decade, many methodologists have raised concerns about the skewed nature of the scientific record. Ioannidis' (2005) highly-cited article claimed that over 50% of the results that are declared statistically significant are false, meaning that they actually reflect a negative (i.e., null) effect. Similar voices are heard in a variety of research fields, including biology and ecology (Csada et al., 1996; Jennions & Møller, 2002), medicine and pharmaceutics (Atkin, 2002; Colom & Vieta, 2011; Dwan et al., 2008; Hopewell et al., 2009; Kyzas et al., 2007), economics (Ioannidis & Doucouliagos, 2013), cognitive sciences (Ioannidis et al., 2014), genetics (Ioannidis, 2003), neurosciences (Jennings & Van Horn, 2012), and psychology (Ferguson & Heene, 2012; Francis, 2013; Laws, 2013).

The abundance of positive results has been attributed to questionable research practices such as selective publication (Dwan et al., 2008; Hopewell et al., 2009; Rothstein et al., 2006), undisclosed exploratory analyses and selective reporting (Chan et al., 2014; Dwan et al., 2008; Kirkham et al., 2010; Simmons et al., 2011), as well as data fabrication (Fanelli, 2009; Moore et al., 2010). These mechanisms are fuelled by an emphasis on productivity (De Rond & Miller, 2005), high rejection rates of journals (Young et al., 2008), and competitive schemes for funding and promotion (Joober et al., 2012). Not just researchers, but also journal editors (Sterling et al., 1995; Thornton & Lee, 2000) and sponsoring/funding parties (Djulbegovic et al., 2000; Lexchin et al., 2003; Sismondo, 2008) have been criticised for favouring positive results over negative ones. It has been argued that certain fields within the social and medical sciences are currently in crisis, meaning that there are so many false positives published that the credibility of entire disciplines is at stake (Kahneman, 2013; Pashler & Harris, 2012; Rouder et al., 2014).

As a reaction to the "excess significance bias" (Ioannidis, 2011), methodologists have emphasized that null results should not remain in the file drawer, and that the decision to publish should be based on methodological soundness rather than novelty or statistical significance (Asendorpf et al., 2013; Dirnagl & Lauritzen, 2010). Methodologists have also warned of the perils of exploratory research, and have encouraged preregistration of research protocols in an attempt to prevent spurious positive findings (Asendorpf et al., 2013; De Angelis et al., 2004). Furthermore, statistical corrections have been introduced that *decrease* observed effects, including corrections for publication bias (Duval & Tweedie, 2000; Rücker et al., 2008; Terrin et al., 2003) and credibility calibration (Ioannidis, 2008).

It is worth noting that in the 1980s and 1990s, the social sciences were also said to be in a crisis. Funding agents threatened to cut budgets, because they were frustrated with psychology's ongoing production of small and inconsistent effect sizes (Hunter & Schmidt, 1996). As an answer, methodologists introduced several artefact corrections (i.e., corrections for range restriction, measurement error, and dichotomization). These artefact corrections almost always *increase* the observed effect sizes (Fern & Monroe, 1996; Schmidt & Hunter, 1996).

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.447v3 | CC-BY 4.0 Open Access | received: 2 Aug 2014, published: 2 Aug

Akin to signal detection theory, measures that decrease false positives will lead to more false negatives (Fiedler et al., 2012). Hence, we ought to ask ourselves whether the alleged crisis reflects the true state of affairs. Have positive results really become more prevalent in recent decades? Furthermore, it is important to determine whether the number of positive results has dropped recently, which could indicate that the methodological recommendations have been effective.

So far, research on longitudinal trends in positive versus negative results has been scarce. An exception is Fanelli (2012), who manually coded 4,656 journal papers. In his article "Negative results are disappearing from most disciplines and countries", Fanelli found that the number of papers providing support for the main hypothesis had increased from 70% in 1990 to 86% in 2007 (it is unclear why Fanelli reported an over 22% increase in the abstract). Fanelli further concluded that the increase was significantly stronger in the social sciences and some biomedical fields than in the physical sciences. He also reported that Asian countries have been producing more positive results than the United States, which in turn have been producing more positive results than European countries.

Fanelli's (2012) paper has some important limitations. First, although his sample size is impressive (considering that coding was done manually), the statistical power does not seem large enough for assessing *differences in growth* between disciplines or countries. We extracted and re-analyzed data shown in Fanelli's figures and found that the 95% confidence intervals (CI) of the regression slopes are overlapping between disciplines (1.44%/year for the social sciences [CI = 0.99, 1.90], 0.92%/year for the biological sciences [CI = 0.53, 1.32], and 0.65%/year for the physical sciences [CI = 0.19, 1.11], cf. Fig. 1). It is unclear how Fanelli (2012) arrived at the conclusion that "the increase was stronger in the social and some biomedical disciplines". Fanelli (2012, see also 2010) claims support for a hierarchy of sciences with physics at the top and social sciences at the bottom, and positive results increasing toward the hierarchy's lower end. However, the results shown in Fig. 1 do not provide statistically convincing support for such hierarchy, with the average percentage of positive results over the period 1991–2007 being 78.3 in physical sciences, 80.1 in biological sciences, and 81.5 in social sciences. A second limitation is that Fanelli's (2012) assessment of positive/negative results is based on the reading of abstracts or full-texts by the author himself. This approach could have introduced bias, especially because Fanelli's coding was not blind to the scientific discipline that the papers belong too. Randomization issues are at play as well, because the coding was first done for papers published between 2000 and 2007, and subsequently for papers published between 1990 and 1999.

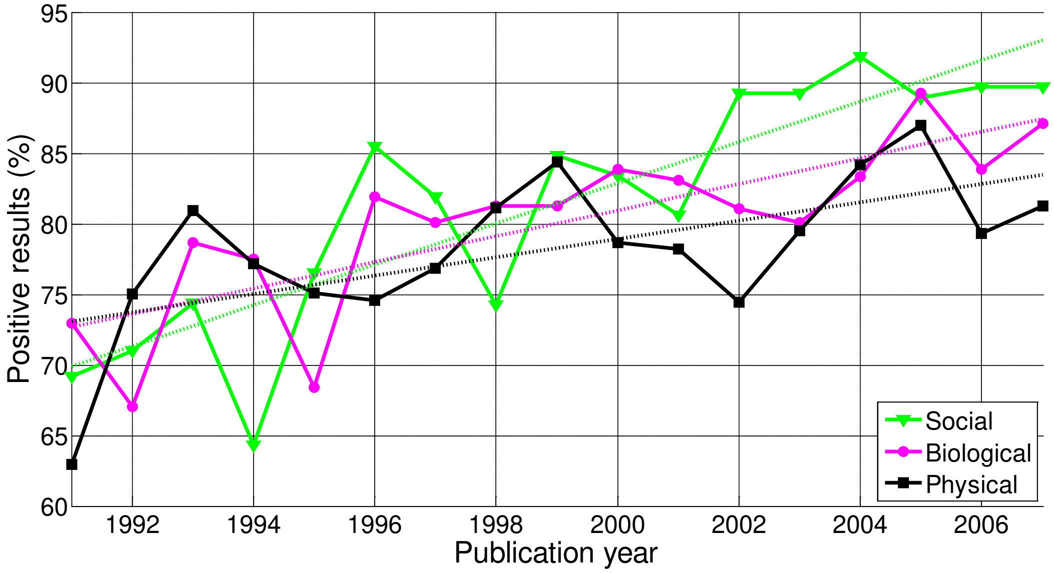

Figure 1. Number of papers reporting a positive result divided by the total number of papers examined (i.e., papers reporting a positive result + papers reporting a negative result) per publication year, for three scientific disciplines. The figure has been created by graphically extracting the data shown in Fanelli's (2012) figures. Dashed lines represent the results of a linear regression analysis.

Pautasso (2010) also studied longitudinal trends in positive versus negative research findings. Opposite to Fanelli's (2012) manual coding, Pautasso searched for the phrases "significant difference" versus "no significant difference" (and variants thereof) in the title and abstract of papers in the (Social) Science Citation Index for 1991–2008, and in

CAB Abstracts and Medline for 1970–2008. Pautasso (2010) found that the prevalence of both positive and negative results has increased over time. At the face of it, Pautasso's results contradict those of Fanelli (2012), because the former reported a clear *increase* of negative results with publication year, whereas the latter stated that negative results are disappearing. Where both authors agree is that the *ratio* of positive to negative results has been increasing over time.

Some of Pautasso's (2010) other conclusions also appear to contradict Fanelli (2012). For example, Pautasso found that the "worsening file-drawer problem" was not apparent for papers retrieved with the keyword 'psychol*', and that the effect was *weaker* in the Social Science Citation Index (ratio of non-significant to significant results reduced from 1.00 in the 1990s to 0.94 in the 2000s) than in the Science Citation Index (ratio of non-significant to significant results reduced from 1.67 in the 1990s to 1.53 in the 2000s; data extracted from Pautasso's figures).

Both Fanelli's (2012) and Pautasso's (2010) analyses require updating, as these studies cover a period up to 2007 and 2008, respectively. Considering the alleged crisis and the exponential growth of the scientific literature (Larsen & Von Ins, 2010), a modern replication of these works seems warranted. Replication is also needed because Fanelli's (2012) work has received ample citations and attention from the popular press (cf. Yong, 2012, stating in *Nature* that psychology and psychiatry are the "worst offenders", based on Fanelli, 2012).

In summary, it is well known that positive results (i.e., results that are statistically significant or in agreement with a hypothesis) are more likely to be published than negative (i.e., null) results (e.g., Hopewell et al., 2009; Smart, 1964; Sterling et al., 1995). However, it is unclear whether the prevalence of negative results is decreasing over time, whether the increase of positive results is stronger in the softer disciplines (i.e., social sciences) as compared to the harder disciplines (i.e., physical sciences), and whether different regions in the world exhibit different tendencies in reporting positive versus negative results.

The aim of this study was to estimate longitudinal trends of positive versus negative results in the scientific literature, and to compare these trends between disciplines and countries. We chose for an automatic search, akin to Pautasso (2010). The quantitative analysis of digitized texts, also known as 'culturomics', has become an established method for investigating secular trends in cultural phenomena (e.g., Michel et al., 2011). Automated analysis is advantageous, because it is free of human biases in coding.

We conducted searches for marginally significant *p*-values (i.e., *p*-values between 0.040 and 0.049) versus marginally non-significant *p*-values (i.e., *p*-values between 0.051 and 0.060). The tacit assumption is that marginally significant *p*-values are often the result of selective analysis and reporting (also called "*p*-hacking" or "fiddling"), whereas marginally non-significant results reflect decent unbiased science done by researchers who resist data massaging (Gadbury & Allison, 2012; Gelman & Loken, 2013; Gerber & Malhotra, 2008; Masicampo & Lalande, 2012; Ridley et al., 2007). We also searched for qualitative statements of statistical significance (i.e., "significant difference" vs. "no significant difference") to replicate Pautasso's (2010) method.

2. Methods

Our searchers were conducted with Elsevier's Scopus. After trying out other search engines (i.e., Web of Science and Google Scholar), we concluded that Scopus offers the most accurate and powerful search and export features.

Scopus classifies papers into 27 subject areas; we grouped these into three scientific disciplines, each discipline including subject areas comparable to Fanelli's (2012) classification, for the sake of replication

1) *Physical sciences*, including the Scopus subject areas of Physics and Astronomy; Engineering; Earth and Planetary Sciences; Chemistry; Chemical Engineering; Materials Science; Energy; Computer Science.
2) *Biological sciences*, including the Scopus subject areas of Agricultural and Biological Sciences; Biochemistry, Genetics and Molecular Biology; Medicine; Neuroscience; Immunology and Microbiology; Pharmacology, Toxicology and Pharmaceutics; Veterinary; Environmental Science; Dentistry.
3) *Social Sciences*, including the Scopus subject areas of Social Sciences; Psychology; Arts and Humanities; Economics, Econometrics and Finance; Decision Sciences; Business, Management and Accounting.

The Mathematics and Multidisciplinary subject areas were excluded, like in Fanelli (2012). Health Professions and Nursing were excluded as well, because we were not sure which discipline they should be classified into; these two subject areas are relatively small anyway, accounting together for 2.5% of all records in Scopus, and it is unlikely that they would have affected our results. Note that a paper can belong to more than one subject area.

To investigate whether temporal changes differ between different regions of the world, we distinguished the following world regions as in Fanelli (2012), namely:

1) United States.
2) Fifteen European countries (EU15): Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, The Netherlands, Portugal, Spain, Sweden, and United Kingdom.
3) Seven Asian countries (AS7): China, Hong Kong, India, Japan, Singapore, South Korea, and Taiwan.

Note that a paper can belonging to multiple world regions (e.g., due to multiple authors with affiliations in countries from different world regions, or due to an author having multiple affiliations in countries from different world regions).

The following queries were conducted using the advanced search function of Scopus for the abstracts of all records in the database as well as for each discipline and world region defined above:

1) ABS({.}), to extract the total number of papers with an abstract.
2) Two queries about reporting of *p*-values, namely:
    a. A query containing *p*-values between 0.040 and 0.049, that is: ABS({p = 0.040} OR {p = .040} OR {p = 0.041} OR {p = .041}… OR {p = 0.049} OR {p = .049}), to extract the number of papers reporting a *p*-value between 0.040 and 0.049, that is, marginally below the typically used alpha value of 0.05.
    b. A query containing *p*-values between 0.051 and 0.060, that is: ABS({p = 0.051} OR {p = .051} OR {p = 0.052} OR {p = .052} … OR {p = 0.060} OR {p = .060}), to extract the number of papers reporting a *p*-value between 0.051 and 0.060, that is, marginally above the typically used alpha value of 0.05.
3) Two queries about textual reporting of statistical (non-)significance, namely:
    a. A query containing typical expressions for reporting significant differences, that is: ABS(({significant difference} OR {significant differences}) AND NOT ({no significant difference} OR {no significant differences} OR {no statistically significant difference} OR {no statistically significant differences})) (cf. Pautasso, 2010), to extract the number of papers with a qualitative statement of significant results.
    b. A query containing typical expressions for reporting no significant differences, that is: ABS({no significant difference} OR {no significant differences} OR {no statistically significant difference} OR {no statistically significant differences}), to extract the number of papers with a qualitative statement of non-significant results.

Note that in Scopus braces ({}) are used for exact searches, in which special characters such as punctuation marks and mathematical symbols are taken into consideration, whereas quotation marks ("") are used for more loose searches, neglecting special characters. All data were extracted on 19 July 2014.

The following measures were calculated for both the reporting of *p*-values and the textual reporting of statistical (non-)significance:
1) the number of papers reporting significant results divided by the total number of papers with an abstract per publication year;
2) the number of papers reporting non-significant results divided by the total number of papers with an abstract per publication year; and
3) the ratio of significant to non-significant results.
All three measures were calculated for all papers in the Scopus database, as well as for papers in each discipline and world region defined above.

All analyses were conducted for papers published in the period 1990–2014. Longitudinal trends were assessed by means of the coefficient estimates and corresponding 95% confidence intervals of a simple linear regression. The employed script is provided as supplementary material.

3. Results

3.1. Total number of papers

According to our searches, Scopus contains a total of 39,421,740 papers with an abstract. 18,726,024 papers belong to the biological sciences, 19,871,102 papers belong to the physical sciences, and 3,543,934 papers belong to the social sciences. U.S. was found in the affiliations of 10,235,548 papers, EU15 in 8,752,150 papers, and AS7 in

8,210,522 papers. The number of papers has increased over time in all three scientific disciplines (Fig. 2). The drop in 2014 can be explained by the fact that that we are currently halfway the year 2014.

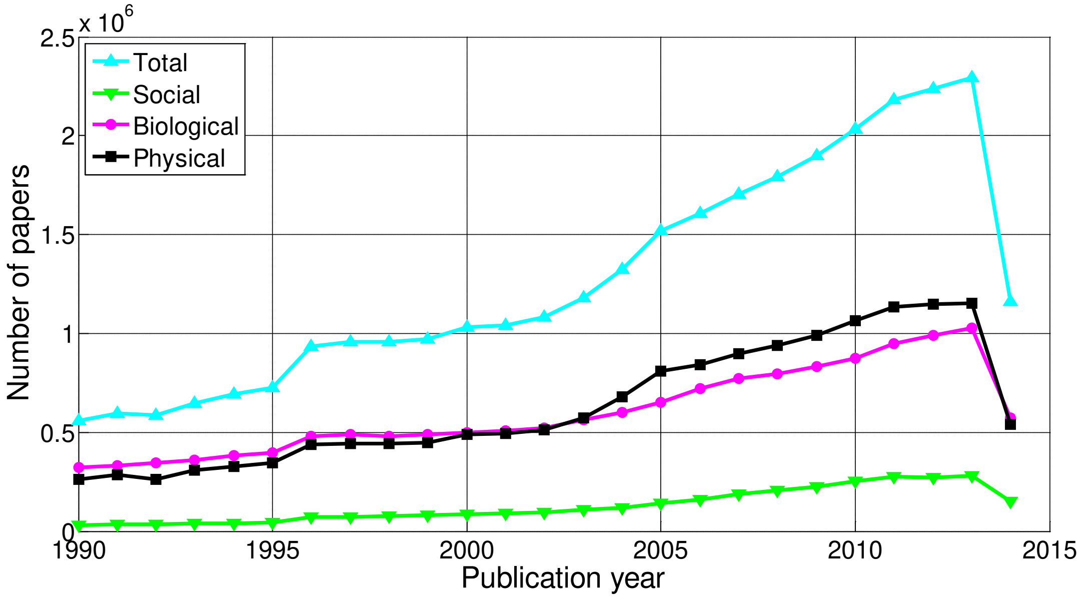

Figure 2. Number of papers per publication year, for three scientific disciplines.

3.2. Longitudinal trends

3.2.1. p-value reporting

Both the significant and non-significant results have increased over time (Fig. 3). In 1990, 0.019% of papers (107 out of 561,194 papers) reported a *p*-value between 0.051 and 0.060. This has risen about 4.3 fold to 0.082% (956 out of 1,161,405 papers) in 2014. Positive results, on the other hand, have increased 13.9-fold in the same period: from 0.031% (175 out of 561,194 papers) in 1990 to 0.432% (5,018 out of 1,161,405 papers) in 2014. In other words, the ratio of significant to non-significant results has increased from 1.6 (i.e., 175/107 papers) in 1990 to 5.2 (i.e., 5,018/956 papers) in 2014 (Fig. 4).

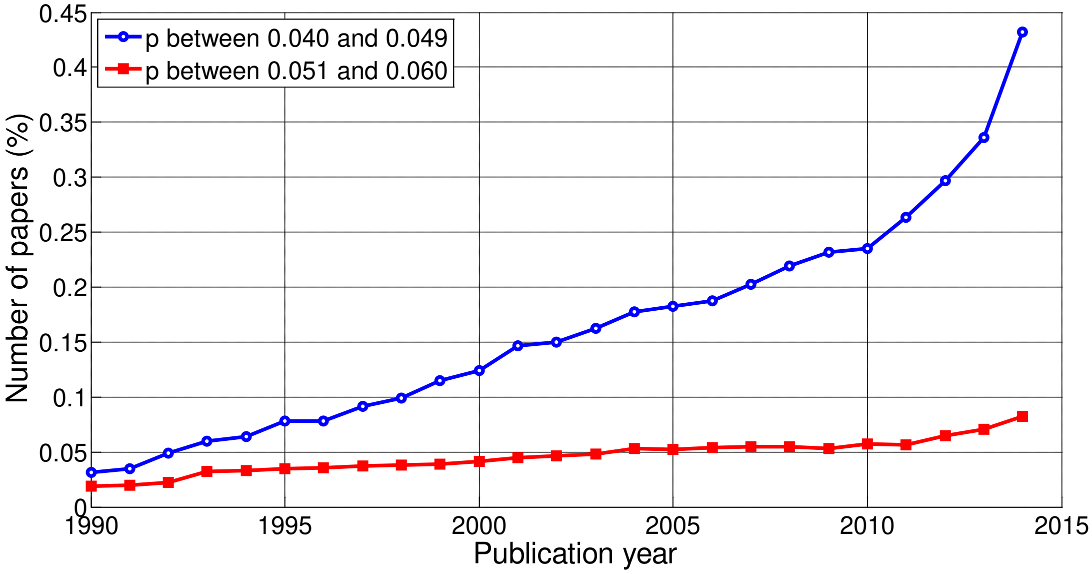

Figure 3. Number of papers reporting a *p*-value between 0.040 and 0.049 (blue line) or a *p*-value between 0.051 and 0.060 (red line) divided by the total number of papers per publication year.

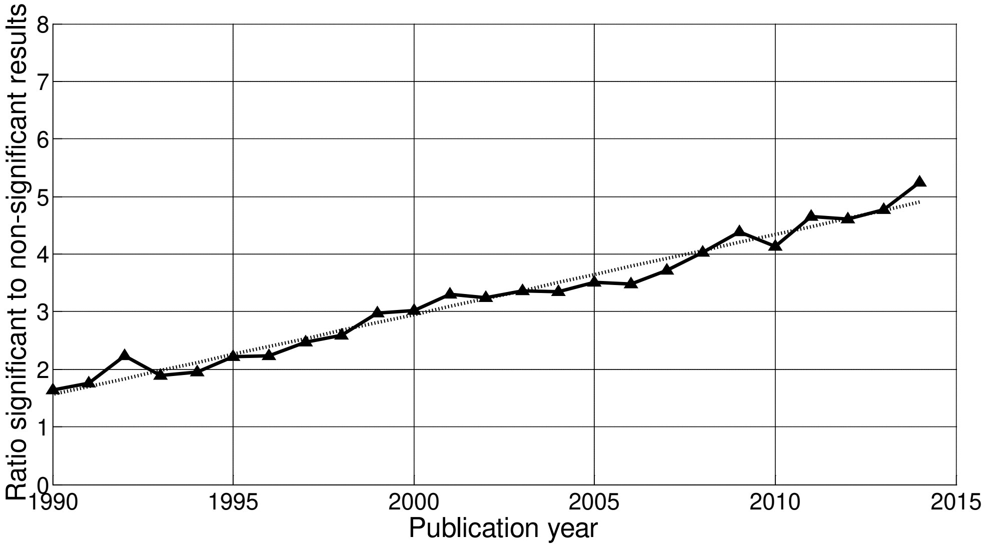

Figure 4. Ratio of significant to non-significant results (*p*-value between 0.040 and 0.049 / *p*-value between 0.051 and 0.060) per publication year. The dashed line represents the result of a linear regression analysis.

3.2.2. Textual reporting of statistical (non-)significance

The analysis of textual reporting of statistical significance and non-significance confirms the findings above on the reporting of *p*-values, namely that both significant and non-significant differences have increased over time (see Fig. S1 in supplementary material); However, the ratio between significant versus non-significant results is smaller and increases less rapidly (Fig 4. vs. Fig. S2). The phrase "no significant difference" is more frequent than the phrase "significant difference".

3.3. Comparison of longitudinal trends between scientific disciplines

3.3.1. p-value reporting

A comparison between disciplines shows that the use of *p*-values is rare in the physical sciences: in 2014, reporting of *p*-values in the social and biological sciences is respectively 6.9 times and 60.6 times more frequent than in the physical sciences (Fig. 5). The 95% confidence intervals of the regression slopes of the ratios of significant to non-significant results in the three disciplines are overlapping (Table 1, see also Fig. 6), indicating there is no difference in growth rates between the three disciplines.

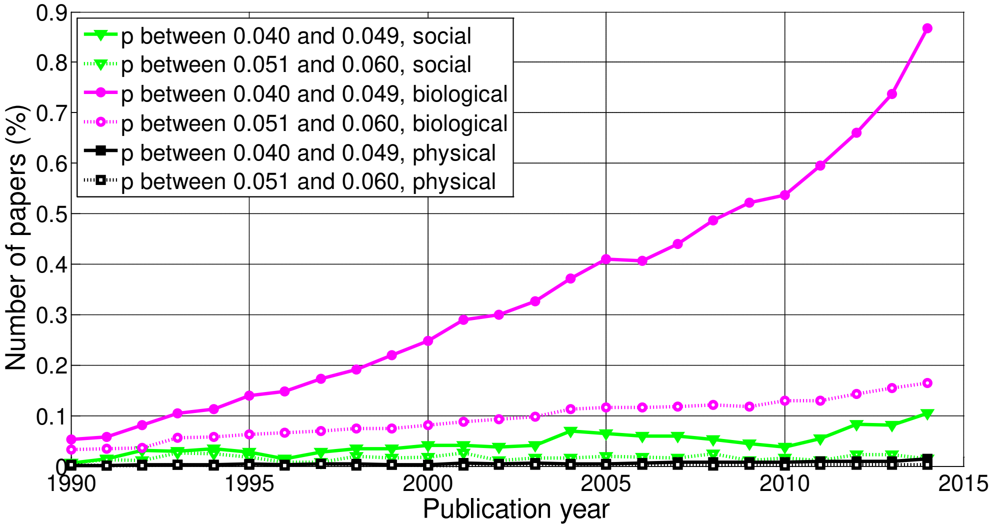

Figure 5. Number of papers reporting a *p*-value between 0.040 and 0.049 (solid lines) or a *p*-value between 0.051 and 0.060 (dotted lines) divided by the total number of papers per publication year, for three scientific disciplines.

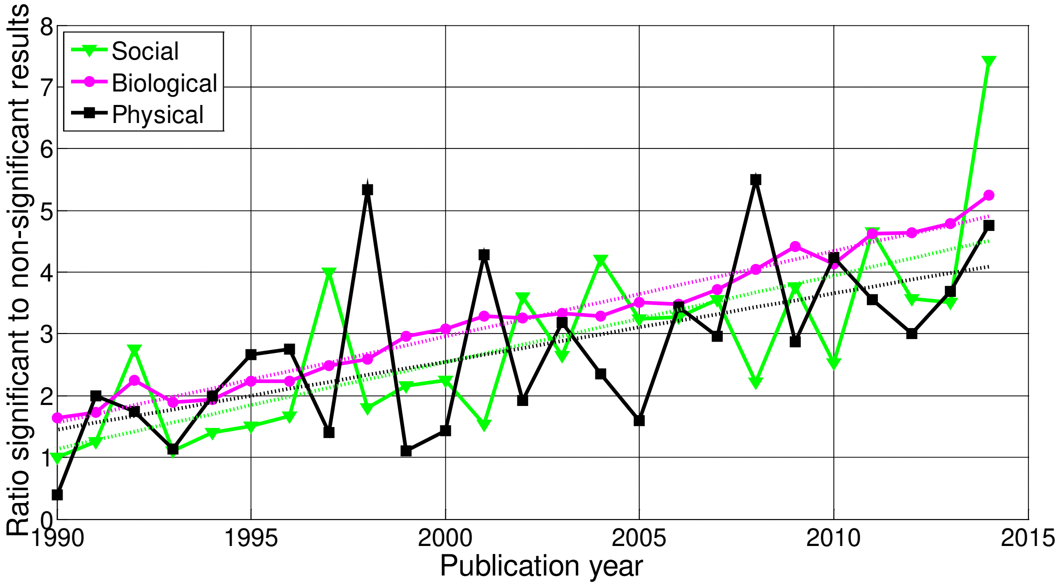

Figure 6. Ratio of significant to non-significant results (*p*-value between 0.040 and 0.049 / *p*-value between 0.051 and 0.060) per publication year, for three scientific disciplines. Dashed lines represent the results of a linear regression analysis.

3.3.2. Textual reporting of statistical (non-)significance

Figure 7 shows that the biological sciences are more likely to use the phrases "significant difference" or "no significant difference" than the social sciences. These phrases are rare in the physical sciences, confirming the results for *p*-values in Fig. 5. The growth rate of the ratio of significant to non-significant results is higher in the social sciences than in the biological sciences, the growth rate of which in turn is higher than that of the physical sciences (Fig. 8 & Table 1). Physical sciences exhibit, however, the highest overall ratio (Fig. 8). Note that the textual analyses are again less sensitive than the *p*-value searches, showing smaller longitudinal trends and smaller differences between disciplines.

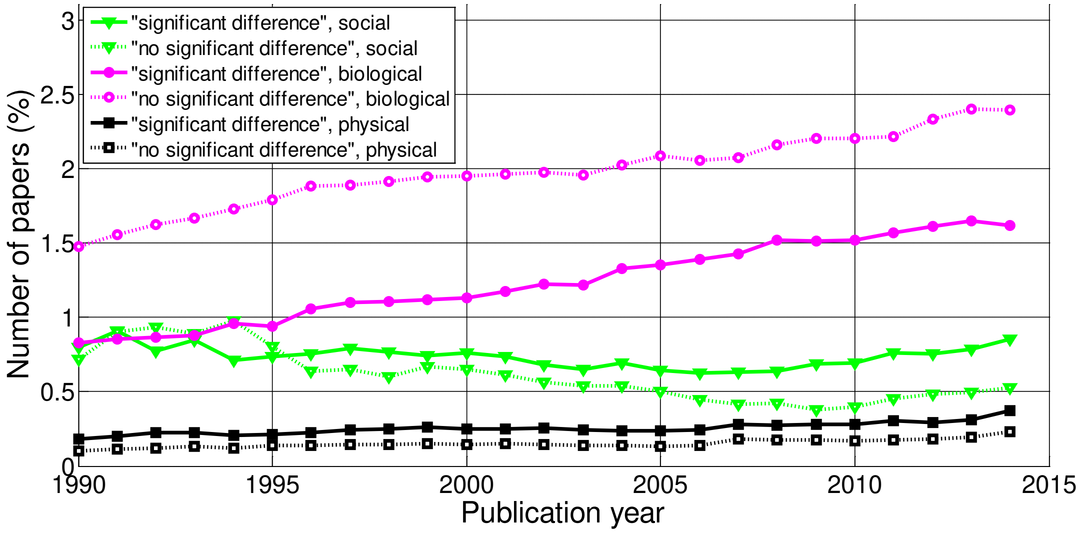

Figure 7. Number of papers containing textual reporting of significance (solid lines) or non-significance (dotted lines) divided by the total number of papers per publication year, for three scientific disciplines.

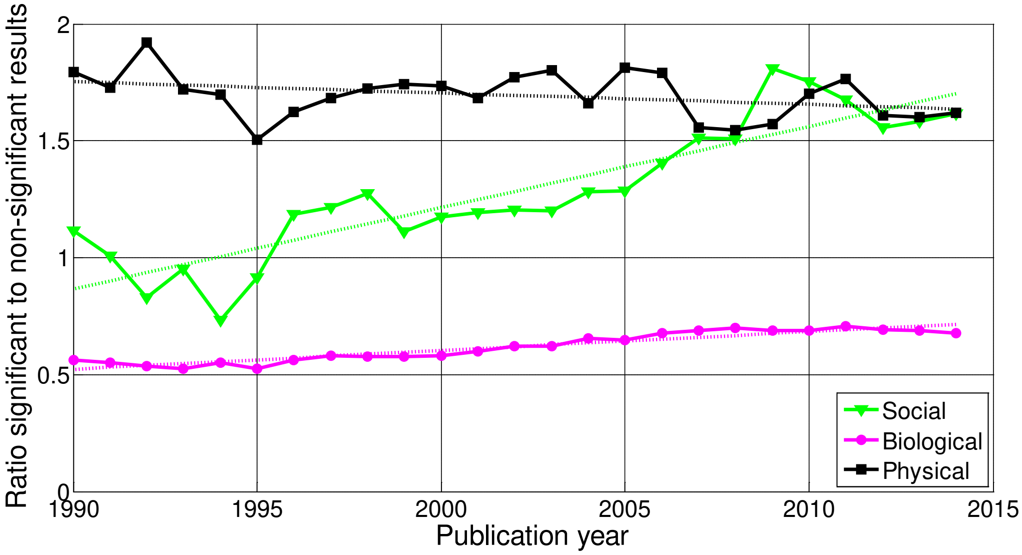

Figure 8. Ratio of significant to non-significant results (textual reporting) per publication year, for three scientific disciplines. Dashed lines represent the results of a linear regression analysis.

Table 1. Slope coefficients (95% confidence interval between brackets) calculated using a simple linear regression, for the ratios of significant (S) to non-significant (NS) results (S/NS) and the percentages of significant results over the sum of significant and non-significant results (100%*S/[S+SN]). Coefficients are reported for all papers, and for papers in three scientific disciplines.

		Total	**Social**	**Biological**	**Physical**
***p*-value**	**S/NS**	0.139 [0.129, 0.149]	0.140 [0.082, 0.198]	0.139 [0.129, 0.150]	0.110 [0.047, 0.174]
	100%*S/(S+SN)	0.844 [0.756, 0.933]	1.075 [0.702, 1.448]	0.845 [0.753, 0.936]	1.083 [0.507, 1.660]
Textual	**S/NS**	0.010 [0.008, 0.011]	0.035 [0.027, 0.042]	0.008 [0.007, 0.009]	−0.005 [−0.010, 0.000]
	100%*S/(S+SN)	0.338 [0.292, 0.383]	0.681 [0.519, 0.843]	0.308 [0.261, 0.355]	−0.067 [−0.142, 0.008]

3.4. Comparison of longitudinal trends between world regions

3.4.1. p-value reporting

For all three world regions, reporting of both significant and non-significant *p*-values has increased over time (see Fig. S3 in supplementary material). The growth of reporting significant results as compared to the reporting of non-significant results is higher in AS7 than in the U.S. and EU15 (Fig. 9; Table 2).

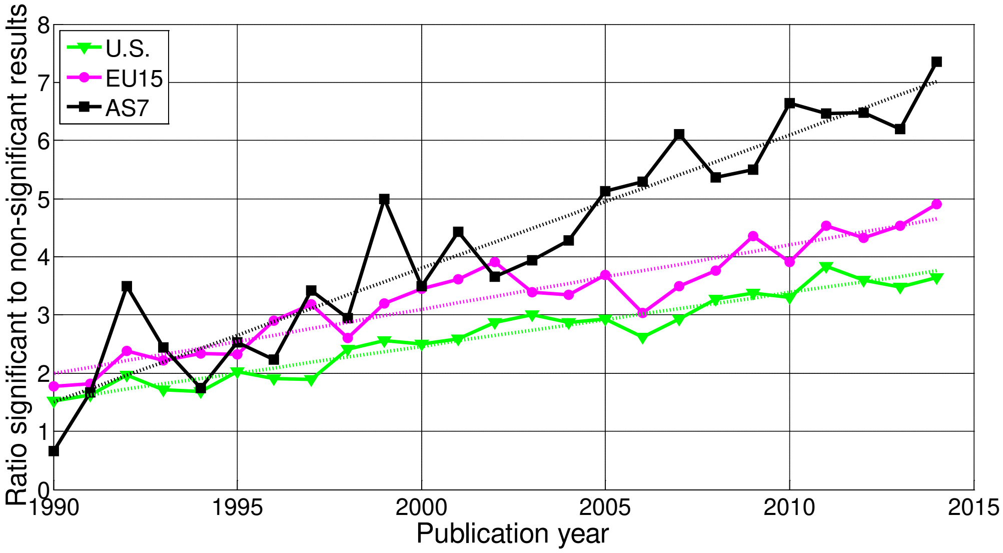

Figure 9. Ratio of significant to non-significant results (*p*-value between 0.040 and 0.049 / *p*-value between 0.051 and 0.060) per publication year, for three world regions. Dashed lines represent the results of a linear regression analysis.

3.4.2. Textual reporting of statistical (non-)significance

The above increases of *p*-value reporting over time are confirmed by the results on textual reporting of statistical significance versus non-significance (see Fig. S4 in supplementary material). The increase of significant results has been faster for EU15 as compared to the U.S. and AS7 (Table 2). Moreover, while AS7 exhibits the highest ratio of significant to non-significant results in terms of *p*-value reporting (Fig. 9), this region also shows the lowest ratio of significant to non-significant results in terms of textual reporting (Fig. 10).

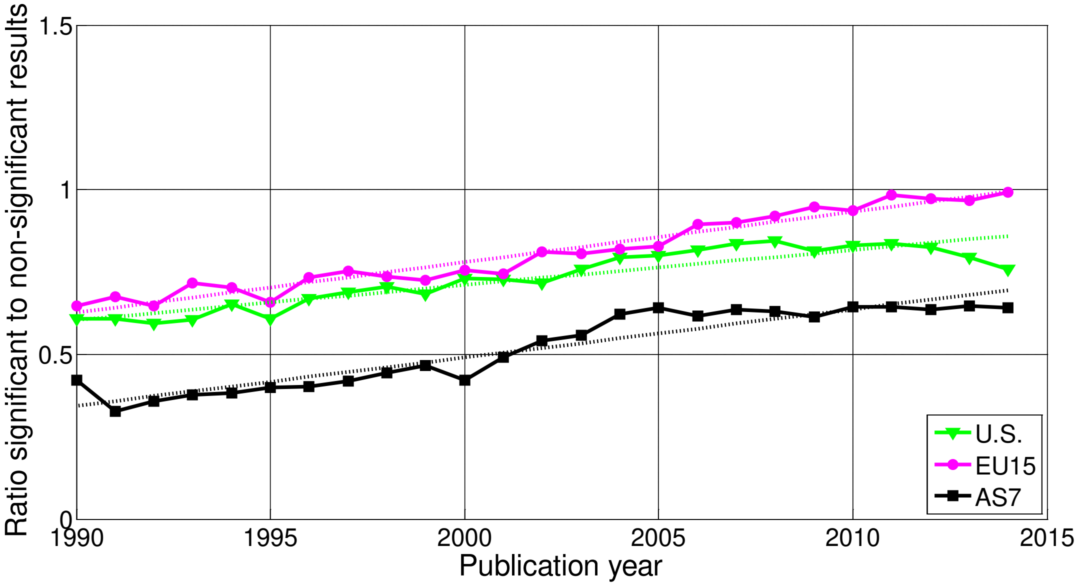

Figure 10. Ratio of significant to non-significant results (textual reporting) per publication year, for three world regions. Dashed lines represent the results of a linear regression analysis.

Table 2. Slope coefficients (95% confidence interval between brackets) calculated using a simple linear regression, for the ratios of significant (S) to non-significant (NS) results (S/NS) and the percentages of significant results over the sum of significant and non-significant results (100%*S/[S+SN]). Coefficients are reported for papers in three world regions.

		U.S.	**EU15**	**AS7**
***p*-value**	**S/NS**	0.092 [0.082, 0.103]	0.111 [0.093, 0.129]	0.230 [0.194, 0.266]
	100%*S/(S+SN)	0.749 [0.644, 0.855]	0.658 [0.530, 0.786]	1.169 [0.798, 1.539]
Textual	**S/NS**	0.011 [0.009, 0.013]	0.015 [0.014, 0.017]	0.015 [0.012, 0.017]
	100%*S/(S+SN)	0.361 [0.262, 0.431]	0.466 [0.420, 0.511]	0.647 [0.542, 0.751]

4. Discussion

We investigated longitudinal trends of positive versus negative results reported in abstracts and compared these trends between disciplines and between world regions. Our analysis showed that the percentage of papers reporting *p*-values between 0.051 and 0.060 has risen with a factor of 4.3 between 1990 and 2014, which indicates that negative results are *not* disappearing. Equally striking is the 13.9-fold increase of *p*-values between 0.040 and 0.049 over the same time period. The large increase of marginally significant *p*-values (as compared to marginally non-significant *p*-values) is consistent with the crisis that certain disciplines are currently experiencing, and suggests that methodologists' recommendations are not heard by the scientific community. Our results for textual reporting of significant differences displays a more modest increase than the results for *p*-value reporting and resembles Pautasso's (2010) findings.

We found no support for the widely discussed idea of a 'hierarchy of sciences'. In our analysis, the differences of the significant to non-significant ratios between the three scientific disciplines are inconsistent. For example, the social sciences show the fastest increase of this ratio, but the physical sciences have the highest ratio overall (Fig. 8). A more salient finding of our analysis is that there is enormous difference in reporting practices between disciplines, with the reporting of *p*-values being 60.6 times more frequent in the biological sciences than in the physical sciences. This effect is even larger when considering that many physics papers that report *p*-values have medical affinity (e.g., a new radiology method being tested in a medical setting, new ultrasound techniques). Indeed, 63% (847 out of

1,344) physical sciences papers retrieved from the search of marginal significant and non-significant *p*-values had "Medicine" appointed by Scopus as an additional subject area. Summarizing, a hierarchy of sciences (if something like that exists) is not characterised by excess significance bias, but rather by differences in null hypothesis significance testing in the first place: the physical sciences almost never use *p*-values.

Researchers in the physical sciences (including engineers) are often able to predict precise numerical values and may not need *p*-values because they encounter low levels of sampling and measurement error (e.g., Hand, 2004; Meehl, 1967). However, physics research is sometimes characterised by excessive noise too. Examples are faint signals picked up in interstellar space (Gurnett et al., 2013), detection of exoplanets (Bean et al., 2010; Robertson et al., 2014), frame dragging experiments (Everitt et al., 2011), or searching for the Higgs Boson (The CMS Collaboration, 2014). Biological and social scientists are trained in experimental methodology and statistics from early college years. Physicists seem to be lagging behind in some aspects of experimental design and statistics, and have introduced methods such as experimenter blinding only recently (Klein & Roodman, 2005). The "look-elsewhere effect" in high-energy physics (Gross & Vitells, 2010) is the equivalent of the multiple comparison problem, a term often used in psychology. Although social scientists have expressed "a desire to imitate physics" (Roberts & Pashler, 2000), such physics envy may be misplaced because (discovery-oriented, innovative) physics encounters the same methodological challenges as those developed by and for the softer scientists.

It is worth noting that the ratio of significant to non-significant differences in our analyses was greater than 1 for *p*-value reporting (Figs. 4, 6 & 9), but mostly smaller than 1 for the textual reporting of significance (Figs. 8 & 10, in line with Pautasso, 2010). So, the absolute number of *p*-values above versus below the alpha (= 0.05) threshold is not a valid measure of bias, and some figures circulating in the social media (e.g., Hankins, 2014, suggesting there is an "immortal hand or eye" pushing *p*-values below 0.05) provide an oversimplified picture. We suspect that researchers often state that a result is "not significant" instead of reporting the specific non-significant *p*-value (e.g., "*p* = 0.055").

We found that the percentage of papers reporting a positive result as well as the percentage of papers reporting a negative result have increased since the 1990s. In a supplementary analysis, we assessed the frequency of various statistically significant *p*-values (0.012, 0.022, 0.032, and 0.042) as a function of publication year (see Fig. S5 in supplementary material). We found that the smaller the significant *p*-value the more frequently it is reported; a rapid increase for all *p*-values can also be seen, suggesting that null hypothesis significance testing has become more widely used over the years (despite widespread criticism against the use of *p*-values, see e.g., Wagenmakers, 2007).

Researchers from the Asian region report *p*-values between 0.040 and 0.049 at a disproportionally high level (Fig. 9), but they are considerably *more* likely to use the phrase "no significant difference" than the other two world regions (Fig. 10). Some studies have found that Asian research is more biased than research elsewhere in the word (Pan et al., 2005; Vickers et al., 1998). Others have argued that it is the U.S. that overestimates effect sizes, especially in softer research (see Fanelli & Ioannidis, 2013, but see Nuijten et al., 2014). Our analysis suggests that it is impossible to 'blame' certain countries for displaying an excess of positive results, since the ratio of significant to non-significant results totally depends on which type of keywords one searches for (i.e., *p*-values vs. textual search). Regional differences may be further moderated by the type of research dominating a particular world region and the fraction of research being indexed in Scopus. For example, papers in the social sciences represent 12% of all papers from the U.S., 8% of all papers from the EU15, and only 5% of all papers from the AS7 region (data retrieved from Scopus). Furthermore, regional differences in excess significance bias are probably obscured by important moderators such as the emergence of China as second publishing power after the U.S. during the last decade (Leydesdorff &Wagner, 2009). Summarizing, differences between world regions in the reporting of significance are too small and inconsistent to draw conclusions on cross-cultural differences in significance bias.

Our automated string-search approach has some important limitations. First, our method may be susceptible to faulty inclusions. For example, the string "p = 0.048" could sometimes appear in a paper that does not test a null hypothesis, but tests something else (e.g., normality). One strength of Fanelli's (2012) work was that he manually examined the abstracts and/or full texts of the selected 4,656 papers testing a hypothesis. So, although Fanelli's sample may not be more representative than ours (as his sample was also automatically generated by searching for the sentence "test* the hypothes*"), his manual checking of the content of each paper may have prevented faulty inclusions, albeit at the cost of objectivity.

Second, Scopus is known to be incomplete for publications prior to 1996 (Elsevier, 2014). This discontinuity can be observed in Fig. 2 showing the number of papers per publication year, but does not seem to have affected the percentages of papers reporting a positive or negative result (i.e., no discontinuity appears in Figs. 3 & 5). We focused the search only on papers with an abstract available in Scopus, to avoid artefacts due to an increased unavailability of abstracts for older records (indeed, we observed that the percentage of Scopus records without an abstract dropped from 45.4% [254,968 out of 561,194 papers] in 1990 to 12.1% [140,986 out of 1,161,405] in 2014).

Third, reporting practices may have changed over time regardless of actual *p*-values. For example, when performing various control checks we found that certain phrases ("the aim of" and "results showed that") have drastically increased over time, whereas other phrases (e.g., "on the other hand", "the properties of", or "room temperature") have remained about constant (see Fig. 11). Our speculation is that the increasing trends in the former category reflect an increase in empiricism and structured reporting in abstracts. Following the same line of reasoning, reporting *p*-values in the abstract may have become increasingly preferred (or even recommended in journal instructions for authors) over a narrative summary of research findings.

Fourth, we did not assess all papers which tested a hypothesis, but only a fraction of these. The papers we assessed represent less than 1% of all papers indexed in Scopus (Figs. 3 & 5). Among these, the results strongly depend on whether one searches for *p*-values or for a textual phrase. In a supplementary analysis, we found that the results also look very different when searching for "p < 0.05" versus "p > 0.05" (see Figs. S6 & S7 in supplementary material). According to this analysis, both the positive ($p < 0.05$) and negative ($p > 0.05$) results have increased, but the increase for negative results has been relatively steeper. This again confirms our main result that negative results are *not* disappearing. We also searched for "$p < 0.05$" versus "$p > 0.05$" for the three scientific disciplines (see Figs. S8 & S9 in supplementary material). According to this analysis, the ratio of significant-over-non-significant results is highest for the social sciences (opposite to our findings for "significance difference(s)" vs. "non-significant difference(s)" in the results section), but it decreases more rapidly for the social sciences than for the physical sciences (see Table S1 in supplementary material), opposite to Fanelli's (2012) observations. The results of a "$p < 0.05$" versus "$p > 0.05$" comparison for the three world regions also differ from both the *p*-value reporting and the textual reporting (see Figs. S10 & S11 as well as Table S2 in supplementary material), with a *decrease* of positive ($p < 0.05$) results for U.S. and EU15 over time and a decreasing ratio of significant to non-significant results for all three world regions.

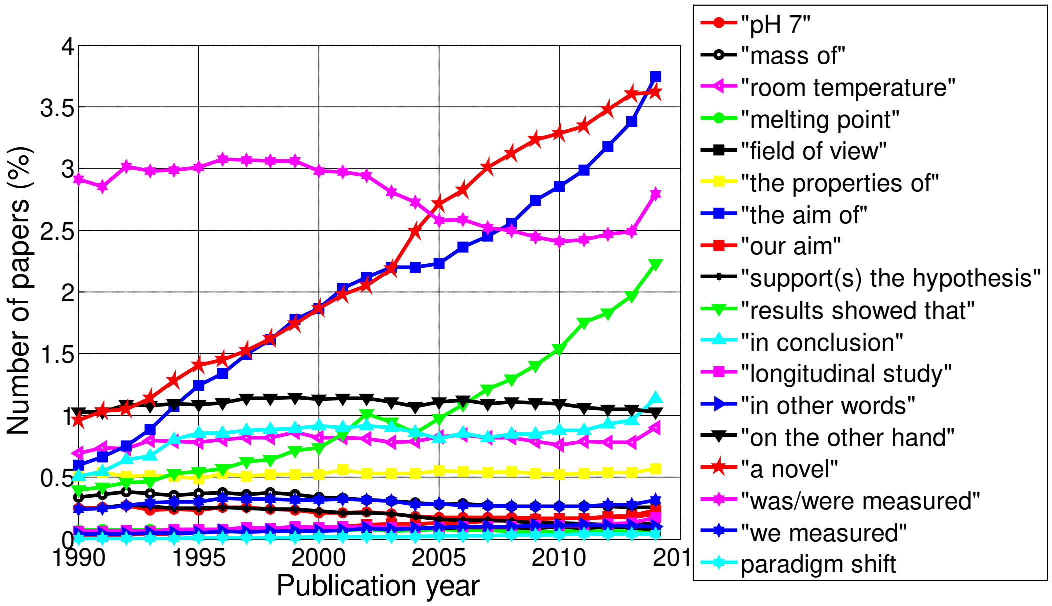

Figure 11. Number of papers reporting certain idiosyncratic expressions typical for technical papers divided by the total number of papers per publication year (the "paradigm shift" expression was used to compare with Atkin, 2002, who reported an exponential increase of this expression in the titles of papers; indeed, the expression was used 36 times more often in 2014 than in 1990).

As mentioned in the introduction, positive results may be caused by questionable research practices such as selective publication, selective analysis, as well as data fabrication. Selective analysis and reporting is probably common in many research fields (Ioannidis, 2010). Some researchers might not be even aware that "data peeking", removing some outliers, or trying out various statistical tests (e.g., parametric and non-parametric ones) and subsequently reporting only the most significant result contributes to the false positive problem (e.g., Bakker & Wicherts, 2014; Strube, 2006). Simulation studies by Simmons et al. (2011) illustrate how flexible analyses can easily result in statistically significant evidence for a false hypothesis. Data fabrication is probably relatively rare, but certainly very harmful. A meta-analytic review by Fanelli (2009) found that 2% of researchers admitted fabricating data, and 14% knew a colleague who fabricated. Fabrication is obviously condemnable and should be prevented by all possible means.

The growth of *p*-values between 0.040 and 0.049 does not imply that questionable practices have become more prevalent. More positive explanations for the observed trends must be considered as well. First, it can be expected that scientists have become more knowledgeable, and therefore better able to formulate accurate predictions and design powerful experiments that disprove a null hypothesis. Second, as mentioned above, scientists have become more likely to use significance testing. The observed longitudinal trends in the reporting of *p*-values might occur because of science becoming more empirical, organized, and quantitative, in an attempt to escape the structuralism and postmodernism of the 1970s and 1980s. Third, false positive results are not necessarily bad, as long as there is a rapid self-correcting mechanism in place (De Winter & Happee, 2013). The observed increase of negative results (Fig. 3) may be caused by a growing replication movement providing a counterforce to positive results, and the increasing use of *p*-values might be a manifestation of dynamic exchange of information within the scientific network. Fourth, as indicated by Popper (1959), theories are "*in an asymmetrical way, falsifiable only*: they are statements which are tested by being submitted to systematic attempts to falsify them". In this line, a single falsification can be worth more than numerous confirmations of an established null hypothesis. So, perhaps some asymmetry between positives and negatives results in innovative research fields is expectable or even desirable.

Supplementary material

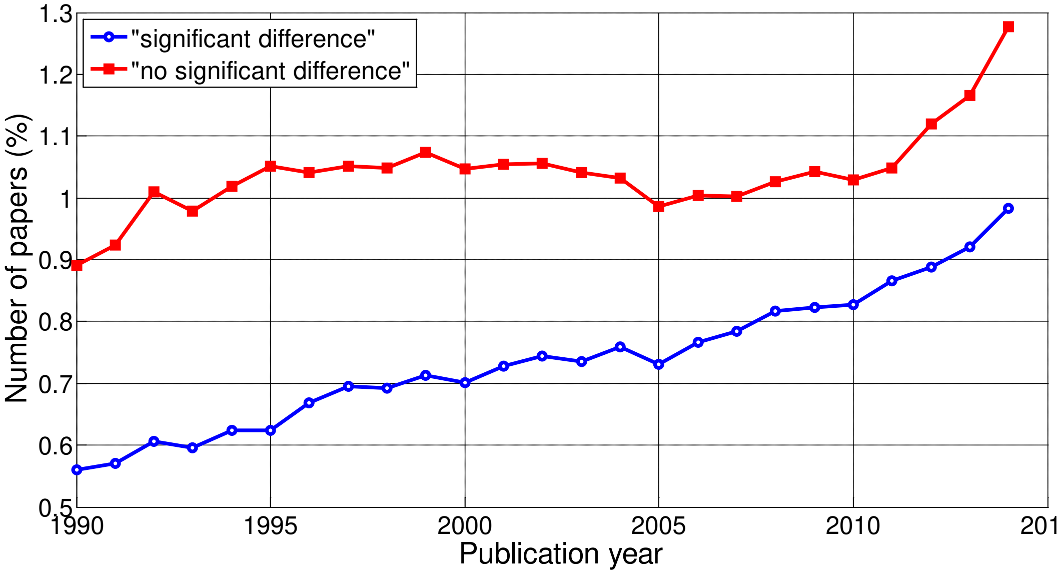

Figure S1. Number of papers containing textual reporting of significance (blue line) or non-significance (red line) divided by the total number of papers per publication year.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.447v3 | CC-BY 4.0 Open Access | received: 2 Aug 2014, published: 2 Aug

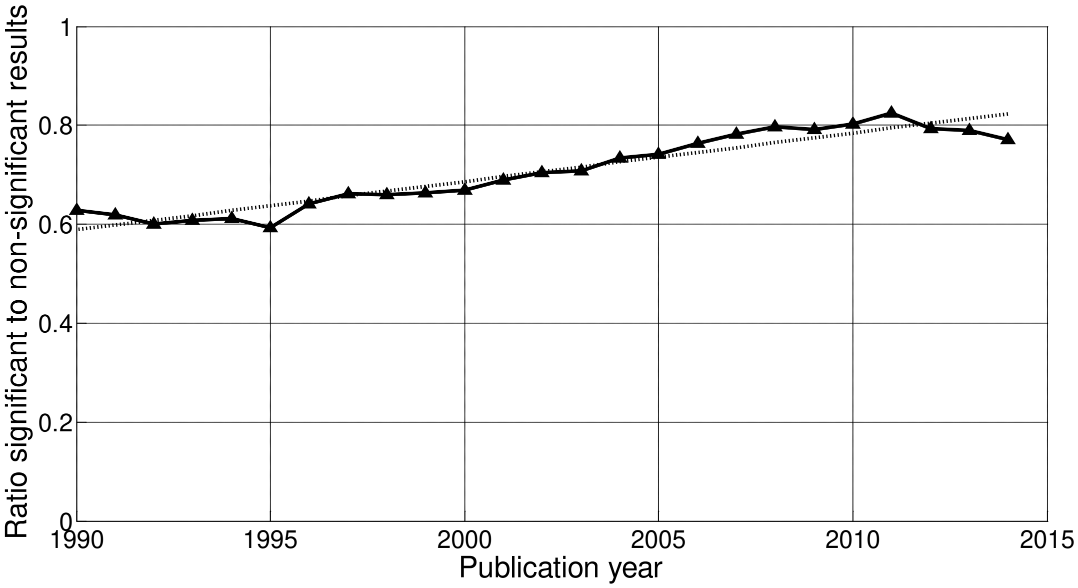

Figure S2. Ratio of significant to non-significant results (textual reporting) per publication year. The dashed line represents the result of a linear regression analysis.

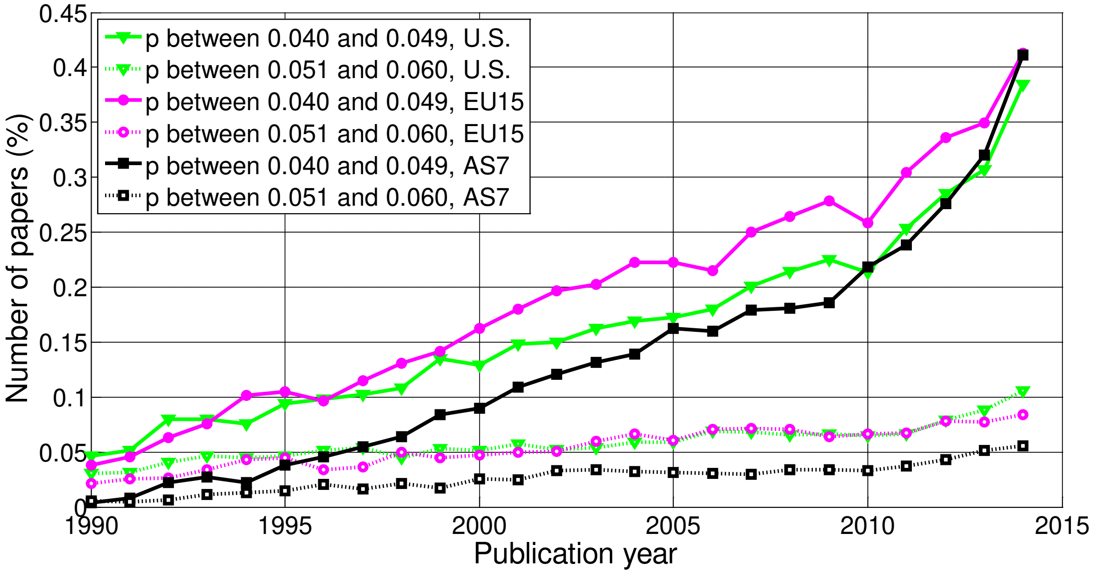

Figure S3. Number of papers reporting a *p*-value between 0.040 and 0.049 (solid lines) or a *p*-value between 0.051 and 0.060 (dotted lines) divided by the total number of papers per publication year, for three world regions.

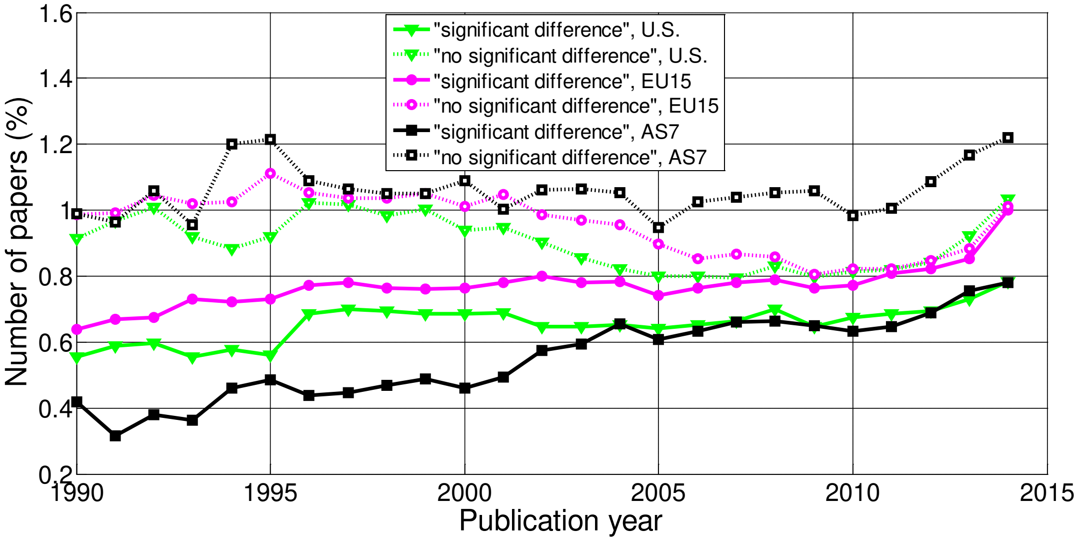

Figure S4. Number of papers containing textual reporting of significance (solid lines) or non-significance (dotted lines) divided by the total number of papers per publication year, for three world regions.

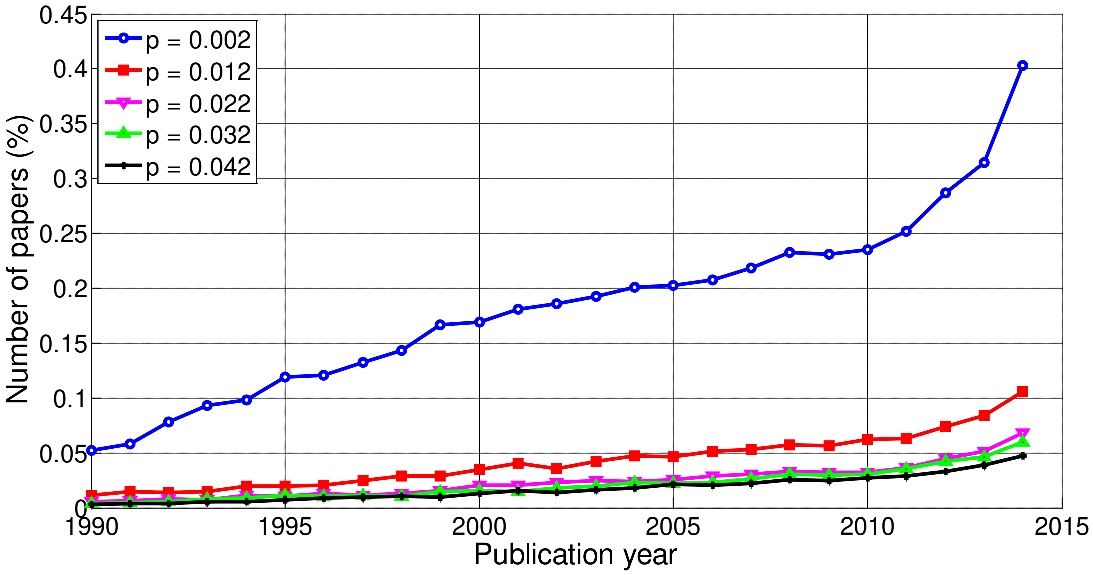

Figure S5. Number of papers reporting a *p*-value between 0.002 and 0.042 by an increment of 0.010 divided by the total number of papers per publication year.

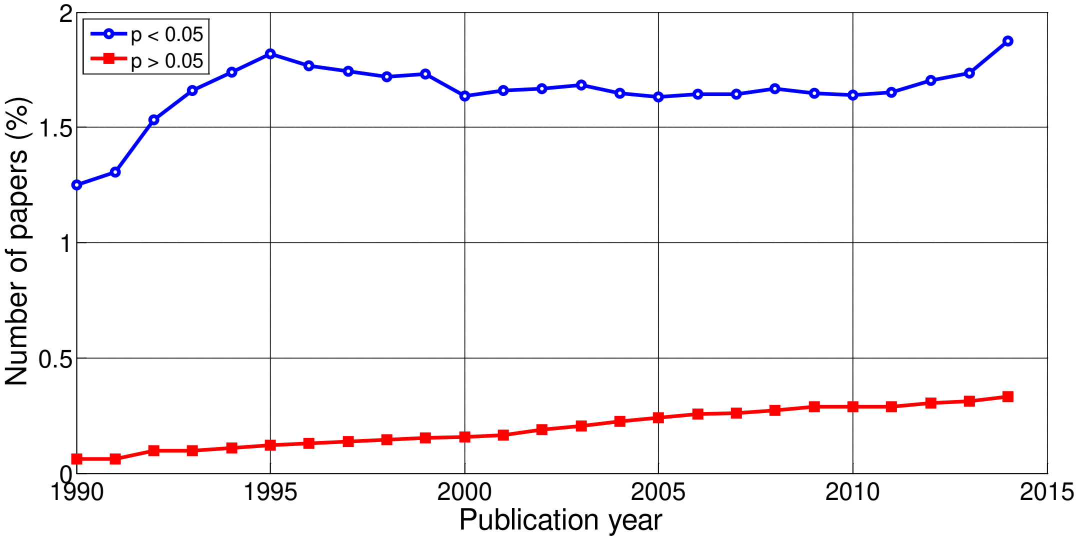

Figure S6. Number of papers reporting "p < 0.05" (or "p < .05") (blue line) or "p > 0.05" (or "p > .05") (red line) divided by the total number of papers per publication year.

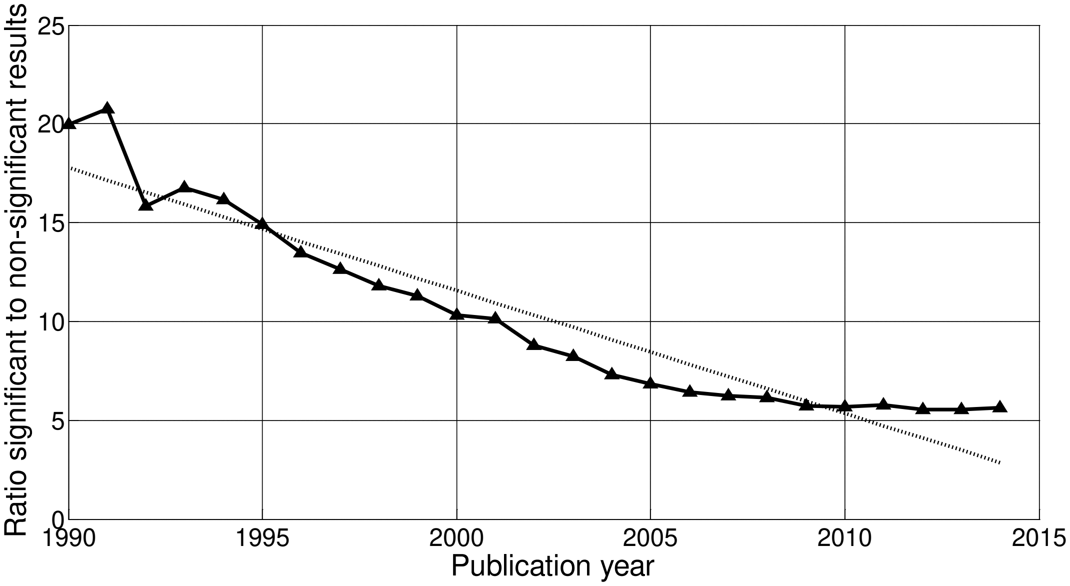

Figure S7. Ratio of significant to non-significant results ($p < 0.05$ / $p > 0.05$) per publication year. The dashed line represents the result of a linear regression analysis.

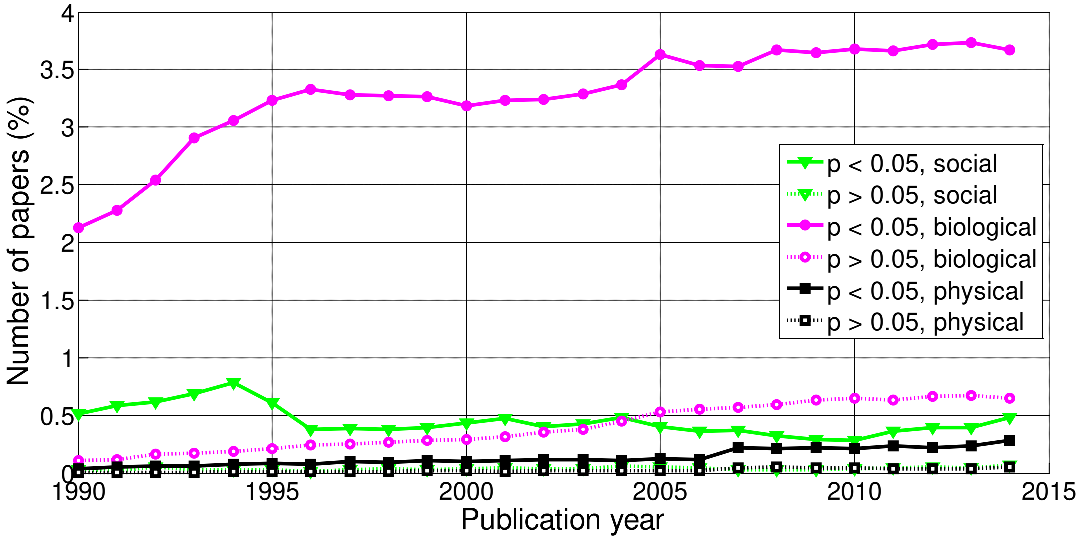

Figure S8. Number of papers reporting "p < 0.05" (or "p < .05") (solid lines) or "p > 0.05" (or "p > .05") (dotted lines) divided by the total number of papers per publication year, for three scientific disciplines.

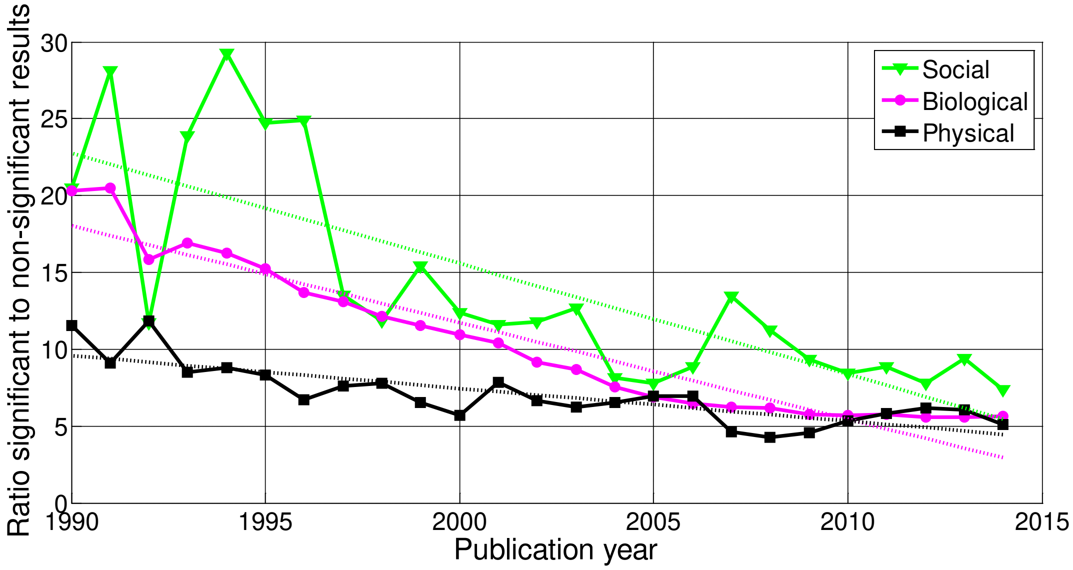

Figure S9. Ratio of significant to non-significant results ($p < 0.05$ / $p > 0.05$) per publication year, for three scientific disciplines. Dashed lines represent the results of a linear regression analysis.

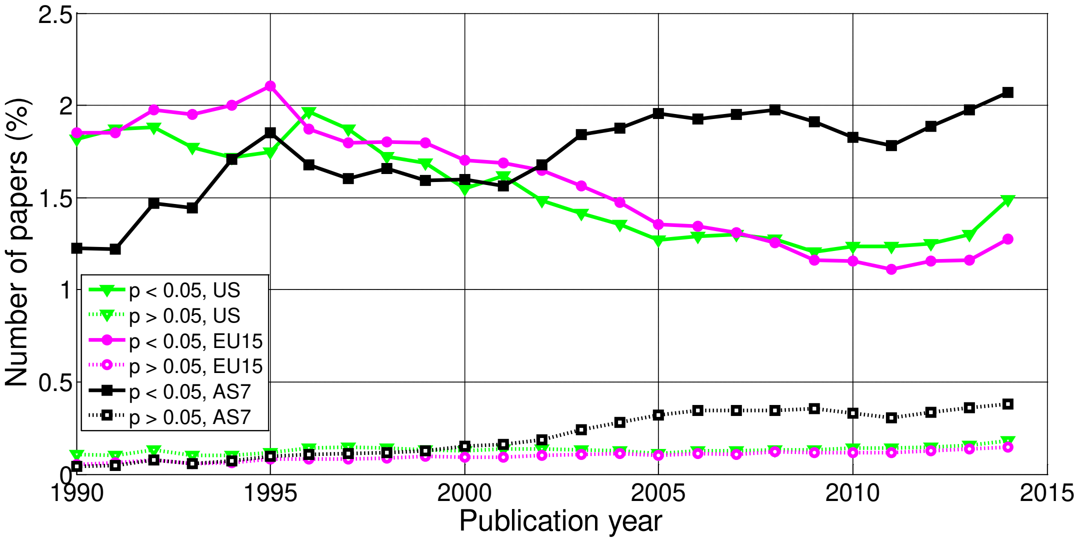

Figure S10. Number of papers reporting "p < 0.05" (or "p < .05") (solid lines) or "p > 0.05" (or "p > .05") (dotted lines) divided by the total number of papers per publication year, for three world regions.

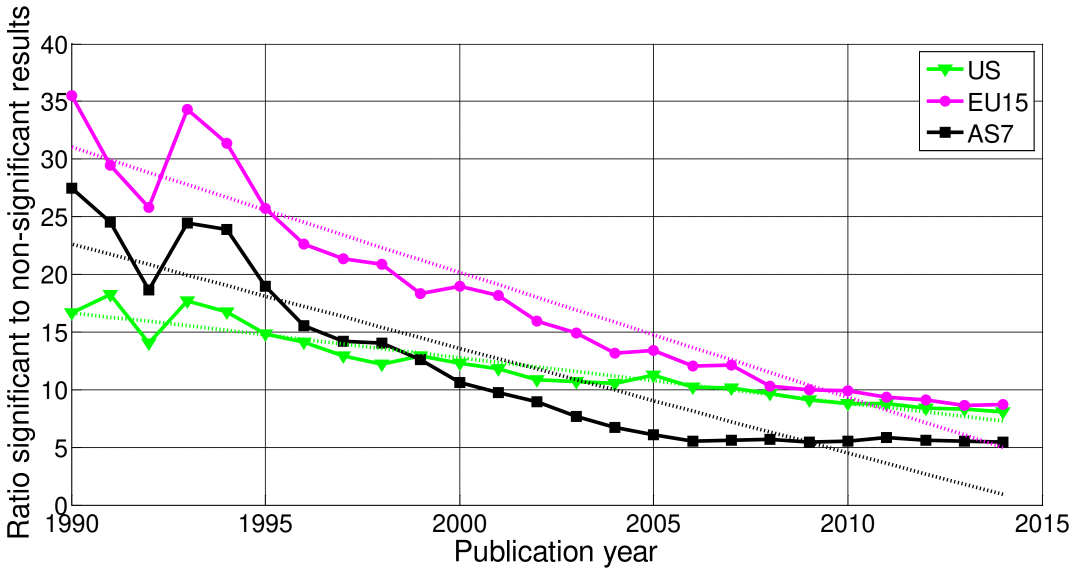

Figure S11. Ratio of significant to non-significant results ($p < 0.05$ / $p > 0.05$) per publication year, for three world regions. Dashed lines represent the results of a linear regression analysis.

Table S1. Slope coefficients (95% confidence interval between brackets) calculated using a simple linear regression, for the ratios of significant (S) to non-significant (NS) results (S/NS) and the percentages of significant results over the sum of significant and non-significant results (100%*S/[S+SN]). Coefficients are reported for all papers, and for papers in three scientific disciplines.

		Total	**Social**	**Biological**	**Physical**
***p* < 0.05 vs.**	**S/NS**	−0.621 [−0.709, −0.534]	−0.721 [−0.974, −0.468]	−0.629 [−0.711, −0.547]	−0.213 [−0.278, −0.148]
***p* > 0.05**	**100%*S/(S+SN)**	−0.515 [−0.553, −0.476]	−0.309 [−0.396, −0.222]	−0.517 [−0.558, −0.476]	−0.316 [−0.417, −0.214]

Table S2. Slope coefficients (95% confidence interval between brackets) calculated using a simple linear regression, for the ratios of significant to non-significant results (S/NS) and the percentages of significant results over the sum of significant and non-significant results (100%*S/[S+NS]). Coefficients are reported for papers in three world regions.

		U.S.	**EU15**	**AS7**
***p* < 0.05 vs.**	**S/NS**	−0.393 [−0.449, −0.338]	−1.085 [−1.237, −0.932]	−0.907 [−1.079, −0.735]
***p* > 0.05**	**100%*S/(S+SN)**	−0.237 [−0.257, −0.217]	−0.342 [−0.367, −0.317]	−0.610 [−0.683, −0.537]

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.447v3 | CC-BY 4.0 Open Access | received: 2 Aug 2014, published: 2 Aug

References

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J., Fiedler, K., ... & Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality, 27*, 108–119.

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

Bakker, M., & Wicherts, J. M. (2014). Outlier removal, sum scores, and the inflation of the type I error rate in independent samples t tests: The power of alternatives and recommendations. *Psychological Methods*.

Bean, J. L., Seifahrt, A., Hartman, H., Nilsson, H., Reiners, A., Dreizler, S., ... & Wiedemann, G. (2010). The proposed giant planet orbiting VB 10 does not exist. *The Astrophysical Journal Letters*, *711*, L19.

Chan, A. W., Song, F., Vickers, A., Jefferson, T., Dickersin, K., Gøtzsche, P. C., ... & Van der Worp, H. B. (2014). Increasing value and reducing waste: addressing inaccessible research. *The Lancet, 383*, 257–266.

Colom, F., & Vieta, E. (2011). The need for publishing the silent evidence from negative trials. *Acta Psychiatrica Scandinavica, 123*, 91–94.

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

De Angelis, C., Drazen, J. M., Frizelle, F. A., Haug, C., Hoey, J., Horton, R., ... & Weyden, M. B. V. D. (2004). Clinical trial registration: a statement from the International Committee of Medical Journal Editors. *New England Journal of Medicine, 351*, 1250–1251.

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

De Winter, J., & Happee, R. (2013). Why selective publication of statistically significant results can be effective. *PloS One, 8*, e66463.

Dirnagl, U., & Lauritzen, M. (2010). Fighting publication bias: introducing the Negative Results section. *Journal of Cerebral Blood Flow and Metabolism, 30,* 1263–1264.

Djulbegovic, B., Lacevic, M., Cantor, A., Fields, K. K., Bennett, C. L., Adams, J. R., ... & Lyman, G. H. (2000). The uncertainty principle and industry-sponsored research. *The Lancet, 356*, 635–638.

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

Dwan, K., Altman, D. G., Arnaiz, J. A., Bloom, J., Chan, A. W., Cronin, E., ... & Williamson, P. R. (2008). Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PloS One, 3*, e3081.

Elsevier (2014). Printing and exporting citation overviews. Retrieved from http://help.scopus.com/Content/h_citovrdoc.htm

Everitt, C. W. F., DeBra, D. B., Parkinson, B. W., Turneaure, J. P., Conklin, J. W., Heifetz, M. I., ... & Wang, S. (2011). Gravity Probe B: Final results of a space experiment to test general relativity. *Physical Review Letters, 106*, 221101.

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

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *PloS One, 5*, e10068.

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics, 90*, 891–904.

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science, 7*, 555–561.

Fanelli, D., & Ioannidis, J. P. (2013). US studies may overestimate effect sizes in softer research. *Proceedings of the National Academy of Sciences, 110*, 15031–15036.

Fern, E. F., & Monroe, K. B. (1996). Effect-size estimates: Issues and problems in interpretation. *Journal of Consumer Research, 23*, 89–105.

Fiedler, K., Kutzner, F., & Krueger, J. I. (2012). The long way from α-error control to validity proper. Problems with a short-sighted false-positive debate. *Perspectives on Psychological Science, 7*, 661–669.

Francis, G. (2013). Replication, statistical consistency, and publication bias. *Journal of Mathematical Psychology, 57*, 153–169.

Gadbury, G. L., & Allison, D. B. (2012). Inappropriate fiddling with statistical analyses to obtain a desirable p-value: tests to detect its presence in published literature. *PloS One, 7*, e46363.

Gelman, A., & Loken, E. (2013). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time.* Unpublished manuscript.

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

Gross, E., & Vitells, O. (2010). Trial factors for the look elsewhere effect in high energy physics. *The European Physical Journal C - Particles and Fields, 70*, 525–530.

Gurnett, D. A., Kurth, W. S., Burlaga, L. F., & Ness, N. F. (2013). In situ observations of interstellar plasma with Voyager 1. *Science, 341*, 1489–1492.

Hand, D. J. (2004). *Measurement: theory and practice.* London: Arnold.

Hankins, M. [mc_hankins] (2014, May 17). What immortal hand or eye, Could frame thy fearful asymmetry? Reported p values in the range 0.041 to 0.059 (Scholar) pic.twitter.com/bPhlljb0B4 [Tweet]. Retrieved from https://twitter.com/mc_hankins/status/467766548162412545

Hopewell, S., Loudon, K., Clarke, M. J., Oxman, A. D., & Dickersin, K. (2009). Publication bias in clinical trials due to statistical significance or direction of trial results. *Cochrane Database of Systematic Reviews, 1*.

Hunter, J. E., & Schmidt, F. L. (1996). Cumulative research knowledge and social policy formulation: The critical role of meta-analysis. *Psychology, Public Policy, and Law, 2*, 324–347.

Ioannidis, J. (2003). Genetic associations: false or true? *Trends in Molecular Medicine, 9*, 135–138.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine, 2*, e124.

Ioannidis, J. (2008). Calibration of credibility of agnostic genome-wide associations. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 147*, 964–972.

Ioannidis, J. (2010). Meta-research: The art of getting it wrong. *Research Synthesis Methods, 1*, 169–184.

Ioannidis, J. P. (2011). Excess significance bias in the literature on brain volume abnormalities. *Archives of General Psychiatry, 68*, 773–780.

Ioannidis, J., & Doucouliagos, C. (2013). What's to know about the credibility of empirical economics? *Journal of Economic Surveys, 27*, 997–1004.

Ioannidis, J., Munafò, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. *Trends in Cognitive Sciences, 18*, 235–241.

Jennings, R. G., & Van Horn, J. D. (2012). Publication bias in neuroimaging research: implications for meta-analyses. *Neuroinformatics, 10*, 67–80.

Jennions, M. D., & Møller, A. P. (2002). Publication bias in ecology and evolution: an empirical assessment using the 'trim and fill' method. *Biological Reviews of the Cambridge Philosophical Society, 77*, 211–222.

Joober, R., Schmitz, N., Annable, L., & Boksa, P. (2012). Publication bias: What are the challenges and can they be overcome? *Journal of Psychiatry & Neuroscience, 37*, 149–152.

Kahneman, D. (2013). A proposal to deal with questions about priming effects. Retrieved from http://www.nature.com/polopoly_fs/7.6716.1349271308!/suppinfoFile/Kahneman%20Letter.pdf

Kirkham, J. J., Dwan, K. M., Altman, D. G., Gamble, C., Dodd, S., Smyth, R., & Williamson, P. R. (2010). The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. *British Medical Journal, 340*, c365.

Klein, J. R., & Roodman, A. (2005). Blind analysis in nuclear and particle physics. *Annual Reviews of Nuclear and Particle Science, 55*, 141–163.

Kyzas, P. A., Denaxa-Kyza, D., & Ioannidis, J. (2007). Almost all articles on cancer prognostic markers report statistically significant results. *European Journal of Cancer*, *43*, 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*, 575–603.

Laws, K. R. (2013). Negativland-a home for all findings in psychology. *BMC Psychology, 1*, 2.

Lexchin, J., Bero, L. A., Djulbegovic, B., & Clark, O. (2003). Pharmaceutical industry sponsorship and research outcome and quality: systematic review. *British Medical Journal, 326*, 1167–1170.

Leydesdorff, L., & Wagner, C. (2009). Is the United States losing ground in science? A global perspective on the world science system. *Scientometrics, 78*, 23–36.

Masicampo, E. J., & Lalande, D. R. (2012). A peculiar prevalence of *p* values just below. 05. *The Quarterly Journal of Experimental Psychology, 65*, 2271–2279.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34*, 103–115.

Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., ... & Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science, 331*, 176–182.

Moore, R. A., Derry, S., & McQuay, H. J. (2010). Fraud or flawed: adverse impact of fabricated or poor quality research. *Anaesthesia, 65*, 327–330.

Nuijten, M. B., van Assen, M. A., van Aert, R. C., & Wicherts, J. M. (2014). Standard analyses fail to show that US studies overestimate effect sizes in softer research. *Proceedings of the National Academy of Sciences, 111*, E712–E713.

Pan, Z., Trikalinos, T. A., Kavvoura, F. K., Lau, J., & Ioannidis, J. P. (2005). Local literature bias in genetic epidemiology: an empirical evaluation of the Chinese literature. *PLoS Medicine, 2*, e334.

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science, 7*, 531–536.

Pautasso, M. (2010). Worsening file-drawer problem in the abstracts of natural, medical and social science databases. *Scientometrics*, *85*, 193–202.

Popper, K. R. (1959). *The logic of scientific discovery.* London: Hutchinson.

Ridley, J., Kolm, N., Freckelton, R. P., & Gage, M. J. G. (2007). An unexpected influence of widely used significance thresholds on the distribution of reported *P*-values. *Journal of Evolutionary Biology, 20*, 1082–1089.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review, 107*, 358–367.

Robertson, P., Mahadevan, S., Endl, M., & Roy, A. (2014). Stellar activity masquerading as planets in the habitable zone of the M dwarf Gliese 581. *Science*, 1253253.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.). (2006). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. West Sussex, England: John Wiley & Sons.

Rouder, J. N., Morey, R. D., Verhagen, J., Province, J. M., Wagenmakers, E. J., & Rouder, J. (2014). *The p<. 05 rule and the hidden costs of the free lunch in inference.* Unpublished manuscript. Retrieved from: http://pcl.missouri.edu/sites/default/files/p_5.pdf

Rücker, G., Schwarzer, G., & Carpenter, J. (2008). Arcsine test for publication bias in meta-analyses with binary outcomes. *Statistics in Medicine, 27*, 746–763.

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1*, 199–233.

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

Sismondo, S. (2008). Pharmaceutical company funding and its consequences: a qualitative systematic review. *Contemporary Clinical Trials, 29*, 109–113.

Smart, R. G. (1964). The importance of negative results in psychological research. *Canadian Psychologist/Psychologie Canadienne, 5*, 225–232.

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. *The American Statistician*, *49*, 108–112.

Strube, M. J. (2006). SNOOP: A program for demonstrating the consequences of premature and repeated null hypothesis testing. *Behavior Research Methods, 38*, 24–27.

Terrin, N., Schmid, C. H., Lau, J., & Olkin, I. (2003). Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine, 22*, 2113–2126.

The CMS Collaboration (2014). Evidence for the direct decay of the 125 GeV Higgs boson to fermions. *Nature Physics*.

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

Vickers, A., Goyal, N., Harland, R., & Rees, R. (1998). Do certain countries produce only positive results? A systematic review of controlled trials. *Controlled Clinical Trials, 19*, 159–166.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review, 14*, 779–804.

Yong, E. (2012). Replication studies: Bad copy. *Nature, 485*, 298–300.

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