The Rules of the Game Called Psychological Science
Marjan Bakker, Annette van Dijk and Jelte M. Wicherts
*Perspectives on Psychological Science* 2012 7: 543
DOI: 10.1177/1745691612459060

The online version of this article can be found at:
http://pps.sagepub.com/content/7/6/543
The Rules of the Game Called Psychological Science

Marjan Bakker¹, Annette van Dijk¹, and Jelte M. Wicherts²
¹University of Amsterdam, The Netherlands, and ²Tilburg University, The Netherlands

Abstract

If science were a game, a dominant rule would probably be to collect results that are statistically significant. Several reviews of the psychological literature have shown that around 96% of papers involving the use of null hypothesis significance testing report significant outcomes for their main results but that the typical studies are insufficiently powerful for such a track record. We explain this paradox by showing that the use of several small underpowered samples often represents a more efficient research strategy (in terms of finding $p < .05$) than does the use of one larger (more powerful) sample. Publication bias and the most efficient strategy lead to inflated effects and high rates of false positives, especially when researchers also resorted to questionable research practices, such as adding participants after intermediate testing. We provide simulations that highlight the severity of such biases in meta-analyses. We consider 13 meta-analyses covering 281 primary studies in various fields of psychology and find indications of biases and/or an excess of significant results in seven. These results highlight the need for sufficiently powerful replications and changes in journal policies.

Keywords

replication, sample size, power, publication bias, false positives

In many ways, science resembles a game (Mahoney, 1976). It involves rules (not cheating), individual players (researchers), competing teams (paradigms), arbiters (reviewers and editors), and the winning of points (publications) and trophies (professorships, grants, and awards). Just like many games, science also involves the laws of chance. This is so specifically because many results are obtained by null hypothesis significance testing (NHST; Kline, 2004). Notwithstanding the criticism it has received (Cohen, 1990, 1994; Gigerenzer, 2004; Kruschke, 2011; Meehl, 1978; Nickerson, 2000; Rosenthal, 1960; Wagenmakers, 2007; Wetzels et al., 2011), NHST continues to be the main method of statistical inference in many fields. In NHST, the researcher defines a null hypothesis of no effect (H0) and then determines the chance of finding at least the observed effect given that this null hypothesis is true. If this collected chance (or $p$ value) is lower than a predetermined threshold (typically .05), the result is called significant. A significant result will increase the possibility of publishing a result (Mahoney, 1977). If science were a game, winning would entail writing the most interesting publications by gathering many $p$ values below .05.

In this article, we discuss the replication crisis in psychology in terms of the strategic behaviors of researchers in their quest for significant outcomes in NHST. In line with previous work (Ioannidis, 2005, 2008b; Simmons, Nelson, & Simonson, 2011), we present the results of simulations in the context of meta-analysis to highlight the potential biases thus introduced. We assess these problems in 13 psychological meta-analyses and discuss solutions.

Authors Are Lucky!

It has long been argued that the combined outcomes of NHST in the scientific literature are too good to be true (Fanelli, 2010; Fiedler, 2011; Ioannidis, 2008a; Sterling, 1959; Vul, Harris, Winkielman, & Pashler, 2009). Fanelli (2010) documented that over 80% of scientific publications in various sciences report positive results and that the psychological literature shows the highest prevalence of positive outcomes. Sterling (1959) and Sterling, Rosenbaum, and Weinkam (1995) showed that in 97% (in 1958) and 96% (in 1986–1987) of psychological studies involving the use of NHST, H0 was rejected at $\alpha = .05$. Although it should be noted that psychological papers report a host of test results (Maxwell, 2004), the abundance of positive outcomes is striking because effect sizes (ESs) in psychology are typically not large enough to be detected by the relatively small samples used in most studies (i.e., studies are often underpowered; Cohen, 1990).
The power of statistical tests depends on the nominal significance level (typically .05), the sample size, and the underlying ES, such as Cohen’s $d$ for between-group mean comparisons. According to Marszalek, Barber, Kohlhart, and Holmes (2011), the median total sample size in four representative psychological journals (Journal of Abnormal Psychology, Journal of Applied Psychology, Journal of Experimental Psychology: Human Perception and Performance, and Developmental Psychology) was 40. This finding is corroborated by Wetzel et al. (2011), who found a median cell size of 24 in both between- and within-subjects designs in their large sample of $t$ tests from Psychonomic Bulletin & Review and Journal of Experimental Psychology: Learning, Memory and Cognition. The average ES found in meta-analyses in psychology is around $d = 0.50$ (Anderson, Lindsay, & Bushman, 1999; Hall, 1998; Lipsy & Wilson, 1993; Meyer et al., 2001; Richard, Bond, & Stokes-Zoota, 2003; Tett, Meyer, & Roese, 1994), which might be an overestimation of the typical ES given the biases we discuss below. Nevertheless, the typical power in our field will average around 0.35 in a two independent samples comparison, if we assume an ES of $d = 0.50$ and a total sample size of 40. This low power in common psychological research raises the possibility of a file drawer (Rosenthal, 1979) containing studies with negative or inconclusive results. Publication bias can have dire consequences, as illustrated recently by clear failures to replicate medical findings (Begley & Ellis, 2012; Prinz, Schlange & Asadullah, 2011). On the basis of surveys of researchers and a study of the fate of studies approved by the institutional review board of a major U.S. university, the percentage of unpublished studies in psychology is estimated to be at least 50% (Cooper, DeNeve, & Charlton, 1997; Coursol & Wagner, 1986; Shadish, Doherty, & Montgomery, 1989), but the problem goes beyond widespread failure to publish. Statistical textbooks advise the use of formal a priori power estimates, but in a recent sample of psychological papers with NHST (Bakker & Wicherts, 2011), only 11% referred to power as a rationale for the choice of sample size or design. Although power estimates can also be done informally, the typical study in psychology appears to be underpowered. If a study’s power equals 0.50, the chance to find a significant result equals that of correctly predicting “heads” in a coin flip. The number of “heads” presented in the psychological literature (and in other literatures) suggests a problem. Although one author has explained this by claiming that researchers are psychic (Bones, 2012), we think that they just act strategically.

A Dozen Replications

The common lack of power is well illustrated by studies of the (positive) association between infants’ habituation to a given stimulus and their later cognitive ability (IQ). One often-cited meta-analysis (McCall & Carriger, 1993) collated 12 studies of the correlation between measures of habituation during children’s first year of life and IQ as measured between 1 and 8 years of age. In the funnel plot (Light & Pillemer, 1984) of Figure 1, these 12 Fisher-transformed (normalized) correlations are plotted against the inverse of the standard error (SE)
in each study. The \( SE \) depends on sample size and equals \( \sqrt{1/N-3} \). The white area represents the 95% confidence interval (or area) under \( H_0 = 0 \), and study outcomes that fall in the grey area are significant at \( \alpha = .05 \) (two-tailed). The straight dotted line represents the estimated underlying ES from a fixed effect meta-analysis (\( Z = .41 \), which corresponds to \( r = .39 \) and \( d = .85 \)) and the curved dotted lines represent the 95% confidence interval around that estimate of the alternative hypothesis (\( H_A \)); study outcomes invariably fell in this interval and so they appear to be homogeneous; \( Q = 6.74, DF = 11, p = .820 \). The upward narrowing of both 95% confidence intervals in the funnel reflects larger power to reject \( H_0 \) in large samples. On the right hand side, we depicted the power to reject \( H_0 \) given the sample size (\( SE \) level) as implied by the estimated underlying ES of .41. For instance, on the top of Figure 1, we depicted the distributions under \( H_0 \) and \( H_A \) for \( N = 50 \) (\( 1/SE = 6.86 \)), which corresponds to a power of .807. As we go down the funnel, \( SEs \) become larger and so outcomes should deviate more strongly from the estimate of \( H_A \). It is noteworthy that (a) all but three of the studies have a power below .80, (b) the correlation differed significantly from zero in all but one study, and (c) the study outcomes are clearly not evenly distributed in the right- and left-hand side of the funnel associated with \( H_A \). In fact, the two largest studies showed the weakest link between infant cognition and later IQ, whereas the smaller studies all lie on the right-hand side of the funnel (i.e., in the grey area where \( p < .05 \)). Such funnel plot asymmetry is awkward and can be tested (Sterne & Egger, 2005) by regressing outcomes on sample sizes (or \( SEs \)) across the 12 studies: \( Z = 2.24, p = .025 \). The median sample size of these studies was 25 and their typical (median) power equaled .488. Under the assumption that studies are independent, the expected number of significant findings on the basis of this power analysis (i.e., the sum of power values) is 6.71, and so a positive outcome in 11 out of such 12 underpowered studies is unlikely. Ioannidis and Trikalinos (2007) proposed a straightforward \( \chi^2 \) test for such an excess of significant findings: \( \chi^2 (DF = 1) = 6.21, p = .013 \). So this meta-analysis shows the typical signs of publication bias and results that are too good to be true (Francis, 2012b). One possible explanation is that studies with nonsignificant or lower correlations were missing from the meta-analysis. In addition, research with infants is not easy and seldom are these analyses carved in stone. Statistical choices concerning exclusion of data points, outliers, and operationalization of dependent variables require choices that are often arbitrary and so provide researchers degrees of freedom (Simmons et al., 2011) in their analyses.

**Playing the Game Strategically**

The excess of significant findings may partly be explained by researchers’ exploitation of these degrees of freedom in their pursuit of significant outcomes (Fiedler, 2011; Wicherts, Bakker, & Molenaar, 2011) and by the fact that it is easier to find a significant effect in multiple small studies rather than one larger study. The use of multiple small studies rather than a larger one gives the researcher the opportunity to make small alterations to the research design and provides ample opportunity for capitalizing on chance.

Simmons et al. (2011) illustrated how easy it is to inflate Type I error rates when researchers employ hidden degrees of freedom in their analyses and design of studies (e.g., selecting the most desirable outcomes, letting the sample size depend on results of significance tests). John, Loewenstein, and Prelec (2012) surveyed over 2,000 psychological researchers and found a majority of them to admit to use at least some of these questionable research practices (QRPs). For instance, the majority admitted to having ever failed to report all of the dependent measures in a study. Forty-eight percent admitted to having only report studies that “worked” (which we take to imply \( p < .05 \)), whereas 57% acknowledged to having used sequential testing (cf. Wagenmakers, 2007) in their work. Such practices lead to inflated ESs and increased false positive rates (Ioannidis, 2005, 2008b; Simmons et al., 2011).

Suppose psychology were a game in which players have to gather a significant result in a particular direction. Players have resources to gather data from \( N \) participants and can choose between these options:

- **Strategy 1.** Perform one large study (with \( N \) as the sample size) with sufficient power and publish it.
- **Strategy 2.** Perform one large study and use some of the QRPs most popular in psychology (John et al., 2012). These QRPs may be performed sequentially until a significant result is found:
  - a. Test a second dependent variable that is correlated with the primary dependent variable (for which John et al. found a 65% admittance rate)
  - b. Add 10 subjects (sequential testing; 57% admittance rate)
  - c. Remove outliers (\(|Z > 2|\)) and rerun analysis (41% admittance rate)
- **Strategy 3.** Perform, at most, five small studies each with \( (N/5) \) as sample size. Players may stop data collection when they find a significant result in the expected direction and only publish the desired result (the other studies are denoted “failed”; 48% admittance rate).
- **Strategy 4.** Perform, at most, five small studies and apply the QRPs described above in each of these small studies if the need arises. Players may report only the first study that “worked.”

Strategies 3 and 4 imply publication bias in the traditional sense, whereas Strategies 2 and 4 relate to the analysis of the data. So what is the winning strategy? We simulated data (see the online Appendix at http://pps.sagepub.com-supplemental
for details) on the basis of sample sizes and ESs that are typical for psychology and found a clear answer. The left panel of Figure 2 gives the proportion of researchers who gather at least one significant finding ($p < .05$) under these four strategies. Note that we simulated one-sided results (i.e., directional hypotheses) but employed two-sided tests, which should be

Fig. 2. Results of simulations showing the optimal strategy for players (in terms of probability of finding at least one significant result; left column) and the bias in estimated ESs (right column) under the four strategies described in the text under a range of genuine ESs ($d = 0$ to $1$). $N$ represents sample sizes for small studies, whereas the larger sample size equals $5^*N$ within each row. Results are one-sided (i.e., directional hypotheses), but tests are two-sided and should be significant in the expected direction.
significant in the expected direction. So the (combined) Type I error rate (i.e., when ES = 0) is .025 for the large study and 1 − (1 − .025)×5 = .119 for the five small studies.

The upshot of the simulation results is that when the true ES is small executing multiple small and underpowered studies represents the optimal strategy for individual players to generate a \( p \) value of less than .05. Furthermore, the use of QRPs pays off, especially with small samples and ESs. Besides, many players need not even perform all five studies—for example, with a cell size of 20 and an ES of 0.5 they need on average 2.58 studies (expected total \( N = 103 \)) without QRPs and 1.14 studies (expected total \( N = 66 \)) with QRPs. Combined with the selection of significant results through publication decisions, these strategies may explain why so many psychological researchers continue to run underpowered studies yet almost always report significant results.

Get Serious!

Science is not a game. The optimal strategies described above lead to inflated Type I errors of up to 40% and inflate genuine effects. We calculated the bias in our simulation as the difference between the average estimated ES and the true ES. Results are presented in the right panel of Figure 2. With large samples (Strategy 1), there is no systematic bias. For large samples with QRPs (Strategy 2), the bias goes to zero with a larger true ES, which is to be expected because of the larger power under this scenario (QRPs are not required for “winning”). The bias is large with multiple small studies (Strategies 3 and 4). In the typical psychological study (cell size 20 and true ES = .5), the biases with and without QRPs are 0.327 and 0.158, respectively. With smaller cell sizes and an ES of .5, the bias can be as large as 0.913. When multiple small studies are combined with QRPs (Strategy 4), the bias is large for small true ES, but decreases with larger true ES, possibly because of the adding of subjects (in Step b). Even those who ignore \( p \) values of individual studies will find inflated ESs in the psychological literature if a sufficient number of researchers play strategically, which indeed many psychological researchers appear to do (John et al., 2012).

Our field lacks clear codes of conduct considering the use of these analytic strategies (Sterba, 2006) and many reviewers tend not to accept \( p \) values above .05 (Mahoney, 1977), possibly because this presents an easy heuristic (Hoekstra, Finch, Kiers, & Johnson, 2006; Nelson, Rosenthal, & Rosnow, 1986). Sole studies are seldom definitive but even knowledgeable researchers tend to underestimate the randomness associated with small samples (Tversky & Kahneman, 1971). At the end of the day it is all about replication.

Another 250+ Replications

A set of reasonably similar replications can be used to determine robustness of findings and to study signs of the use of the strategies described above. In our simulations, we used fixed effects and applied Sterne and Egger’s (2005) test of funnel plot asymmetry and Ioannidis and Trikalinos’ (2007) test for an excess of significant findings. Both methods are described above and functioned well in the simulation (see the online Appendix at http://pps.sagepub.com/supplemental and Fig. 4), although it is important to note that they are sensitive to actual heterogeneity of the underlying ESs (Ioannidis & Trikalinos, 2007; Sterne & Egger, 2005). For instance, funnel plot asymmetry may arise if smaller studies tap on stronger underlying effects because they are done in relatively more controlled settings. Therefore, these tests are best applied to relatively homogenous sets of studies as defined in the realm of meta-analysis.

To gather a representative sample of sets of psychological studies that concern the same phenomenon or at least highly similar phenomena, we retrieved from PsycARTICLES all 108 peer-reviewed articles published in 2011 that contained the strings “research synthesis,” “systematic review,” or “meta-anal*” in title and/or abstract. Subsequently, we randomly selected 11 useful meta-analyses (10% of the total). We only included meta-analyses that reported the ESs and standard errors (or sample sizes) of primary studies. From each meta-analysis, we retrieved the subset (as selected by the authors of the meta-analyses) of at least 10 primary studies that was the most homogenous subset in terms of Higgins’ \( I^2 \). We assumed that the meta-analysts employed rigorous inclusion and exclusion criteria and that they correctly determined ESs, and we feel confident that the primary studies in each of the fields are sufficiently comparable to be considered replications.

The selected (subsets from) meta-analyses are given in Table 1 together with tests for homogeneity, excess of significant findings, and funnel plot asymmetry. The average impact factor of the journals in which the meta-analyses appeared was 4 (see the online Appendix at http://pps.sagepub.com/supplemental for full references). Meta-analyses were from clinical, counseling, educational, evolutionary, developmental, family, and industrial/organizational psychology. The medians of the sample sizes align with those found in the wider literature, although the median ES \( (d = .37) \) was slightly lower than \( d = .50 \), as described earlier.

Figure 3 depicts the funnel plots of the 11 meta-analyses. Tests for funnel plot asymmetry (with an \( \alpha = .10 \) as suggested by various authors; Ioannidis & Trikalinos, 2007) were significant in four instances (36%). In three instances, we found signs of an excess of significant results (27% at \( \alpha = .10 \)). These results replicate earlier indicators of the prevalence of funnel plot asymmetry in 99 psychological meta-analyses (Ferguson & Brannick, 2012; Levine, Asada, & Carpenter, 2009) and the finding of an excess of significant results in four areas of psychological research (Francis, 2012a, 2012b, in press).

To get a feel for the likelihood of biases in the actual meta-analyses, we simulated results for 16 meta-analyses with 100 studies each. These meta-analyses are presented in Figure 4 (see the online Appendix at http://pps.sagepub.com/
<table>
<thead>
<tr>
<th>Reference</th>
<th>Subgroup</th>
<th>ES (d)</th>
<th>(SE)</th>
<th>k</th>
<th>N</th>
<th>Med. N</th>
<th>Q (p)</th>
<th>I²</th>
<th>Exp.</th>
<th>Obs.</th>
<th>χ² (p)</th>
<th>Regtest (p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alfieri et al. (2011)</td>
<td>Enhanced discovery:</td>
<td>0.20</td>
<td>(.042)</td>
<td>24</td>
<td>2350</td>
<td>61</td>
<td>39.35 (.018)</td>
<td>41.6</td>
<td>3.85</td>
<td>8</td>
<td>5.34 (.021)</td>
<td>0.64 (.519)</td>
</tr>
<tr>
<td></td>
<td>Children</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Benish et al. (2011)</td>
<td>All</td>
<td>0.37</td>
<td>(.067)</td>
<td>21</td>
<td>933</td>
<td>39</td>
<td>16.28 (.699)</td>
<td>0</td>
<td>4.72</td>
<td>5</td>
<td>0.02 (.884)</td>
<td>1.48 (.140)</td>
</tr>
<tr>
<td>Berry et al. (2011)</td>
<td>Self-other</td>
<td>0.37</td>
<td>(.070)</td>
<td>21</td>
<td>3502</td>
<td>146</td>
<td>147.29 (&lt;.001)</td>
<td>86.4</td>
<td>19.75</td>
<td>20</td>
<td>0.05 (.819)</td>
<td>0.77 (.440)</td>
</tr>
<tr>
<td>Card et al. (2011)</td>
<td>Externalizing</td>
<td>0.02</td>
<td>(.019)</td>
<td>11</td>
<td>1295</td>
<td>138</td>
<td>47.59 (&lt;.001)</td>
<td>79</td>
<td>0.66</td>
<td>1</td>
<td>0.19 (.663)</td>
<td>-0.75 (.454)</td>
</tr>
<tr>
<td>Farber &amp; Doolin (2011)</td>
<td>All</td>
<td>0.21</td>
<td>(.032)</td>
<td>18</td>
<td>1027</td>
<td>38</td>
<td>44.81 (&lt;.001)</td>
<td>62.1</td>
<td>5.71</td>
<td>9</td>
<td>2.78 (.095)</td>
<td>2.47 (.014)</td>
</tr>
<tr>
<td>Green &amp; Rosenfeld (2011)</td>
<td>Average SIRS simulators</td>
<td>1.78</td>
<td>(.091)</td>
<td>12</td>
<td>718</td>
<td>53</td>
<td>19.12 (.059)</td>
<td>42.5</td>
<td>11.97</td>
<td>12</td>
<td>0.03 (.862)</td>
<td>2.24 (.025)</td>
</tr>
<tr>
<td></td>
<td>versus nonclinical</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hallion &amp; Ruscio (2011)</td>
<td>Posttest</td>
<td>0.14</td>
<td>(.042)</td>
<td>44</td>
<td>2311</td>
<td>48</td>
<td>38.36 (.672)</td>
<td>0</td>
<td>3.45</td>
<td>2</td>
<td>0.66 (.416)</td>
<td>3.16 (.002)</td>
</tr>
<tr>
<td>Lucassen et al. (2011)</td>
<td>All</td>
<td>0.12</td>
<td>(.028)</td>
<td>16</td>
<td>1355</td>
<td>80</td>
<td>13.11 (.594)</td>
<td>0</td>
<td>3.07</td>
<td>3</td>
<td>0.00 (.965)</td>
<td>0.44 (.658)</td>
</tr>
<tr>
<td>Mol &amp; Bus (2011)</td>
<td>Grades 1-12 basics</td>
<td>0.26</td>
<td>(.030)</td>
<td>18</td>
<td>1164</td>
<td>51</td>
<td>20.46 (.251)</td>
<td>16.9</td>
<td>8.56</td>
<td>10</td>
<td>0.46 (.498)</td>
<td>0.70 (.487)</td>
</tr>
<tr>
<td>Woodin (2011)</td>
<td>Satis.-Hosti.</td>
<td>-0.63</td>
<td>(.037)</td>
<td>40</td>
<td>3008</td>
<td>51</td>
<td>41.48 (.363)</td>
<td>6.0</td>
<td>25.27</td>
<td>26</td>
<td>0.06 (.811)</td>
<td>-2.01 (.045)</td>
</tr>
<tr>
<td>Woodley (2011)</td>
<td>All</td>
<td>0.02</td>
<td>(.022)</td>
<td>12</td>
<td>2056</td>
<td>193</td>
<td>37.26 (&lt;.001)</td>
<td>70.5</td>
<td>0.72</td>
<td>3</td>
<td>7.66 (.006)</td>
<td>-0.62 (.532)</td>
</tr>
<tr>
<td>Greenwald, Poel-hman, Uhman, &amp;</td>
<td>Race IAT</td>
<td>0.22</td>
<td>(.025)</td>
<td>32</td>
<td>1699</td>
<td>38</td>
<td>30.17 (.509)</td>
<td>0</td>
<td>10.17</td>
<td>16</td>
<td>4.90 (.027)</td>
<td>3.80 (&lt;.001)</td>
</tr>
<tr>
<td>Banaij (2009)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>McCall &amp; Carriger (1993)</td>
<td>Habitation</td>
<td>0.41</td>
<td>(.085)</td>
<td>12</td>
<td>447</td>
<td>25</td>
<td>6.74 (.820)</td>
<td>0</td>
<td>6.71</td>
<td>11</td>
<td>6.21 (.013)</td>
<td>2.24 (.025)</td>
</tr>
</tbody>
</table>

Note. ES = Effect size in Cohen's d, except for those in bold which are Fisher correlations with Cohen's d between the brackets; k = number of studies; N = total sample size; Med. N = median sample size; Q = test for homogeneity; I² = Higgins' I²; Exp. = expected number of rejections of H0 under ES; Obs. = observed number of rejections of H0; χ² = test of the excess of significant results; Regtest = Sterne and Egger's (2005) test for funnel plot asymmetry.
**Fig. 3.** Funnel plots of 11 (subsets of) meta-analyses from 2011 and Greenwald, Poehlman, Uhlman, and Banaji (2009). I-Chi(1) represents Ioannidis and Trikalinos’ (2007) test for an excess of significant results and BIAS Z represents Sterne and Egger’s (2005) test for funnel plot asymmetry.
Fig. 4. Funnel plots of simulations under true ESs of 0, .2, .5, and .8 (rows A through D) for the four strategies (columns 1 through 4) with varying sample sizes. I-Chi(1) represents Ioannidis and Trikalinos’ (2007) test for an excess of significant results and BIAS Z represent Sterne and Egger’s (2005) test for funnel plot asymmetry.

A

1

1 Large Study; True ES = 0

2

1 Large Study With QRPs; True ES = 0

3

5 Small Studies; True ES = 0

4

5 Small Studies With QRPs; True ES = 0

(continued)
Fig. 4 (continued)

1 Large Study; True ES = 0.5

Est ES = 0.497
I−Chi(1)=2.163, p=0.141
Z=1.616, p=0.106

1 Large Study With QRPs; True ES = 0.5

Est ES = 0.509
I−Chi(1)=11.11, p<.001
Z=2.985, p=0.003

5 Small Studies; True ES = 0.5

Est ES = 0.787
I−Chi(1)=7.908, p=0.005
Z=3.957, p<.001

5 Small Studies With QRPs; True ES = 0.5

Est ES = 0.779
I−Chi(1)=37.794, p<.001
Z=5.083, p<.001

5 Small Studies; True ES = 0.8

Est ES = 0.81
I−Chi(1)=0.996, p=0.318
Z=2.022, p=0.043

5 Small Studies With QRPs; True ES = 0.8

Est ES = 0.81
I−Chi(1)=3.574, p=0.059
Z=2.022, p=0.043

1 Large Study; True ES = 0.8

Est ES = 0.81
I−Chi(1)=0.996, p=0.318
Z=2.022, p=0.043

5 Small Studies; True ES = 0.8

Est ES = 0.955
I−Chi(1)=22.939, p<.001
Z=4.841, p<.001

5 Small Studies With QRPs; True ES = 0.8

Est ES = 0.897
I−Chi(1)=34.557, p<.001
Z=4.431, p<.001
supplemental for details), based on the four strategies as described above (Column 1 through 4) and four levels of true ES ($d = 0.0, 0.2, 0.5, \text{ or } 0.8; \text{ Rows A through D}). As can be seen, the pattern of results in the habituation–IQ studies (Fig. 1) looks highly similar to results simulated under Strategy 4 and an ES greater than 0 (Fig. 4; Panels B4 and C4). The funnel plot from Alfierei et al. (2011; educational psychology) resembles that from Strategy 2 under a small ES (Panel B2). In Woodley’s meta-analysis (2011; evolutionary psychology), the overall effect is close to zero and Strategy 2 appears to be at play (Panel A2). In Farber and Doodlin’s meta-analysis (2011; psychotherapy works better with positive regard), samples sizes were small and there is an indication of the use of Strategy 3 (Panel B3). In Hallion and Ruscio’s meta-analysis (2011), the effects of cognitive bias modification on stress and anxiety appear small and based on too many underpowered studies (Panel A3 or B3). Correlations between relationship conflict and hostility of partners as studied by Woodin (2011) appear to be substantial but may also be inflated by publication bias and the use of small samples (Panel C3). In these research lines, additional studies with larger sample sizes are clearly welcome.

We also included in Figure 3 and Table 1 a recent meta-analysis on the predictive validity of the Implicit Association Test (IAT). The subset of studies that concerned racial discrimination is another example of an excess of significant results and funnel plot asymmetry. The results from the 32 studies collated by Greenwald, Poehlman, Uhlman, and Banaji (2009) are based on small samples sizes, considerable freedom in the analysis, and a high degree of faddism, all of which may conspire to bring about inflated effects (Ioannidis, 2005). Further studies with larger sample sizes should be added to the database to accurately determine IAT’s validity.

**Improving the Game**

Without any clear rules concerning the use (and documentation) of multiple small studies and QRPs, strategic behaviors by researchers can lead literatures astray. The best way to separate the wheat from the chaff in psychology is to (a) end the pretense that small studies are definitive, (b) improve reporting standards, (c) start considering and publishing nonsignificant results, and (d) introduce a distinction between exploratory and confirmatory studies into journal policies (Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011). The prevalence of underpowered studies in psychological research hints at the common use of Strategies 3 and 4 in which researchers conduct a series of small studies up to the point that one turns out to be significant. Sample sizes should be based on a priori power analyses that take into account the potential inflation of effects in earlier small studies. Later publications of unpublished results and/or of statistical outcomes will be facilitated by including experimental material, raw data, and computer code (in the case of nonstandard analyses) as online supplements. In our view, researchers should be open about having not found what they looked for in early phases of research, while in the confirmatory phases, they should conduct studies that are possibly preregistered, sufficiently powerful, and analyzed in ways that are explicated in advance. The ideal paper then is not one with one or a few small studies with $p$ values just below .05, but one in which all small pilot studies are reported in a meta-analytic summary and tested for homogeneity and/or moderation and in which one major study lends clear support. Small and underpowered studies may lead to biases of different kinds (Ledgerwood & Sherman, 2012) and some have even argued to simply exclude them from meta-analyses (Kraemer, Gardner, Brooks, & Yesavage, 1998).

We found indications of bias in nearly half of the psychological research lines we scrutinized. The ambition of players in the game of science does not always sit well with the goal of the scientific enterprise. Optimal strategies for individual researchers introduce biases that we can only counter by improving the rules of the game. The arbiters in the game (peer reviewers and editors) are in an ideal position to do so.

**Declaration of Conflicting Interests**

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

**Funding**

The preparation of this article was supported by Netherlands Organization for Scientific Research (NWO) Grants 400-08-214 and 016-125-385.

**References**

References marked with an asterisk indicate studies included in the study sample


