**Supplementary Information**

**Study 1**

**Career-stage differences (or, similarities).**

**Is the field more replicable now than it was 10 years ago?**Fifty percent of participants answered “yes” when asked if the field was more replicable now than it was 10 years ago. Perceptions of improved replicability over time, however, varied as a function of career stage, χ2(4, *N* = 920) = 24.10, *p* < .001, *Cramer’s V = .16*. Graduate students and associate professors were most likely to believe that replicability improved over time (58% and 53%, respectively), followed by non-tenure track PhDs and assistant professors (each at 48%). Full professors were markedly the least likely to believe replicability improved over time (35%).

**Confidence.**On average, participants’ confidence that the majority of findings in social psychology will replicate was *M* = 2.04, *SD =* 0.83, which registered just over *slightly confident.* Confidence that the majority of findings will replicate was moderated by career stage, *F*(4, 925) = 2.78, *p* = .03, ω2 =.01. Tukey tests indicated that non-tenure track PhDs were the least confident (*M* = 1.96, *SD* = 0.77) and full professors the most confident (*M* = 2.17, *SD* = 0.93) that the majority of findings would replicate. Graduate students, assistant professors, and associate professors fell between these two groups and were statistically indistinguishable.

**Attitudes toward the status of our science conversation.** On average, participants’ perception that the status of our science discussion has been a good or bad thing was *M* = 4.34 (*SD* = 1.50) or roughly the *neither good nor bad* response option. This perception was not qualified by stage of career, *F*(4, 924) = 2.23, *p* = .06, ω2 < .01.

**Has the discussion improved research?** The average perception that the status of our science discussion had improved research in the field was *M* = 2.78 (*SD* = 0.96), or slightly less than *moderately*. The perception that the discussion has led to some improvement of research was not qualified by stage of career, *F*(4, 925) = 1.66, *p* = .12, ω2 < .01.

**Has the discussion changed the way you do research?** Participants reported a moderate change in the way that they do research as a result of the status of our science discussion, *M* = 2.83, *SD* = 1.12, a result that was qualified by stage of career, *F*(4, 923) = 8.53, *p* < .001, ω2 < .01. Tukey tests indicated that full professors were least likely (*M* = 2.59, *SD* = 1.17) and graduate students were most likely (*M* = 3.08, *SD* = 1.03) to report that they changed the way they do research given the status of our science discussion. Non-tenure track PhDs, assistant professors, and associate professors fell between these two groups and were statistically indistinguishable.

 …

**Intentions to change*.*** The practices that participants reported as most acceptable (i.e., not making data publicly available, not conducting power analyses, and not pre-registering hypotheses) were interestingly also the practices participants’ reported they were most likely to change in response to the status of our science discussion. As can be seen in Figure 4, 50% or more of participants reported that their intentions to make data publicly available, conduct power analyses, and to report effect sizes have increased as a result of the status of our science discussion. Many participants also reported that their intentions to engage in various QRPs have decreased as a function of the status of our science conversation (see Figure for more detail). As can be in seen in the figure below, graduate students were the most likely to report that their intentions to use QRPs had decreased as a function of the ongoing conversation about best practices. In contrast, full professors were the least likely to report an intention to increase their use of “best” practices, relative to researchers at other career stages.

**Study 2**

**Design Information**

Table S1 summarizes the types of designs reported over all types of studies conducted (correlational, experimental, and quasi-experimental). In Table S2, we removed correlational studies (10.10% and 15.17% for 2003-2004 and 2013-2014, respectively) most of the studies which we could not code the design type was removed. In summary, most of the studies (about 70%) in our sample were between-subject designs.

Table S1.

*Design type of all studies analyzed in manuscript.*

|  |  |  |
| --- | --- | --- |
|  | **2003-2004** | **2013-2014** |
|  | Frequency | **%** | Frequency | **%** |
| Missing | 110 | 21.32% | 203 | 20.53% |
| Between | 267 | 51.74% | 562 | 56.83% |
| Mixed  | 76 | 14.73% | 115 | 11.63% |
| Within  | 63 | 12.21% | 109 | 11.02% |
| Sum | 516 | 100% | 989 | 100% |

Table S2.

*Design type of all experimental or quasi-*e*xperimental studies analyzed in manuscript.*

|  |  |  |
| --- | --- | --- |
|  | **2003-2004** | **2013-2014** |
|  | Frequency | **%** | Frequency | **%** |
| Missing | 9 | 1.57% | 7 | 0.83% |
| Between | 394 | 68.64% | 613 | 73.06% |
| Mixed  | 121 | 21.08% | 145 | 17.28% |
| Within  | 50 | 8.71% | 74 | 8.82% |
| Sum | 574 | 100% | 839 | 100% |

**P-Curve collapsing across all studies**

We looked at the distribution of calculated *p*-values for all studies reported separately for the original P-curve and ambitious *P*-curve by treating our sample as if it were one paper with 954 studies. Using the original *P*-curve, we found that there were significantly more *p*-values that were smaller (i.e., less than .04, and therefore consistent with evidentiary value) than larger (i.e., between .04 and .05, which estimates the degree of *p-*hacking), in both 2003-2004 (92.84% vs 7.17%, *p* < .0001) and 2013-2014 (91.63 % vs 8.37%, *p* < .0001). Moreover, a two-way χ2 test comparing these distributions over the two time periods revealed that there was no difference between original *P*-curves reported in 2003-2004 and 2013-2014, χ2 (1, *N* = 954) = 0.28, *p* = .6, *Cramer’s V* = .02.

Using the ambitious *P*-curve definition, we also found that there were significantly more *p*-values that were smaller (i.e., less than .025) than larger (i.e., between .025 and .05), in both 2003-2004 (81.31% vs 18.69%, *p* < .0001) and 2013-2014 (78.99% vs 21.01%, *p* < .0001). Moreover, a two-way χ2 test comparing these distributions over the two time periods similarly revealed that there was no difference between *p*-values reported in 2003-2004 and 2013-2014, χ2 (1, *N* = 954) = 0.57, *p* = .45, *Cramer’s V* = .02. Thus, when collapsing across all studies, roughly 80%-95% of studies demonstrate evidentiary value within each time period and no difference between the time periods, according to the original and ambitious *P*-curve types, respectively. *P*-curve analysis collapsing across all studies therefore suggests it is *not so bad* and that it is not *getting better* or *getting worse*.

**Citation Rate Analysis**

In addition to these variants on the p-curve analyses, we also tried to identify visible, influential, and famous papers within our sample by coding how frequently those papers had been cited and then looking within each time period to see how citation (one proxy measure for scientific influence and fame) relates to the replicability metrics. We approached this question in two ways. First we examined the correlation between the citation rate and sample size, effect size, and power. If researchers were citing bad apples papers, those papers would have unusually small sample sizes and/or large effect sizes. Second, we examined the replicability metrics use examined in Table 4 of the paper (P-curve, N-pact, a priori power, observed power, TIVA Z-curve, & R-index) the top 90th percentile of papers in terms of citation rate (relative to each time period) and ran the non-parametric inferential test we used in the paper (i.e., BCA bootstrap) between the middle of the distribution (45-55th percentile) of cited papers. If the well cited papers (90th percentile) are rotten apples they should have smaller samples, higher effect sizes, lower reported power, and lower replicability metrics.

**Correlation approach**. The number of citations was not correlated in other time period with sample size (2003-2004: Kendall’s τ(505) = .035; 2013-2014: Kendall’s τ(978) = -.007) or estimated effect size (2003-2004: Kendall’s τ(341) = .015; 2013-2014: Kendall’s τ(655) = .043), post-hoc power (2003-2004: Kendall’s τ(356) = -.012; 2013-2014: Kendall’s τ(683) = .019). In addition there were no differences between time periods for those correlations for sample size, *z* = 0.76, *p* = .45, effect size, *z* = .41, *p* = .68 or post-hoc power *z* = .47, *p* = .64. These results suggest that highly cited papers are no more likely to be “rotten” than papers not cited as highly.

**Bootstrap approach**. Tables S3 and S4 shows the results of the analysis comparing the 90th and 45-55th percentile of papers in terms of citation rate separately for each time period. Results suggest no differences between the 90th and 45-55th percentile of papers. Again, these results are inconsistent with the idea that a highly cited papers are no more likely to be “rotten” than papers not cited as highly.

Table S3.

*Summary statistics for citation rate groupings for 2013-2014 with 95% BCa CIs.*

|  |  |
| --- | --- |
|  | **2003-2004** |
| **Citation Percentile:** | **45-55th Percentile** | **90+th  Percentile** |
|  | Arithmetic | 95% BCa CI | Arithmetic | 95% BCa CI |
|   | Statistic | LCI | UCI | Statistic | LCI | UCI |
| **Published Methods** |  |  |  |  |  |  |
| Original P-Curve  | 100 | 100 | 100 | 100 | 100 | 100 |
| Ambitious P-Curve  | 55.56 | 20 | 87.5 | 60.00 | 0 | 100 |
| N-Pact (median sample size) | 118 | 95 | 141 | 100.50 | 76 | 118 |
| A Priori Power (% .8 power at *d* = .43) | 39.02 | 23.81 | 53.6 | 34.62 | 17.36 | 55 |
| Observed (Post-hoc) Power (median) | 0.81 | 0.63 | 0.93 | 0.96 | 0.70 | 0.99 |
| **Unpublished Methods** |  |  |  |  |  |  |
| TIVA | 196.56 | 106.44 | 335.95 | 336.92 | 203.76 | 437.53 |
| Z-Curve (median) | 2.57 | 2.25 | 3.04 | 2.87 | 2.21 | 4.30 |
| R-Index  | 0.65 | 0.49 | 0.93 | 0.64 | 0.20 | 0.98 |

Note: \**Bold values indicate where the 95% BCa CIs do not overlap.*

Table S4.

*Summary statistics for citation rate groupings for 2013-2014 with 95% BCa CIs.*

|  |  |
| --- | --- |
|  | **2013-2014** |
| **Citation Percentile:** | **45-55th Percentile** | **90+th  Percentile** |
|  | Arithmetic | 95% BCa CI | Arithmetic | 95% BCa CI |
|   | Statistic | LCI | UCI | Statistic | LCI | UCI |
| **Published Methods** |  |  |  |  |  |  |
| Original P-Curve  | 95 | 83.33 | 100 | 100 | 100 | 100 |
| Ambitious P-Curve  | 30.43 | 14.29 | 52.57 | 66.67 | 20 | 90 |
| N-Pact (median sample size) | 94 | 81 | 104 | 115.5 | 94.5 | 163 |
| A Priori Power (% .8 power at *d* = .43) | 34.31 | 25.96 | 44.24 | 42.37 | 29.64 | 55.56 |
| Observed (Post-hoc) Power (median) | 0.84 | 0.81 | 0.87 | 0.85 | 0.80 | 0.93 |
| **Unpublished Methods** |  |  |  |  |  |  |
| TIVA | 311.07 | 170.87 | 516.11 | 360.58 | 224.12 | 514.76 |
| Z-Curve (median) | 2.75 | 2.50 | 2.88 | 2.79 | 2.45 | 3.28 |
| R-Index  | 0.63 | 0.48 | 0.71 | 0.61 | 0.4 | 0.83 |

Note: \**Bold values indicate where the 95% BCa CIs do not overlap.*

**Estimated effect size.** We plotted the standard effect size (R2) using a KDE density plot that we were able to create given the statistics reported. As depicted in the top panel of Figure S1, both distributions were positively skewed, but in 2003-2004 there was a higher proportion of studies with R2 of .25 and larger. Using the entropy density equality test, we found that the two density plots differed significantly across time periods, S*p* = .0353, *p* < .0001. Mean and median effect size decreased from 2003-2004 to 2013-2014. To help put these values into context, the change in median value R2= .109 in 2003-2004 and the median value R2 = .075 in 2013-2014 is the equivalent to a change in Cohen’s d from .70 to .57, respectively. If this were an independent samples t-test, it would mean that 48% more participants would be needed to maintain the same level of power (.80). These analyses do not directly specify whether research has improved or not, but do highlight the importance of evaluating numerous aspect of studies beyond just sample size or power.

**Figure S1**. Top panel is the KDE density plot for observed effect size (*R*2) from reported test statistics (*t*, *F*, *r*) by time period and the bottom panel is a forest plot with 95% BCa CI for measures of central tendency.

**Post-Hoc Power.** Figure S2 shows the KDE distributions of *post-hoc* power by year (see Hoenig & Heisey, 2001). As with Observed power seen in Figure 8 of the paper, the distribution was highly negatively skewed with peaks at power of 1, suggesting publication bias. The entropy density equality test did not differ significantly across time periods for observed power, *Sp* = .0138, *p* = .155. In other words, the post hoc estimate of power has remained stable from 2003-2004 to 2013-2014. Although post hoc power estimates are extremely upwardly biased and should be interpreted with great caution, the median values were just below Cohen’s .80 threshold for both time periods, a conclusion more consistent with an interpretation of *it’s not so bad* than *it’s rotten to the core*.

**Figure S2**. Top panel is the KDE density plot for post-hoc power from reported test statistics (*t*, *F*, *r*) by time period and the bottom panel is a forest plot with 95% BCa CI for measures of central tendency.

**Journal-by-journal comparison of estimated power.**

Following Fraley and Vazire (2014), we estimated the percentage of studies conducted that should have power of .80, assuming the average effect size in social/personality psychology of *r* = .21 or *d* = .43 and assuming all designs are between-subjects *t*-tests (Richard, Bond, & Stokes-Zoota, 2003). Under these assumptions, studies with sample sizes at or greater then *n* = 172 would be sufficient sample size to claim evidentiary value at *a priori* power = .80. Under these admittedly conservative assumptions, the percentage of studies with a sufficient sample to obtain a power of .80 significantly increased from 15.20% [CI95% = 11.70, 19.00] in 2003-2004 to 24.27% [CI95% = 20.92, 27.18] in 2013-2014.[[1]](#footnote-1) Additionally, these changes are not the same at each of the four journals examined. Instead, *Psychological Science* shows the greatest (and only significant) increase over the time period from 35.19% [CI95% 29.16%, 42.48%] in 2003-2004 to 58.81% [CI95% 49.98, 67.76%] of studies having power of .80 in 2013-2014. While the other journals showed some increase, none were beyond the confidence internals (see Figure below).

**Figure.** The power of each journal to detect an effect of size of d = .430, based on each journals median sample size. The drop line represents the sample median collapsing across journals within each time period.

****

**TIVA conversation to power**

BecauseTIVA basically is measurement of the variance of scores, it means we can covert it over to the effect size, R2 (Explained variance / Explained + Chance variance). We assumed that chance variance in the perfect case would be a uniform distribution of *Z*-scores, which is also the null case in *P*-Curve. We bounded the *Z*-scores of the uniform distribution to the max and min of our data and used it as the chance value in our calculation: R2 = Variance Z-scores / Variance Z-scores + Variance Uniform Distribution of Z-scores). We used this value with lowest bound of median sample size (*n* = 84) in a between-subjects t-test to estimate power. See Line 1028 of our R script for the function.

1. This analysis only includes studies that used χ2, *F*, *t*, and *r* tests. [↑](#footnote-ref-1)