Examining publication bias – A simulation-based evaluation of statistical tests on publication bias

Andreas Schneck [Corresp. 1]

[1] Department of Sociology, Ludwig-Maximilians-Universität München, Munich, Germany

Corresponding Author: Andreas Schneck
Email address: andreas.schneck@lmu.de

Background

Publication bias is a form of scientific misconduct. It threatens the validity of research results and the credibility of science. Although several tests on publication bias exist, no in-depth evaluations are available that suggest which test to use for the specific research problem.

Methods

In the study at hand four tests on publication bias, Egger's test (FAT), p-uniform, the test of excess significance (TES), as well as the caliper test, were evaluated in a Monte Carlo simulation. Two different types of publication bias, as well as its degree (0%, 50%, 100%), were simulated. The type of publication bias was defined either as *file-drawer*, meaning the repeated analysis of new datasets, or *p-hacking*, meaning the inclusion of covariates in order to obtain a significant result. In addition, the underlying effect (β = 0, 0.5, 1, 1.5), effect heterogeneity, and the number of observations in the simulated primary studies (N =100, 500), as well as in the number of observations for the publication bias tests (K =100, 1000), were varied.

Results

All tests evaluated were able to identify publication bias both in the *file-drawer* and *p-hacking* condition. The false positive rates were, with the exception of the 15%- and 20%-caliper test, unbiased. The FAT had the largest statistical power in the *file-drawer* conditions, whereas under *p-hacking* the TES was, except under effect heterogeneity, slightly better. The caliper test was, however, inferior to the other tests under effect homogeneity and had a decent statistical power only in conditions with 1000 primary studies.

Discussion

The FAT is recommended as a test for publication bias in standard meta-analyses with no or only small effect heterogeneity. If no clear direction of publication bias is suspected the TES is the first alternative to the FAT. The 5%-caliper tests is recommended under conditions of effect heterogeneity, which may be found if publication bias is examined in a discipline-wide setting when primary studies cover different research problems.

1 **Examining publication bias – A simulation-based evaluation of**
2 **statistical tests on publication bias**
3

4

5 Andreas Schneck[1]

6 [1] Department of Sociology, Ludwig-Maximilians-Universität München, Munich, Germany

7

8 Corresponding Author:

9 Andreas Schneck[1]

10 Konradstraße 6, Munich, 80801, Germany

11 Email address: andreas.schneck@lmu.de

12

Introduction

All scientific disciplines try to uncover truth by systematically examining their surrounding environment (Descartes 2006: 17). Natural scientists try to observe regularities in nature, whereas social scientists try to uncover patterns in the social behaviour of humans. This could be, for example, the development of pharmaceuticals or the evaluation of political interventions, such as the effect of minimum wages on employment. The success, as well as the reputation, of science rests on the accuracy as well as unbiasedness of scientific results. Publication bias, the publication of only positive results confirming the researcher's hypothesis (cf. Dickersin & Min 1993: 135), threatens this validity. Under publication bias results showing either statistical significance and/or the desired direction of the effects are published. The published literature in this case is merely a selective (and too optimistic) part of all existing scientific knowledge.

The study at hand examines the performance of four methods to identify publication bias: Egger's Test/FAT (Egger et al. 1997; Stanley & Doucouliagos 2014), p-uniform (PU; van Aert et al. 2016; van Assen et al. 2015), the test for excess significance (TES; Ioannidis & Trikalinos 2007a) and the caliper test (CT; Gerber & Malhotra 2008a; Gerber & Malhotra 2008b). In order to compare the performance of these tests, the false positive rate (α-error, type I error) and the statistical power (true positive rate) were examined with a Monte Carlo approach. This makes it possible to assess the performance of the four tests under different conditions of publication bias (*file-drawer* vs. *p-hacking*), as well as study settings (underlying true effect, effect heterogeneity, number of observations in primary studies and in meta-analyses).

The issue of publication bias

The classification of inferential statistics relies on the truth table (Table 1). An estimator (rows) tries to derive conclusions about the underlying true data (columns). The diagonal from the top left

36   to the lower right (in bold) describes a situation in which the estimator describes the underlying

37   data correctly. This can be either stating no existing effect (true negative) or stating an existing

38   effect (true positive). In the opposite situation the estimator states the wrong result, either an effect

39   if none is present (false positive) or no effect if one is present (false negative).

40   The false positive rate of a test (commonly called p-value) is the probability of the estimator

41   rejecting H_0 despite this being true. The p-value is therefore the probability that the observed

42   estimate is at least as extreme given there is no effect as assumed by H_0 (Wasserstein & Lazar

43   2016). The larger the p-value the higher the risk of assuming an effect if none exists in the data. p-

44   values below a certain threshold are called statistically significant, whereas values above the

45   threshold are labelled as non-significant. In the empirical sciences the 5%-significance threshold

46   is mostly used (Cohen 1994; Labovitz 1972; Nuzzo 2014). The difference between 0.049 and 0.051

47   in the error probability is, however, marginal. Nevertheless, from the standpoint of the 5%-

48   significance threshold the first would be a significant effect, whereas the latter would be a non-

49   significant effect. In both of these two cases on average around 1 in 20 null-hypotheses of no

50   difference would be rejected, albeit true. If empirical researchers select their data/models until they

51   find, just by chance, significant evidence that seems worth publishing, publication bias is on the

52   rise, leading to inflated or even artificial effects.

53   Rosenthal (1979) constructs a worst case scenario in which only the 5% of false positive studies

54   that are "significant" solely by pure chance are published. In this case, misinterpreted results shape

55   the scientific discourse and finally result in (medical or political) interventions. Although

56   Rosenthal's example is extreme, a multitude of evidence for publication bias exists in various

Peer J Preprints

57 disciplines and research fields.[1] Godlee (2012) therefore warns that scientific misconduct may also

58 physically harm patients. Chalmers (1990) also counts publication bias among general forms of

59 scientific misconduct because the consequences for the society as well as for science are similar.

60 In addition to the societal consequences, publication bias also has severe implications for the

61 evolution of knowledge. All scientific progress relies on the rejection of theories (Popper 1968:

62 215) but under publication bias no such rejection occurs, which leads to a state of "undead theory"

63 (Ferguson & Heene 2012: 559) where all existing theories are confirmed irrespective of their

64 truth.[2]

65 Statistically significant results furthermore stress the originality of research findings (Merton

66 1957). Both authors and scientific journals[3] have large incentives to maximise their significant

67 results to survive in a publish or perish research environment. Authors especially want to increase

68 their publication chances, notably in top-tier journals where low acceptance rates of 5%–10% are

69 quite common (for the top interdisciplinary ournals Nature 2017; Science 2017; cf. for the political

70 sciences Yoder & Bramlett 2011: 266). There are, in particular, two distinct strategies to achieve

71 significant results by means of publication bias practices. Firstly, non-significant findings can be

72 suppressed (cf. the classical file-drawer effect described by Rosenthal 1979) and significant results

73 are then searched for in another dataset. Secondly, small bits in the model of analysis can be

74 changed (e.g. adding covariates) until a significant result is obtained – this method is known as *p*-

---

[1] One of the most prominent examples in medicine is the case of Tamiflu, which is commonly used to treat influenza. The meta-analysis of Jefferson et al. (2012) shows that the results in the published literature are far too optimistic, especially considering side effects. But the problem is also of concern in the social sciences, where in the case of minimum wage research the literature mostly suggests no or only very low minimum wages in order to prevent negative side effects on employment, whereas no such negative effects exist (Doucouliagos & Stanley 2009).

[2] Existing true effects are then indistinguishable from false positives. Schoenfeld & Ioannidis (2013) demonstrate this in their meta-analysis in which they unsurprisingly find that in most of the studies they included in their meta-analysis nearly all commonly used cooking ingredients are labelled carcinogenic.

[3] For evidence on publication bias by reviewers and editors (Coursol & Wagner 1986; Epstein 1990, 2004; Mahoney 1977). In contrast other studies found no evidence (Dickersin et al. 1992; Lee et al. 2006; Olson et al. 2002).

75   *hacking* (cf. "fishing" Gelman 2013; or "researchers degree of freedom" Simmons et al. 2011:

76   1359). Whereas the *file-drawer* strategy can be utilised by authors as well as by editors and

77   reviewers, *p-hacking* can only be committed by authors/researchers. Nonetheless, *p-hacking*

78   strategies can be recommended by actors other than authors (e.g. editors, reviewers, etc.).[4]

79   Evidence on the prevalence of publication bias
80   So far, there are two strategies for identifying publication bias: the first traces studies through the

81   publication process, the second asks authors, reviewers, or editors about their publication practices

82   via surveys. In the first strategy, most of the analyses trace conference papers or ethics committee

83   decisions if those results get published or remain in the file-drawer. Callaham et al. (1998) trace

84   all papers submitted to a medical conference and find that significant findings have nearly twice

85   the chance of being published. Coursol & Wagner (1986) report from a retrospective survey of

86   psychological studies that in total positive findings are over three times more likely to be published.

87   However, full conference papers may already be biased because authors might submit their results

88   to a conference only if they are already significant (Callaham et al. 1998: 256). This problem of

89   underestimating publication bias might be overcome if the starting point is set more early on in the

90   research process.[5] Therefore, another approach is to trace the studies directly after an ethics

91   committee vote. The studies of Dickersin et al. (1992) and Easterbrook et al. (1991) confirm the

92   previous findings by reporting 2.32 higher chance of getting published and 2.54 higher publication

93   rates for significant studies. Also, Ioannidis (1998), who traces the study protocols of a large

---

[4] In contrast to fraudulently manipulated data, although publication bias is heavily punished, it is nearly impossible to detect at the individual level (Stroebe et al. 2012: 681) and in the case of *p-hacking* it is almost wholly without any costs, as data analysis tools/packages become increasingly easy to apply (Paldam 2013). Feigenbaum & Levy (1996) therefore even postulate the technological obsolescence of fraud. For evidence on the prevalence of fraud see (Baerlocher et al. 2010; John et al. 2012; Nuijten et al. 2016).

[5] Timmer et al. (2002) and Hua et al. (2016) report no evidence of unequal publication chances in the medical field of gastroenterology and dentistry. Both of these studies rely only on conference abstracts and therefore might be restricted in respect of the research results, because abstracts have the disadvantage that only a selection of results are presented in them, while in full papers all results are reported.

94   medical network over 10 years from 1986–1996, finds that significant studies have, beside their

95   3.7 times higher publication rate, a substantially higher publication speed, meaning a shorter time

96   between completion of the study and the final publication.

97   The second approach asks directly about the publication practices of the involved actors. In a

98   survey of psychologists that used a sensitive question technique up to 50% of the respondents

99   claimed that they exercised publication bias (John et al. 2012: 525). Franco et al. (2014) also note

100 that most non-significant findings go to the file-drawer right after the analysis and are not even

101 written up. Also, other less harshly sanctioned forms of misbehaviour, like optional stopping

102 (stopping data collection when significance is reached) or erroneous rounding of p-values to reach

103 significant results, are alarmingly widespread (prevalence rate arround 22.5% John et al. 2012:

104 525). These results are in line with the survey of Ulrich & Miller (2017: 9), who report that

105 researchers in the field of psychology prefer significant results over non-significant results, and,

106 furthermore, attribute more value to results with smaller p-values. These estimates may even be

107 conservative because it is known from the survey literature that sensitive behaviours like scientific

108 misconduct may be underreported (Kreuter et al. 2008: 848). According to the presented research

109 results *file-drawer* and *p-hacking* behaviour is therefore quite widespread.

110 ## Methods

111 Statistical tests on publication bias
112 So far, the presented detection strategies ask either directly for publication preferences or examine

113 the publication fate of conference papers. Both approaches have the weakness that they either rely

114 on the potentially biased answers of the actors involved or require an immense effort to follow the

115 publication process, while publication bias may have happened before the paper is submitted to a

116 conference. Statistical tests on publication bias circumvent this problem by relying only on the

117   published literature. In the paper at hand the regression-based Egger's test (Egger et al. 1997;

118   Stanley & Doucouliagos 2014), PU (van Assen et al. 2015), an extended version of p-curve

119   (Simonsohn et al. 2014a, b; Simonsohn et al. 2015), the TES (Ioannidis & Trikalinos 2007b) and

120   the CT (Gerber & Malhotra 2008a, b) are discussed.[6]

121   All of these tests are applied mostly in a discipline-specific context: the Egger's test is routinely

122   used in classical meta-analyses across all disciplines (cf. the Cochrae Handbook Higgins & Green

123   2008: 314), PU (for applications see Blázquez et al. 2017; Head et al. 2015; Simmons &

124   Simonsohn 2017), as well as the TES (for applications see Francis 2012a, b, c, d, e, 2013) are more

125   widely used in psychology. The CT is mostly implemented in the general social sciences (for

126   further applications in Soiology see Auspurg & Hinz 2011; Auspurg et al. 2014; Berning & Weiß

127   2015; in Psychology see Hartgerink et al. 2016; Kühberger et al. 2014). The discipline-specific

128   use of the tests is therefore to a certain degree path dependent on the practices involved in testing

129   publication bias in the specific fields.

130   Funnel asymmetry test (FAT)
131   The first class of tests makes it possible to address publication bias by the association of the effect

132   sizes and their variance. Because the variance (se^2) of an effect size in a primary study (es) is

133   strongly related to the sample size, small studies with a low number of observations (N) show an

134   increased variation of effects around the unobserved true effect. The larger the N, the smaller the

135   variation and thus the more precise is the effect size of the study. Under publication bias small

136   non-significant studies are mostly omitted, whereas small but precise effects with a large N still

137   remain in the analysis. When this pattern for a small positive effect is represented through a

---

[6] Because for Fail-save-N (Rosenthal 1979) only rules of thumbs instead of a formal statistical test exist it was not included in the simulation at hand. Although it is still widely applied (Banks et al. 2012: 183; Ferguson & Brannick 2012: 4), this benchmark is not recommended in the *Cochrane Handbook*, a guideline for conducting meta-analyses (Higgins & Green 2008: 321f.).

138   scatterplot graph a typical inverted funnel-shaped pattern can be observed (called "funnel plot"

139   Light & Pillemer 1984: 63-69). In the exemplary Figure 1 on the right, studies in the lower left

140   side are missing because of publication bias with a preference for significant positive effects. On

141   the left side, in contrast, a symmetric funnel with no publication bias is shown.

142   Relying only on subjective graphical information, as provided by funnel plots, might be misleading

143   (Tang & Liu 2000). Begg & Mazumdar (1994: 1089) examine the rank correlation of the

144   standardised effect ($t = es/se$) and its variance (se^2). A similar approach by Egger et al. (1997)[7]

145   regresses t on the inverse standard error ($1/se$). t is chosen as the dependent variable in order to

146   account for the unequal variance across the effects (heteroscedasticity) by weighting each

147   observation by the inverse of its variance. Compared to the regression of se on es this changes the

148   interpretation.

149
$$t_i = \beta_0 + \beta_1 \frac{1}{se_i} + \varepsilon_i$$

150   The constant β_0 is the test on publication bias (FAT stating publication bias if $\beta_0 \neq 0$), whereas β_1

151   makes it possible to identify a true empirical effect controlling for publication bias (Egger et al.

152   1997: 632). In the left graph of Figure 2 a primary study (depicted as dots), with almost no

153   precision, is not able to find an effect (H_0: $\beta_0 = 0$ could not be rejected). In contrast, in the right

154   graph under publication bias studies with no precision also find a substantial effect.

155   In the following simulation only the FAT is used because of its better statistical power, as found

156   in previous simulations (Hayashino et al. 2005; Kicinski 2014; Macaskill et al. 2001; Sterne et al.

[7] This estimator is equivalent to the bivariate FAT-PET recommended by Stanley & Doucouliagos (2014). The FAT-PET furthermore makes it possible to also include "potential effect modifiers" (Deeks et al. 2008: 284) in a meta-regression model. This is especially necessary if the literature being studies has, besides its theoretical meaningful overall effect, systematic differences (e.g. different implementations of an experimental stimulus, different experimental populations, etc.).

157   2000), compared to the rank correlation test of Begg & Mazumdar (1994).[8] Despite its strengths,

158   the central weaknesses of the FAT lies in its low statistical power in a setting with only a small

159   number of primary studies (Macaskill et al. 2001 simmulated the performace only based on 20

160   primary studies).[9]

161   p-uniform (PU)

162   The tests discussed so far focus on the empirical effect sizes, whereas the p-curve method,

163   proposed by Simonsohn et al. (2014b), and the similar PU, a method proposed by van Assen et al.

164   (2015), focus entirely on the distribution of significant *p*-values. All non-significant values are

165   therefore dropped from the analysis. The sample is, furthermore, restricted to the direction of

166   suspected publication bias: that means only positive or negative effects are examined (Simonsohn

167   et al. 2014a: 677). In the first step the *p*-value of the estimate in the primary study is rescaled in

168   respect to the significance threshold. For the present study the 5%-significance threshold ($p = 0.05$)

169   rescales the *pp*-values to the range [0,1]. This *p*-value of *p*-values (*pp*-value) reflects the probability

170   under the null hypothesis of a non-existing effect that a *p*-value would be as small as, or even

171   smaller than, the observed one.[10]

172
$$pp_i = \frac{p_i}{0.05} = \frac{1 - \Phi\left(\frac{es_i}{se_i}\right)}{0.05} \quad if \; p_i < 0.05$$

---

[8] The regression-based test also shows superior properties compared to the trim and fill technique (Bürkner & Doebler 2014; Kicinski 2014; Moreno et al. 2009; Renkewitz & Keiner 2016), which tries to obtain a symmetrical funnel plot by imputing studies that might be missing due to publication bias (Duval & Tweedie 2000).

[9] In addition to the performance of the FAT, multiple simulation studies (Alinaghi & Reed 2016; Paldam 2015; Reed 2015) also examine the unbiasedness of the effect estimate (PET - the estimated underlying effect size corrected on publication bias) which is not of interest in the study at hand. The PET is especially threatened by an increased false positive rate under effect heterogeneity (Deeks et al. 2005; Stanley 2017), the properties of the FAT in these conditions have not yet been examined.

[10] Φ represents the standard normal distribution

173   In a second step the skewness of the *pp*-distribution is tested (Simonsohn et al. 2015: 1149). Right

174   skewness shows an overrepresentation of findings with a substantial statistical significance and

175   indicates a genuine empirical effect. Left skewness, in contrast, shows an overrepresentation of

176   just significant estimates that barely pass the significance threshold (in this case 5%) and indicates

177   publication bias under the null hypothesis (Simonsohn et al. 2014b: 536).

178   Whereas p-curve by Simonsohn et al. (2014b) only allows to identify publication bias under a true

179   underlying null effect, PU (van Assen et al. 2015) allows to also identify publication bias under an

180   empirically observed effect. Therefore, p-curve is a special case of PU. For PU the underlying

181   effect has to be estimated empirically by a fixed-effect meta-analysis (FE-MA)[11] with all primary

182   studies. In a second step, and equivalent to p-curve, only k estimates with $p < 0.05$ and the direction

183   of the suspected publication bias remain in the analysis (van Aert et al. 2016: 727). By adjusting

184   on the existing underlying effect, the fixed-effect estimate μ, it is possible to test the skewness of

185   the distribution conditional on the underlying empirical effect (van Assen et al. 2015). In the

186   numerator, the effect size estimate is conditioned on the underlying effect (μ), similar to a one-

187   sample z-test. The denominator of the *pp*-value is not fixed to 0.05 as in p-curve, but is also

188   conditioned on the underlying effect (μ), which is subtracted from the effect threshold (*et*) an effect

189   has to reach to become statistically significant given its standard error (*se*).

190
$$pp_i^\mu = \frac{1 - \Phi\left(\dfrac{es_i - \mu}{se_i}\right)}{1 - \Phi\left(\dfrac{et_i - \mu}{se_i}\right)} \quad if\ p_i < 0.05$$

[11] Mean effect size across all included studies weighted by the inverse study variance.

191 The test statistic is gamma-distributed with k degrees of freedom.[12] Because the skewness is now

192 conditional on the underlying empirical effect left skewness observed by PU identifies publication

193 bias across all underlying empirical effects, as depicted in Figure 3.

194 Because PU rests on the average effect size estimated by a fixed-effects meta-analysis it is sensible

195 to effect heterogeneity. The degree of heterogeneity which invalidates the test is, however, unclear.

196 Whereas Simonsohn et al. (2014a: 680) state that p-curve, and therefore also PU, is also

197 appropriate under effect heterogeneity, van Aert et al. (2016: 718) note exactly the opposite.

198 van Assen et al. (2015) evaluate the performance of PU and the TES (a publication bias test,

199 discussed in the next section), and trim-and-fill, and conclude that PU has a greater statistical

200 power than the other methods (van Assen et al. 2015: 303). Also, Renkewitz & Keiner (2016)

201 evaluate the PU publication bias test and observe its slightly better performance compared to the

202 FAT and the TES. However, in both studies the number of studies in the meta-analyses (max. 160),

203 as well as the number of observations (max. 80) in the primary studies, is relatively small.[13]

204 Test for excess significance (TES)
205 The TES (Ioannidis & Trikalinos 2007b; also called ic-index see Schimmack 2012) builds on the

206 observed power of every single study to uncover the true total effect. This true effect is estimated

207 by a fixed-effect meta-analysis, as in PU. Observed power analyses make it possible to compute

---

[12] $p = \Gamma\left(k, -\Sigma_{i=1}^{k} \log\left(pp_i^{\mu}\right)\right)$

[13] Similar to the FAT-PET, evaluations of PU center mainly on the estimated overall effect. While van Assen et al. (2015) show a good coverage of the estimated overall effect, McShane et al. (2016) state, in contrast, that while "p-curve and p-uniform approaches have increased awareness about the consequences of publication bias in meta-analysis, they fail to improve upon, and indeed are inferior to, methods proposed decades ago" (McShane et al. 2016: 744).

208   the post hoc power (pw_i) of a study. This allows to specify the expected number of significant

209   effects E, given the average effect as well as the significance threshold (in this case $\alpha = 0.05$).[14]

210
$$E = \sum_{i=1}^{k} (pw_i)$$

211   E may even be a conservative estimate of the expected number of significant studies because it

212   heavily relies on the fixed-effect estimate, which suffers from an eventual publication bias. In

213   relation to O, the empirically observed number of significant studies ($p_i < 0.05$) the TES tests

214   whether more significant results than expected are reported in the literature. To test whether the

215   share of observed positive outcomes $\left(\frac{O}{K}\right)$ is larger than the share of expected positive outcomes $\left(\frac{E}{K}\right)$

216   a one-sided binomial test is used (Ioannidis & Trikalinos 2007b: 246).

217   On exemplary datasets the TES performs considerably better under moderate effect heterogeneity

218   in large meta-analyses, where the FAT in particular failed to uncover publication bias (Ioannidis

219   & Trikalinos 2007b: 248). Nevertheless, Johnson & Yuan (2007: 254) ask if the TES makes it

220   possible to dissect between publication bias and study-heterogeneity accurately. Therefore, the

221   authors of the *Cochrane Handbook* (Higgins & Green 2008: 323) express the need for further

222   evaluations.

223   Caliper test (CT)
224   In contrast to the aforementioned three tests, the CT, developed by Gerber and Malhotra (2008a,

225   b), ignores most of the information provided by the studies included and looks only at a narrow

226   interval (caliper = c) around the significance threshold (th) in a distribution of absolute z-values.

227   In case of a continuous distribution of z-values, studies in the interval below the significance

---

[14] Although Hoenig & Heisey (2001) criticise the application of post-hoc power analyses in primary studies for the good reason that the observed power estimate may be biased, meta-analyses circumvent this critique because a distribution of power estimates allows to infer more accurately the power of a set of studies.

228   threshold (in the so-called over-caliper; $x_z = 1$) should be as likely as just non-significant studies

229   (in the so-called under-caliper; $x_z = 0$).

230
$$x_z = \begin{cases} 0 \; if \; th - c * th < z \leq th \\ 1 \; if \; th < z < th + c * th \end{cases}$$

231   Gerber and Malhotra (2008a, b) use a 5%, 10%, 15% and 20% interval (c) proportional to the

232   significance threshold (th). In particular, the widest 20% caliper may be too wide because the 10%-

233   significance level that could be another target threshold for publication bias is fully overlapped.

234   The higher the overrepresentation in the over-caliper, the higher the likelihood of publication bias.

235   This is also shown in Figure 4: in the left graph with no publication bias no discontinuities are seen

236   around the arbitrary 5% significance threshold (dashed line), whereas in the right graph a stepwise

237   increase of just significant results indicates publication bias. As with the TES, a one-sided binomial

238   test is used to test the equal distribution of z-values in the over- and under-caliper.[15]

239   Publication bias tests in comparison
240   In order to compare the different publication bias tests presented, four different criteria have to be

241   established: the measurement level, the sample used, the assumptions in connection with the test

242   method, and its according limitations (cf. Table 2).

243   Whereas the FAT and PU explicitly model the test value distribution, the TES and the CT rely

244   only on dichotomous classifiers. The PU and the CT furthermore restrict their sample to either

245   significant estimates of positive or negative sign (PU) or estimates in a close interval around the

246   significance threshold (CT). Both criteria lead to a hierarchy of information: the FAT relies on all

247   available information, whereas the TES and PU, and the CT most strongly, rely only on limited

---

[15] Masicampo & Lalande (2012) and Leggett et al. (2013) test the deviance of values around the significance threshold from a fitted exponential curve on p-values in a broader range from $0.1 - 0.10$ to counter the huge loss of observations in the CT. This may be problematic, because a single distributive function may not be able to describe the pattern well enough across the suspected jump points (cf. Lakens 2015). In the case of substantial effect heterogeneity this problem would be aggravated even further.

---

248 information on the published estimates. This may reduce the statistical power of the tests despite

249 being a useful means of circumventing certain limitations or fulfilling assumptions, as discussed

250 later on.

251 The FAT has the assumption that study precision drives publication bias (there is more publication

252 bias in smaller and less precise studies). This has the disadvantage that variation in the number of

253 observations in primary studies is necessary. Another disadvantage is that only directed publication

254 bias either in favour of a positive or negative significant effect can be tested. In the extreme case

255 of only a significant positive and significant negative effects due to publication bias no publication

256 bias can be detected by the test. Publication bias in this case is nonetheless visible in the funnel

257 plot. In addition, PU is only able to detect directed publication bias. The TES and the CT do not

258 have this limitation of either a preference for positive or negative estimates. In contrast to the FAT,

259 all the other tests are only able to test for publication bias in respect of a specific significance

260 threshold (e.g. 5%).

261 The FAT has the central assumption that all variation of the effect should be independent of the

262 study precision and therefore N (number of observations in the primary studies). PU has the

263 assumption that every left skewness in the pp-value distribution is caused by publication bias. This

264 assumption is, however, grounded mainly on the -effect estimate, which is very sensitive to effect

265 heterogeneity. The same problem applies to the TES, which also relies on a fixed-effect estimate.

266 Despite having the disadvantage of a vast amount of information and thus statistical power, the

267 CT has the advantage of being unaffected by underlying effect distribution, as well as publication

268 bias direction. Therefore, no assumption, despite being continuous, has to be made. In particular,

269 jumps in the distribution around the significance threshold should therefore be highly unlikely.

270 The assumption of a uniform distribution ($P(x_z) = 0.5$) of the under- and over-caliper is stronger,

271   the narrower the caliper (c) is set. This is because narrower calipers are less sensitive to the overall

272   shape of the z-value distribution.

273   No evaluation of these tests exists based on a larger number of primary studies. In particular, the

274   newer publication bias tests like PU, as well as the TES and the CT, are in need of an evaluation

275   under different conditions. For the CT also no studies exist regarding the best caliper width to use.

276   Despite the existence of some simulation studies on publication bias tests, so far no direct

277   comparison exists that evaluates the performance of all available publication bias tests. Such an

278   evaluation can particularly guide the choice of publication bias tests under substantial effect

279   heterogeneity.

280   Simulation setup

281   In order to examine the performance of the four publication bias tests, a Monte Carlo simulation

282   approach is used. For the simulation two different processes have to be distinguished: firstly, the

283   data generation process (DGP), and, secondly, the meta-analytical estimation method (EM). The

284   DGP provides the ground for the hypothetical data used by the simulated actors, as well as the

285   results they report, whereas the EM applies the tests on publication bias reported in the previous

286   section. The central advantage of using Monte Carlo simulations is that controlling the DGP allows

287   us to identify which simulated studies suffer from publication bias and which do not. Similar to

288   the case in experiments, different conditions can be defined to ensure a controlled setting. The

289   performance of the estimators can then be examined under the different conditions.

290   The first step of the DGP (cf. Table 3) defines different effect size conditions that underlie the

291   analyses of the simulated actors. In order to cover low to medium effects, as defined by the large-

292   scale literature survey of Bosco et al. (2015: 436), as a first condition the underlying true effect

293   was specified by a linear relationship with $\beta = 0, 0.5, 1.0, 1.5$. In addition to the homogenous

294   conditions with a common effect size, a heterogeneous condition was added that assumes no fixed

295   distribution of an underlying effect but a uniform mixture of all four effect sizes, as defined above,

296   plus an additional effect of $\beta = 2.0$ in order to ensure enough variation. The specified linear

297   relationship between the dependent variable y and the independent variable x had a normally

298   distributed regression error term of $\varepsilon = \Phi(0,10)$, while the variation of the independent variable

299   was defined as $\sigma_x = 2$ (for a similar setup see Alinaghi & Reed 2016; Paldam 2015).

300   In order to quantify effect heterogeneity the I^2 measure (Higgins & Thompson 2002) is used. This

301   approach allows us to differentiate between the random variation that is driven by N and the

302   variation that is caused by underlying effect heterogeneity. I^2 allows us to specify the share of

303   variation caused by true effect heterogeneity and the total variation that consists of random and

304   true effect heterogeneity. Although effect heterogeneity is best addressed directly in meta-

305   regression models this approach is not possible if the sources of heterogeneity are unknown or too

306   diverse. This may be the case when analysing publication bias in a discipline-wide literature with

307   no common underlying effect (e.g. sociology, psychology, etc.).

308   Because the FAT is based on study precision, which is mainly driven by the number of

309   observations (N) of the primary studies, N was computed as a second condition by an absolute

310   normal distribution with a mean of 100 (small N) or 500 (large N) and a standard deviation of 150.

311   In order to ensure an adequate statistical analysis for the primary studies, Ns equal to or smaller

312   than 30 were excluded.[16] This procedure resulted in a right skewed distribution with a mean N of

313   roughly 500 for the large N, and 165 for the small N condition.

---

[16] The statistical power in studies with $N \leq 30$ is very small and therefore the normality assumption of the regression error term is not met.

314   Due to the two manipulated conditions, the true effect and the study N, the statistical power to

315   detect an underlying true effect varied widely across the simulated primary studies with an

316   underlying true effect, ranging from effectively powerless studies (9.7%) to very powerful ones

317   (92.7%). On average, a study had a statistical power of 42.6%. 16.6% of the studies were

318   adequately powered with at least 80% power (Cohen 1988: 56). The setting produced by the DGP

319   also reflects the results of Ioannidis et al. (2016: 15), who report that only 10% of the studies in

320   economics are adequately powered.

321   In addition to the number of observations in the primary studies (*N)* the number of primary studies

322   that were included in the meta-analysis and form the basis of the publication bias tests (*K*) was

323   varied in the third condition. A setting with 100 studies was used as a lower condition, whereas

324   1000 studies were set as an upper condition. The small *K* as well as large *K* condition define the

325   space in which meta-analyses are applicable with an adequate statistical power.

326   Building on this data setup stage of the DGP the behavioural setup adds publication bias to the

327   simulation in a fourth step. In the simulation setup publication bias is defined as the willingness to

328   collect new data or run additional analyses if statistical significance failed ($p \geq 0.05$) or a negative

329   effect occurred. Five different conditions have to be distinguished. Firstly, the condition without

330   publication bias: in this ideal case all estimates (βx) are estimated by a bivariate ordinary least

331   squares (OLS) model and afterwards published. Publishing in terms of the simulation model means

332   that all estimates enter the final meta-analysis. Therefore, in the condition without publication bias

333   either 100 or 1000 regression results are estimated and enter the meta-analysis. It is important to

334   note that publication bias in the simulation model at hand is only the intention to commit

335   publication bias. The actual publication bias depends on the data setup itself: how large is the true

336   effect size (β) and the number of observations (N) in the primary studies? Or, in short: is there

337   already a significant positive result which does not need a publication bias treatment?

338   In the second and third conditions publication bias is present with a 50% probability. That means

339   that 50% of the actors are willing to run additional analyses in order to obtain significant results.

340   These conditions seem closest to the behavioural benchmark of the empirical studies presented.

341   If a non-significant result is obtained, actors operating under the second condition choose to collect

342   new data in order to obtain significant results that can be published. This second condition

343   therefore models publication bias under the *file-drawer* scenario, because the datasets not used

344   remain unpublished. An actor tries to run analyses on the basis of up to nine additional datasets

345   and only stops earlier if a significant result with a positive sign is obtained. If none of the 10

346   datasets yields a significant relationship with a positive sign, the estimate which is closest to the

347   significant threshold is published. This rule serves two purposes: firstly, it seems plausible that an

348   actor who has tried that many analyses wants to get the results published in the end to compensate

349   for the invested effort and to avoid sunk costs (Thaler 1980). Secondly, from a technical point of

350   view, this allows to keep the number of observations in a meta-analysis K constant across all

351   simulation conditions.

352   In the third condition an actor does not try to achieve significant results by running the same

353   bivariate analysis on different samples, but rather tries to run different model specifications on the

354   same data by including control variables (z_j) to achieve statistical significance of the coefficient of

355   interest (βx). The third condition therefore models publication bias as *p-hacking*, because the

356   existing dataset is optimised to receive a significant p-value. The actor is able to add three different

357   control variables to the model. The control variables are defined as collider variables that are both

358   an effect of x as well as y, which biases the effect of interest (Cole et al. 2009; Greenland et al.

359   1999). The effect of x and y on z_j is, however, only small ($\gamma = 0.5$). The error term of the equation

360   defining z is normally distributed $\Phi(0,10)$. With three available control variables z_j an actor has

361   seven different combinations to improve the research results to obtain a significant effect of x on

362   y.

363   In contrast to the second and third conditions, where 50% of the actors have the intention to commit

364   publication bias, in the fourth and fifth conditions all actors have the intention to engage in

365   publication bias practices, once again either through *file-drawer* (fourth condition) or *p-hacking*

366   behaviour (fifth condition). Part from the higher degree of intention to engage in publication bias

367   practices the settings remain the same. Although the two conditions where all actors have the intent

368   to engage in publication bias are far too pessimistic, they allow us to evaluate the performance of

369   the tests in the most extreme publication bias environment. Tests that are not able to detect

370   publication bias even under such extreme conditions are of low utility to the research community

371   to identify publication bias.

372   The resulting design matrix has 100 different combinations resulting from 20 data setup conditions

373   multiplied by the five publication bias conditions (see Table 3). In order to obtain reliable estimates

374   similarly to in an experiment (Carsey & Harden 2013: 4f.), every single cell of the design matrix

375   has to be replicated multiple times. In order to specify the numbers of replications that are

376   necessary to achieve a sufficient statistical power of at least 80% (Cohen 1988: 56) a power

377   analysis was conducted for the statistical power, as well as the false positive rate estimates (see

378   Table 4). For the false positive rate a small deviation of 1 percentage point from the set 5%-false

379   positive rate has to be correctly identified with at least an 80% chance. To achieve this goal, every

380   condition without publication bias had to be supported with 3,729 runs. As deviations in power

381   are, though important, not as essential as the false positive rate (Cohen 1988: 56) a difference of 3

382   percentage points is set as acceptable. In order to identify a 3 percentage point deviation from the

383   target power of 80% each of the 80 conditions with existing publication bias needed 1,545 runs.

384   In total, 198,080 runs were necessary, resulting in nearly 109 million primary studies that in the

385   case of publication bias contained up to 10 different regression models.[17]

386   The aim of the simulation study at hand is to compare the performance of the four tests in respect

387   of: A) their capability to detect publication bias if present (true positive, statistical power), as well

388   as B) consistent false positive classification (α-error) as shown in the second row of the truth table

389   (cf. Table 1). Because the conditions with and without publication bias are known in a simulation

390   study, the power of the tests as well as the false positive rate is computable (Mooney 1997: 77-

391   79). In a first step, a dummy variable (s) is constructed, with the value 1 for a significant test result

392   below the significance threshold (5% significance level). To obtain the power, the first estimate is

393   restricted to the conditions with publication bias and sums up the indicator variable s over all runs

394   (r) in that condition. The power is than defined as the proportion of significant results s in respect

395   to r (see Table 4). The false positive rate is computed equivalently but only in conditions without

396   any publication bias.

397   Results

398   The simulation process took about five days on a medium-performance computer that executed

399   more than 232 million regressions. Because the publication bias in the experimental setup was

400   implemented by the intent to commit publication bias (which does not necessarily need to result

401   in actual publication bias), two further variables are useful to interpret the results: the share of

---

[17] The simulation routine was written in Stata and builds heavily on the fastreg.ado of Geertsema (2014), which makes it possible to speed up the process immensely by providing a stripped-down OLS-regression command. The analysis files are available for replication purposes.

402 actual studies per meta-analysis that suffer from publication bias (if $p < 0.05$ or negative result in

403 the first trial) and the share of studies that achieve their goal of a significant positive result by

404 publication bias.

405 In the two conditions where 50% of the actors have the intention to commit publication bias to

406 reach their goal of a significant positive effect, on average 21.7% did so (cf. Table 5). Also in the

407 two conditions where 100% are willing to commit publication bias only 43.4% engaged in

408 publication bias practices because they already had achieved significant results beforehand. In

409 other words: over all conditions, 56.6% of the primary studies already achieved significant results

410 that did not call for a publication bias treatment. The success rate of those actors committing

411 publication bias were, because of the limited number of trials (either by the maximum number of

412 datasets or control variable combinations), with around 58% for both *file-drawer* and *p-hacking*

413 far from a guaranteed success. In the 50% publication bias condition, on average 12.5% of the

414 studies achieved significant results with the help of publication bias practices, while in the 100%

415 publication bias condition 25% did so. Neither of these results differ between *file-drawer* and *p-*

416 *hacking*. This means that the leverage of both publication bias conditions to get significant results

417 does not differ in the chosen setting, which allows us to compare the performance of the

418 publication bias tests to identify both strategies.

419 In the condition with no effect heterogeneity, on average only a negligible 4.3% of the variation

420 was attributed to effect heterogeneity, whereas in the condition with effect heterogeneity 73.3% of

421 the variation was attributed in this way. In terms of Higgins & Thompson (2002: 1553), an I^2 larger

422 than 50% has to be modelled explicitly in meta-analyses and cannot be ignored.

423 Table 6 shows the false positive rates of the publication bias tests across all simulated conditions.

424 The false positive rate of the test was fixed in the simulation setting to 0.05 (again, see Table 4),

425   so all false positive rates should be equal to, or even smaller than, 0.05. Positive deviations from

426   0.05 point to inflated false positive rates, which lead to more false conclusions than expected. In

427   Table 6 these values are highlighted in bold. Over all conditions the FAT, PU, the TES, as well as

428   the narrower CTs (3%, 5%), had a consistent false positive rate. The FAT was closest to the

429   expected 5% error rate. PU and the TES, as well as the 3% and 5% CTs, in contrast, are in most

430   cases very conservative because they fall far below 0.05. This over-conservatism may be

431   problematic in respect to a decreased statistical power, a matter which is discussed later on. The

432   wider 10% and 15% CTs suffered under inflated false positive rates because, due to the large

433   caliper width, the assumption of a uniform distribution in both calipers was violated.[18] For the 10%

434   CT the specified false positive rate doubles to more than 10%, whereas in case of the 15% CT it

435   more than quadruples.

436   Looking at conditions with 50% publication bias in the *file-drawer* condition (see Table 7), the

437   FAT had a superior power compared to other tests in 14 of 20 conditions, as indicated by the

438   underlined numbers. The FAT is, however, closely followed by the TES, which had a larger

439   number of conditions with a satisfactory power (> 0.8) compared to the FAT (7 vs. 6). In the first

440   condition with $N = 100$ as well as $K = 100$ the TES was superior in the case of an underlying small

441   or moderate effect ($\beta = 0.5; 1; 1.5$). The large variability of the primary study effect, which was

442   caused by the low-N and low-K in the meta-analyses, resulted in an overall minor statistical power.

443   A sufficient power (highlighted in bold) was only reached in conditions with a low or moderate

444   underlying true effect ($\beta = 0.5, 1$). None of the CTs yielded a sufficient power. This picture changes

445   if more studies were included in the meta-analysis. With $K = 1000$ most of the tests yielded a

446   sufficient power. In particular, the FAT had a statistical power close to 100%, also under effect

---

[18] This means that an asymmetry between over- and under-caliper is not caused by publication bias rather than by an underlying effect distribution that is skewed in the caliper width.

447   heterogeneity. The PU and the TES failed to uncover *file-drawer* behaviour under effect

448   heterogeneity, but performed well under homogeneity. PU was only able to discover *file-drawer*

449   behaviour under low underlying true effects. The CTs profited the most from an increased K, the

450   wider caliper (10, 15%) had a larger statistical power than the narrower ones but also had inflated

451   false positive rates (see Table 6) that might invalidate the conclusions (grey shaded area). The

452   narrower caliper had a sufficient power only in studies with no or small underlying effects ($\beta = 0$;

453   0.5). $K = 100$ and $N = 500$ decreased the power of all tests drastically. In this condition the FAT

454   had the largest, but still not satisfactory power. With $K = 1000$ a sufficient power is yielded in

455   conditions with a low overall effect ($\beta = 0$; 0.5).

456   The statistical power of the tests increased if the intent to engage in *file-drawer* behaviour is set to

457   100% (see Table 8). Overall, more publication bias tests achieved a satisfactory statistical power

458   to detect publication bias. Also, in these conditions, the FAT dominated in 13 of 20 conditions. As

459   before, neither the TES nor the PU were able to detect publication bias under effect heterogeneity.

460   The TES was, furthermore, not able to detect publication bias with an underlying null effect.

461   Similar to the 50% *file-drawer* condition, the CTs showed a drastically decreased power in

462   conditions with $K = 100$.

463   The dominance of the FAT weakened when looking at the 50% *p-hacking* condition (see Table 9).

464   Instead, the TES was besides the 15% CT superior under most conditions but had the advantage

465   that its false positive rate was not inflated. The overall pattern was, however, quite similar: both

466   PU and TES had almost no power to detect *p-hacking* under effect heterogeneity. Also, the

467   statistical power was only satisfactory for PU when $K = 100$. With a large number of included

468   studies, however, the power of the CT was close to, or even outperformed, the FAT, PU and the

469   TES.

470   In the 100% *p-hacking* condition (see Table 10) the FAT caught up with the TES and yielded an

471   increased power, especially in the case of $K = 100$. Despite the dominance of the 15% CT, the TES

472   and the FAT closely followed. The CT had a similar strength to that demonstrated in the earlier

473   conditions under effect heterogeneity and $K = 1000$. The underperformance of all tests in the

474   condition with $N = 500$ and moderate underlying effects ($\beta = 1; 1.5$) is caused by the already

475   existing significance of most results in this condition.

476   Overall, the FAT dominated under the *file-drawer* condition. The TES, in contrast, had a slightly

477   higher statistical power than the FAT under the *p-hacking* condition without effect heterogeneity.

478   However, the differences between both tests were quite small. The CTs performed well under the

479   *file-drawer* as well as *p-hacking* condition with heterogeneous effect sizes and large numbers of

480   studies included ($K = 1000$). Although the 10% and 15% caliper had the highest power to detect

481   *p-hacking* these tests should not be applied due to their increased false positive rate.

482   In order to evaluate the tests on publication bias by their underlying risk factors (already significant

483   results), rather than the conditions of the simulation, a regression model limited either on non-

484   publication bias conditions or publication bias conditions was run. The dependent variable in both

485   cases was the dummy variable (s in Table 4) for $p < 0.05$ for each publication bias test under

486   examination. Linear probability models were used for the estimation.[19]

487   Table 11 shows the false positive rate in dependence of the number of studies included ($K =$

488   100|1000) and the effect heterogeneity (I^2). In the constant condition of a meta-analysis with $K =$

489   100 and no effect heterogeneity none of the tests had larger false positive rates than the expected

490   0.05. In particular, the TES and the 3% and 5% CTs were very conservative. A larger meta-

---

[19] Although the effect of the different conditions could be achieved with logistic regressions and average marginal effects, an intercept cannot be estimated in these models.

491   analytical sample increased the false positive rates for the TES and the CTs. The broadest 15% CT

492   missed the expected significance threshold of 5%, with 7.8%. Increasing effect heterogeneity

493   resulted in more conservative false positive rates for PU and 15% CT, and to a smaller extent also

494   for the FAT, the TES and the 10% CT. The narrower 3% and 5% CTs were unaffected by effect

495   heterogeneity. The overall influence of the varied conditions on the false positive rate was small,

496   as can be seen by the small R^2 ($< 1.7\%$).

497   The following regression model (Table 12) addresses the statistical power. Starting from the

498   baseline condition of a meta-analysis with $K = 100$, a mean share of publication bias committed

499   (32.6%. mean of the 50% and 100% publication bias condition, cf. Table 5), as well as successfully

500   applied (18.8%, cf. Table 5) via a *file-drawer* procedure and no effect heterogeneity, the FAT had

501   a superior power of 56.9%, followed by the TES (51.5%) and the PU (48.3%). The CTs performed

502   worst and yielded only a power of 0.0%–38.6%. The underperformance of the CTs is largely

503   explained by the small number of studies in the meta-analyses. With $K = 100$ hardly any study

504   falls within the small caliper around the significance threshold. This limitation on just significant

505   or non-significant effects also led to missing values, because without observations in the caliper

506   no CT could be performed. The underperformance of the CT changed if 1000 studies were

507   included, which improved the estimated power substantially, by 30.7–57.3 percentage points,

508   while smaller calipers profited most. The FAT, as well as the TES and the PU, profited moderately

509   from an increased number of studies, by 24.4, 23.8 and 16.5 percentage points, respectively. When

510   focussing on the influence of heterogeneity in the meta-analyses the PU and the TES showed a

511   drastic drop in power, by 6.5 and 6.4 percentage points, if the heterogeneity rose by 10 percentage

512   points. This decrease in power shows that neither PU nor TES were able to cope with

513   heterogeneity. In contrast, the FAT and the CTs actually showed a slight increased statistical

514   power. Varying the publication bias procedure from a *file-drawer* mechanism to *p-hacking,* which

515   is less related to the standard error of the effect estimates, increased the power of PU, TES and the

516   CTs. The CTs profited most, increasing the statistical power by around 18 percentage points. The

517   TES and PU showed a smaller increase of power, by 7.5 and 4.8 percentage points. The FAT, in

518   contrast lost about 11 percentage points compared to its power under a *p-hacking* procedure.

519   The structural difference between tests based on a continuous effect distribution (FAT, PU) and

520   tests that focus only on a dichotomous classification (TES, CTs)[20] becomes clear when looking at

521   the effect of the proportion of studies that underwent a publication bias treatment in the simulation

522   and the proportion of studies that had a successful outcome after publication bias. Increasing the

523   share of studies under publication bias lifted the power by 3.0 (FAT) and 5.1 (PU) percentage

524   points. A 10 percentage point increase in studies successfully applying publication bias increases

525   the power by 9.9 (FAT) and 10.3 percentage points (PU). The TES and the CTs, however, were

526   only able to detect successful publication bias. An increase only in studies committing publication

527   bias (whether successful or not) reduced the statistical power. Both tests were therefore not able

528   to detect all outcomes of publication bias. This is especially problematic as non-successful

529   publication bias may also increase the overall estimated effect in meta-analyses. All effects

530   presented are statistically significant ($p < 0.05$).

531   In contrast to the influence of the varied conditions on the false positive rate, the influence on

532   statistical power was substantial, varying from 30.6% in the case of the FAT to 57.2% for the PU.

533   This finding underlines the fact that all publication bias tests have their strengths and weaknesses

534   in specific conditions.

535   Discussion & Conclusions

---

[20] Significant or not (TES) over- or under-caliper (CTs).

536   In the simulation at hand, the performance of four different tests (PU, FAT, TES, CTs) was

537   evaluated by a Monte Carlo simulation. Different conditions were varied: the underlying true effect

538   size, including effect heterogeneity, the number of observations in the primary studies, the number

539   of studies in the meta-analyses, and the degree of publication bias and its form as either *file-drawer*

540   or *p-hacking*. Based on a Monte Carlo simulation with 100 different conditions and nearly 200,000

541   simulated meta-analyses the following recommendations can be made (cf. Table 13).

542   Firstly, for different research settings in a meta-analysis and with publication bias favouring only

543   effects in one direction (directional), and irrespective of effect heterogeneity and the number of

544   primary studies (K) in the meta-analysis or all significant estimates, the FAT is recommended due

545   to its most consistent false positive rate as well as its superior statistical power in most conditions.

546   Secondly, the TES should be preferred to the FAT if *p-hacking* is suspected. The application of

547   the FAT and PU is limited in situations where the direction of publication bias is defined. If

548   publication bias focusses not on positive or negative results but on both, the FAT loses its

549   diagnostic value completely.

550   Therefore, thirdly, the TES is recommended under effect homogeneity if publication bias is

551   suspected to be non-directional. Fourthly, in the case of heterogeneous effect sizes and a sufficient

552   number of observations in the meta-analysis the 5% caliper provides the best trade-off between a

553   conservative false positive rate and a decent statistical power. This test is therefore best used to

554   identify publication bias in an effect heterogeneous discipline-wide setting which relies per

555   definition on completely different underlying effects but offers enough studies to compensate for

556   the low statistical power. Because the wider 10% and 15% CTs yield inflated false positive rates,

557   at least in some conditions, they are not recommended to identify publication bias.

558   Identifying publication bias in substantial meta-analyses as well as focussing on publication as a

559   general problem within the scientific domain is necessary in order to establish and retain trust in

560   scientific results. Further research, however, should not only focus on the diagnosis of publication

561   bias, but also examine the risk factors, either on the side of the authors or with regard to the

562   incentive structure within the discipline (see for example Auspurg et al. 2014).

563   ## Acknowledgements

564   I thank Katrin Auspurg for her valuable comments on drafts of this article.

565   ## References

566   Alinaghi N, and Reed WR. 2016. Meta-Analysis and Publication Bias: How Well Does the FAT-
567   PET-PEESE Procedure Work? https://ideas.repec.org/p/cbt/econwp/16-26.html.
568   Auspurg K, and Hinz T. 2011. What Fuels Publication Bias? Theoretical and Empirical Analyses
569   of Risk Factors Using the Caliper Test. *Journal of Economics and Statistics* 231:636-660.
570   Auspurg K, Hinz T, and Schneck A. 2014. Ausmaß und Risikofaktoren des Publication Bias in
571   der deutschen Soziologie.
572   Baerlocher MO, O'Brien J, Newton M, Gautam T, and Noble J. 2010. Data Integrity, Reliability
573   and Fraud in Medical Research. *European Journal of Internal Medicine* 21:40-45.
574   doi:10.1016/j.ejim.2009.11.002
575   Banks GC, Kepes S, and McDaniel MA. 2012. Publication Bias: A Call for Improved Meta-
576   Analytic Practice in the Organizational Sciences. *International Journal of Selection and*
577   *Assessment* 20:182-196. doi:10.1111/j.1468-2389.2012.00591.x
578   Begg CB, and Mazumdar M. 1994. Operating Characteristics of a Bank Correlation Test for
579   Publication Bias. *Biometrics* 50:1088-1101. doi:10.2307/2533446
580   Berning CC, and Weiß B. 2015. Publication Bias in the German Social Sciences: an Application
581   of the Caliper Test to Three Top-Tier German Social Science Journals. *Quality &*
582   *Quantity* 50:901-917. doi:10.1007/s11135-015-0182-4
583   Blázquez D, Botella J, and Suero M. 2017. The Debate on the Ego-Depletion Effect: Evidence
584   from Meta-Analysis with the p-Uniform Method. *Frontiers in Psychology* 8:197.
585   doi:10.3389/fpsyg.2017.00197
586   Bosco FA, Aguinis H, Singh K, Field JG, and Pierce CA. 2015. Correlational Effect Size
587   Benchmarks. *Journal of Applied Psychology* 100:431-449. doi:10.1037/a0038047
588   Bürkner P-C, and Doebler P. 2014. Testing for publication bias in diagnostic meta-analysis: a
589   simulation study. *Statistics in Medicine* 33:3061-3077. doi:10.1002/sim.6177
590   Callaham ML, Wears RL, Weber EJ, Barton C, and Young G. 1998. Positive-Outcome Bias and
591   Other Limitations in the Outcome of Research Abstracts Submitted to a Scientific
592   Meeting. *JAMA* 280:254-257.

593   Carsey TM, and Harden JJ. 2013. *Monte Carlo Simulation and Resampling Methods for Social*
594   *Science*: Sage Publications.
595   Chalmers l. 1990. Underreporting Research is Scientific Misconduct. *JAMA* 263:1405-1408.
596   doi:10.1001/jama.1990.03440100121018
597   Cohen J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, N.J.: L.
598   Erlbaum Associates.
599   Cohen J. 1994. The Earth Is Round (p<.05). *American Psychologist* 49:997-1003.
600   doi:10.1037/0003-066x.49.12.997
601   Cole SR, Platt RW, Schisterman EF, Chu H, Westreich D, Richardson D, and Poole C. 2009.
602   Illustrating Bias Due to Conditioning on a collider. *International Journal of*
603   *Epidemiology* 39:417-420. doi:10.1093/ije/dyp334
604   Coursol A, and Wagner EE. 1986. Effect of Positive Findings on Submission and Acceptance: A
605   Note of Meta-Analysis Bias. *Prof Psychol Res Parctice* 17. doi:10.1037/0735-
606   7028.17.2.136
607   Deeks JJ, Higgins JPT, and Altman DG. 2008. Analysing data and undertaking meta-analyses.
608   *Cochrane handbook for systematic reviews of interventions*: John Wiley & Sons, Ltd,
609   243-296.
610   Deeks JJ, Macaskill P, and Irwig L. 2005. The Performance of Tests of Publication Bias and
611   Other Sample Size Effects in Systematic Reviews of Diagnostic Test Accuracy Was
612   Assessed. *Journal of Clinical Epidemiology* 58:882-893.
613   doi:10.1016/j.jclinepi.2005.01.016
614   Descartes R. 2006. *A Discourse on the Method of Correctly Conducting One's Reason and*
615   *Seeking Truth in the Sciences*. Oxford ; New York: Oxford University Press.
616   Dickersin K, and Min Y-I. 1993. Publication Bias - the Problem that Won't Go Away. *Annals of*
617   *the New York Academy of Sciences* 703:135-148. doi:10.1111/j.1749-
618   6632.1993.tb26343.x
619   Dickersin K, Min Y-I, and Meinert CL. 1992. Factors Influencing Publication of Research
620   Results. Follow-up of Applications Submitted to Two Institutional Review Boards. *JAMA*
621   267:374-378. doi:10.1001/jama.267.3.374
622   Doucouliagos H, and Stanley TD. 2009. Publication Selection Bias in Minimum-Wage
623   Research? A Meta-Regression Analysis. *British Journal of Industrial Relations* 47:406-
624   428. doi:10.1111/j.1467-8543.2009.00723.x
625   Duval S, and Tweedie R. 2000. Trim and Fill: A Simple Funnel-Plot–Based Method of Testing
626   and Adjusting for Publication Bias in Meta-Analysis. *Biometrics* 56:455-463.
627   doi:10.1111/j.0006-341X.2000.00455.x
628   Easterbrook PJ, Berlin JA, Gopalan R, and Matthews DR. 1991. Publication Bias in Clinical
629   Research. *Lancet* 337:867-872.
630   Egger M, Smith GD, Schneider M, and Minder C. 1997. Bias in Meta-Analysis Detected by a
631   Simple, Graphical Test. *British Medical Journal* 315:629-634.
632   Epstein WM. 1990. Confirmational Response Bias among Social-Work Journals. *Science*
633   *Technology & Human Values* 15:9-38. doi:10.1177/016224399001500102
634   Epstein WM. 2004. Confirmational Response Bias and the Quality of the Editorial Processes
635   among American Social Work Journals. *Research on Social Work Practice* 14:450-458.
636   doi:10.1177/1049731504265838
637   Feigenbaum S, and Levy DM. 1996. The Technological Obsolescence of Scientific Fraud.
638   *Rationality and Society* 8:261-276. doi:10.1177/104346396008003002

639   Ferguson CJ, and Brannick MT. 2012. Publication Bias in Psychological Science: Prevalence,
640   Methods for Identifying and Controlling, and Implications for the Use of Meta-Analyses.
641   *Psychological Methods* 17:120-128. doi:10.1037/a0024445
642   Ferguson CJ, and Heene M. 2012. A Vast Graveyard of Undead Theories: Publication Bias and
643   Psychological Science's Aversion to the Null. *Perspectives on Psychological Science*
644   7:555-561. doi:10.1177/1745691612459059
645   Francis G. 2012a. Evidence That Publication Bias Contaminated Studies Relating Social Class
646   and Unethical Behavior. *Proceedings of the National Academy of Sciences of the United*
647   *States of America* 109:E1587-E1587. doi:10.1073/pnas.1203591109
648   Francis G. 2012b. The Psychology of Replication and Replication in Psychology. *Perspectives*
649   *on Psychological Science* 7:585-594. doi:10.1177/1745691612459520
650   Francis G. 2012c. Publication Bias and the Failure of Replication in Experimental Psychology.
651   *Psychonomic Bulletin & Review* 19:975-991. doi:10.3758/s13423-012-0322-y
652   Francis G. 2012d. The Same Old New Look: Publication Bias in a Study of Wishful Seeing. *i-*
653   *Perception* 3:176-178. doi:10.1068/i0519ic
654   Francis G. 2012e. Too Good to Be True: Publication Bias in Two Prominent Studies from
655   Experimental Psychology. *Psychonomic Bulletin & Review* 19:151-156.
656   doi:10.3758/s13423-012-0227-9
657   Francis G. 2013. Publication Bias in 'Red, Rank, and Romance in Women Viewing Men,' by
658   Elliot et al. (2010). *Journal of Experimental Psychology: General* 142:292-296.
659   doi:10.1037/a0027923
660   Franco A, Malhotra N, and Simonovits G. 2014. Publication Bias in the Social Sciences:
661   Unlocking the File Drawer. *Science*. doi:10.1126/science.1255484
662   Geertsema P. 2014. Stata, Fast and Slow: Why Running Many Small Regressions in a Large
663   Dataset Takes so Long; and What to Do about It. *Available at SSRN 2423171*.
664   Gelman A. 2013. Too Good to Be True. *Available at*
665   *http://www.slate.com/articles/health_and_science/science/2013/07/statistics_and_psycho*
666   *logy_multiple_comparisons_give_spurious_results.html* (accessed 6.1.2017).
667   Gerber AS, and Malhotra N. 2008a. Do Statistical Reporting Standards Affect What Is
668   Published? Publication Bias in Two Leading Political Science Journals. *Quarterly*
669   *Journal of Political Science* 3:313-326. doi:10.1561/100.00008024
670   Gerber AS, and Malhotra N. 2008b. Publication Bias in Empirical Sociological Research.
671   *Sociological Methods & Research* 37:3-30. doi:10.1177/0049124108318973
672   Godlee F. 2012. Research Misconduct is Widespread and Harms Patients. *BMJ* 344:e14.
673   doi:10.1136/bmj.e14
674   Greenland S, Pearl J, and Robins JM. 1999. Causal Diagrams for Epidemiologic Research.
675   *Epidemiology* 10:37-48. doi:10.1097/00001648-199901000-00008
676   Hartgerink CHJ, van Aert RCM, Nuijten MB, Wicherts JM, and van Assen MALM. 2016.
677   Distributions of p-values smaller than .05 in psychology: what is going on? *Peerj*
678   4:e1935. doi:10.7717/peerj.1935
679   Hayashino Y, Noguchi Y, and Fukui T. 2005. Systematic Evaluation and Comparison of
680   Statistical Tests for Publication Bias. *Journal of Epidemiology* 15:235-243.
681   doi:10.2188/jea.15.235
682   Head ML, Holman L, Lanfear R, Kahn AT, and Jennions MD. 2015. The Extent and
683   Consequences of P-Hacking in Science. *PLoS Biol* 13:e1002106.
684   doi:10.1371/journal.pbio.1002106

Higgins JPT, and Green S. 2008. *Cochrane Handbook for Systematic Reviews of Interventions*: Wiley Online Library.

Higgins JPT, and Thompson SG. 2002. Quantifying Heterogeneity in a Meta-Analysis. *Statistics in Medicine* 21:1539-1558. doi:10.1002/sim.1186

Hoenig JM, and Heisey DM. 2001. The Abuse of Power. *The American Statistician* 55:19-24. doi:10.1198/000313001300339897

Hua F, Walsh T, Glenny A-M, and Worthington H. 2016. Thirty Percent of Abstracts Presented at Dental Conferences Are Published in Full: A Systematic Review. *Journal of Clinical Epidemiology* 75:16-28. doi:10.1016/j.jclinepi.2016.01.029

Ioannidis JP. 1998. Effect of the Statistical Significance of Results on the Time to Completion and Publication of Randomized Efficacy Trials. *JAMA* 279:281-286. doi:10.1001/jama.279.4.281

Ioannidis JPA, Stanley TD, and Doucouliagos H. 2016. The Power of Bias in Economics Research Economic Series. Deakin University.

Ioannidis JPA, and Trikalinos TA. 2007a. Comments on 'An exploratory Test for an Excess of Significant Findings' by JPA Ioannidis and TA Trikalinos - Authors' Response to V Johnson and Y Yuan. *Clinical Trials* 4:256-257. doi:10.1177/1740774507079433

Ioannidis JPA, and Trikalinos TA. 2007b. An Exploratory Test for an Excess of Significant Findings. *Clinical Trials* 4:245-253. doi:10.1177/1740774507079441

Jefferson T, Jones MA, Doshi P, Del Mar CB, Heneghan CJ, Hama R, and Thompson MJ. 2012. Neuraminidase Inhibitors for Preventing and Treating Influenza in Healthy Adults. *Cochrane database of systematic reviews*. doi:10.1002/14651858.CD008965.pub3

John LK, Loewenstein G, and Prelec D. 2012. Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science* 23:524-532. doi:10.1177/0956797611430953

Johnson V, and Yuan Y. 2007. Comments on `An Exploratory Test for an Excess of Significant Findings' by JPA loannidis and TA Trikalinos. *Clinical Trials* 4:254-255. doi:10.1177/1740774507079437

Kicinski M. 2014. How Does Under-Reporting of Negative and Inconclusive Results Affect the False-Positive Rate in Meta-Analysis? A Simulation Study. *BMJ open* 4:1-8. doi:10.1136/bmjopen-2014-004831

Kreuter F, Presser S, and Tourangeau R. 2008. Social Desirability Bias in CATI, IVR, and Web Surveys: The Effects of Mode and Question Sensitivity. *Public Opinion Quarterly* 72:847-865. doi:10.1093/poq/nfn063

Kühberger A, Fritz A, and Scherndl T. 2014. Publication Bias in Psychology: A Diagnosis Based on the Correlation between Effect Size and Sample Size. *PloS one* 9:e105825. doi:10.1371/journal.pone.0105825

Labovitz S. 1972. Statistical Usage In Sociology: Sacred Cows and Ritual. *Sociological Methods & Research* 1:13-37. doi:10.1177/004912417200100102

Lakens D. 2015. What R-hacking Really Looks Like: A Comment on Masicampo and LaLande (2012). *The Quarterly Journal of Experimental Psychology* 68:829-832. doi:10.1080/17470218.2014.982664

Lee KP, Boyd EA, Holroyd-Leduc JM, Bacchetti P, and Bero LA. 2006. Predictors of Publication: Characteristics of Submitted Manuscripts Associated with Acceptance at Major Biomedical Journals. *Medical Journal of Australia* 184:621-626.

730 Leggett NC, Thomas NA, Loetscher T, and Nicholls MER. 2013. The Life of P: "Just
731   Significant" Results Are on the Rise. *Quarterly Journal of Experimental Psychology*
732   66:2303-2309. doi:10.1080/17470218.2013.863371
733 Light RJ, and Pillemer DB. 1984. *Summing Up: The Science of Reviewing Research*. Cambridge,
734   Mass.: Harvard University Press.
735 Macaskill P, Walter SD, and Irwig L. 2001. A Comparison of Methods to Detect Publication
736   Bias in Meta-Analysis. *Statistics in Medicine* 20:641-654. doi:10.1002/sim.698
737 Mahoney MJ. 1977. Publication Prejudices: An Experimental Study of Confirmatory Bias in the
738   Peer Review System. *Cognitive Therapy and Research* 1:161-175.
739   doi:10.1007/BF01173636
740 Masicampo EJ, and Lalande DR. 2012. A Peculiar Prevalence of P Values just Below .05.
741   *Quarterly Journal of Experimental Psychology* 65:2271-2279.
742   doi:10.1080/17470218.2012.711335
743 McShane BB, Bockenholt U, and Hansen KT. 2016. Adjusting for Publication Bias in Meta-
744   Analysis: An Evaluation of Selection Methods and Some Cautionary Notes. *Perspectives*
745   *on Psychological Science* 11:730-749. doi:10.1177/1745691616662243
746 Merton RK. 1957. Priorities in Scientific Discovery: A Chapter in the Sociology of Science.
747   *American sociological review* 22:635-659. doi:10.2307/2089193
748 Mooney CZ. 1997. *Monte Carlo Simulation*. Thousand Oakes: Sage Publications.
749 Moreno SG, Sutton AJ, Ades AE, Stanley TD, Abrams KR, Peters JL, and Cooper NJ. 2009.
750   Assessment of Regression-based Methods to Adjust for Publication Bias through a
751   Comprehensive Simulation Study. *Bmc Medical Research Methodology* 9.
752   doi:10.1186/1471-2288-9-2
753 Nature. 2017. Getting published in Nature: The editorial process. *Available at*
754   *http://www.nature.com/nature/authors/get_published/* (accessed 05.01.).
755 Nuijten MB, Hartgerink CHJ, van Assen MALM, Epskamp S, and Wicherts JM. 2016. The
756   Prevalence of Statistical Reporting Errors in Psychology (1985–2013). *Behavior*
757   *Research Methods* 48:1205-1226. doi:10.3758/s13428-015-0664-2
758 Nuzzo R. 2014. Scientific Method: Statistical Errors. *Nature* 506:150-152. doi:10.1038/506150a
759 Olson CM, Rennie D, Cook D, Dickersin K, Flanagin A, Hogan JW, Zhu Q, Reiling J, and Pace
760   B. 2002. Publication Bias in Editorial Decision Making. *Jama-Journal of the American*
761   *Medical Association* 287:2825-2828. doi:10.1001/jama.287.21.2825
762 Paldam M. 2013. Regression Costs Fall, Mining Ratios Rise, Publication Bias Looms, and
763   Techniques Get Fancier: Reflections on Some Trends in Empirical Macroeconomics.
764   *Econ Journal Watch* 10:136-156.
765 Paldam M. 2015. Simulating an Empirical Paper by the Rational Economist. *Empirical*
766   *Economics*:1-25. doi:10.1007/s00181-015-0971-6
767 Popper KR. 1968. *Conjectures and Refutations: The Growth of Scientific Knowledge*. New York:
768   Harper & Row.
769 Reed WR. 2015. A Monte Carlo Analysis of Alternative Meta-Analysis Estimators in the
770   Presence of Publication Bias. *Economics-the Open Access Open-Assessment E-Journal* 9.
771   doi:10.5018/economics-ejournal.ja.2015-30
772 Renkewitz F, and Keiner M. 2016. How to Detect Publication Biases from Published Data? A
773   Monte Carlo Simulation of Different Methods. 50 Kongress der Deutschen Gesellschaft
774   für Psychologie. Leipzig.

775 Rosenthal R. 1979. The File Drawer Problem and Tolerance for Null Results. *Psychological*
776   *Bulletin* 86:638-641. doi:10.1037/0033-2909.86.3.638
777 Schimmack U. 2012. The Ironic Effect of Significant Results on the Credibility of Multiple-
778   Study Articles. *Psychological Methods* 17:551-566. doi:10.1037/a0029487
779 Schoenfeld JD, and Ioannidis JPA. 2013. Is Everything We Eat Associated with Cancer? A
780   Systematic Cookbook Review. *American Journal of Clinical Nutrition* 97:127-134.
781   doi:10.3945/ajcn.112.047142
782 Science. 2017. The Science Contributors FAQ. *Available at*
783   *http://www.sciencemag.org/site/feature/contribinfo/faq/#pct_faq* (accessed 05.01.).
784 Simmons JP, Nelson LD, and Simonsohn U. 2011. False-Positive Psychology: Undisclosed
785   Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant.
786   *Psychological Science* 22:1359-1366. doi:10.1177/0956797611417632
787 Simmons JP, and Simonsohn U. 2017. Power Posing. *Psychological*
788   *Science*:0956797616658563. doi:10.1177/09567976166658563
789 Simonsohn U, Nelson LD, and Simmons JP. 2014a. P-Curve and Effect Size: Correcting for
790   Publication Bias Using Only Significant Results. *Perspectives on Psychological Science*
791   9:666-681. doi:10.1177/1745691614553988
792 Simonsohn U, Nelson LD, and Simmons JP. 2014b. P-curve: A Key to the File Drawer. *Journal*
793   *of Experimental Psychology-General* 143:534-547. doi:10.1037/a0033242
794 Simonsohn U, Simmons JP, and Nelson LD. 2015. Better P-Curves: Making P-Curve Analysis
795   More Robust To Errors, Fraud, and Ambitious P-Hacking, A Reply To Ulrich and Miller
796   (2015). *Journal of Experimental Psychology-General* 144:1146-1152.
797   doi:10.1037/xge0000104
798 Stanley TD. 2017. Limitations of PET-PEESE and Other Meta-Analysis Methods. *Social*
799   *Psychological and Personality Science*:1948550617693062.
800   doi:10.1177/1948550617693062
801 Stanley TD, and Doucouliagos H. 2014. Meta-Regression Approximations to Reduce Publication
802   Selection Bias. *Research Synthesis Methods* 5:60-78. doi:10.1002/jrsm.1095
803 Sterne JAC, Gavaghan D, and Egger M. 2000. Publication and Related Bias in Meta-Analysis:
804   Power of Statistical Tests and Prevalence in the Literature. *Journal of Clinical*
805   *Epidemiology* 53:1119-1129. doi:10.1016/S0895-4356(00)00242-0
806 Stroebe W, Postmes T, and Spears R. 2012. Scientific Misconduct and the Myth of Self-
807   Correction in Science. *Perspectives on Psychological Science* 7:670-688.
808   doi:10.1177/1745691612460687
809 Tang JL, and Liu JLY. 2000. Misleading Funnel Plot for Detection of Bias in Meta-Analysis.
810   *Journal of Clinical Epidemiology* 53:477-484. doi:10.1016/s0895-4356(99)00204-8
811 Thaler R. 1980. Toward a Positive Theory of Consumer Choice. *Journal of Economic Behavior*
812   *& Organization* 1:39-60. doi:10.1016/0167-2681(80)90051-7
813 Timmer A, Hilsden RJ, Cole J, Hailey D, and Sutherland LR. 2002. Publication Bias in
814   Gastroenterological Research – A Retrospective Cohort Study Based on Abstracts
815   Submitted to a Scientific Meeting. *Bmc Medical Research Methodology* 2:7.
816   doi:10.1186/1471-2288-2-7
817 Ulrich R, and Miller J. 2017. Some Properties of p-Curves, With an Application to Gradual
818   Publication Bias. *Psychological Methods*. doi:10.1037/met0000125

819   van Aert RCM, Wicherts JM, and van Assen MALM. 2016. Conducting Meta-Analyses Based
820   on p Values. *Perspectives on Psychological Science* 11:713-729.
821   doi:10.1177/1745691616650874
822   van Assen MALM, van Aert RCM, and Wicherts JM. 2015. Meta-Analysis Using Effect Size
823   Distributions of only Statistically Significant Studies. *Psychological Methods* 20:293-
824   309. doi:10.1037/met0000025 10.1037/met0000025.supp (Supplemental)
825   Wasserstein RL, and Lazar NA. 2016. The ASA's Statement on p-Values: Context, Process, and
826   Purpose. *American Statistician* 70:129-131. doi:10.1080/00031305.2016.1154108
827   Yoder S, and Bramlett BH. 2011. What Happens at the Journal Office Stays at the Journal
828   Office: Assessing Journal Transparency and Record-Keeping Practices. *Ps-Political*
829   *Science & Politics* 44:363-373. doi:10.1017/S1049096511000217
830

831

Table 1(on next page)

Truth table

Estimator	Data	
	No effect (false)	Effect (true)
No effect detected	**True negative (1-α)**	False negative (β)
Effect detected	False positive (α)	**True positive (1-β)**

1

Figure 1

Funnel plot
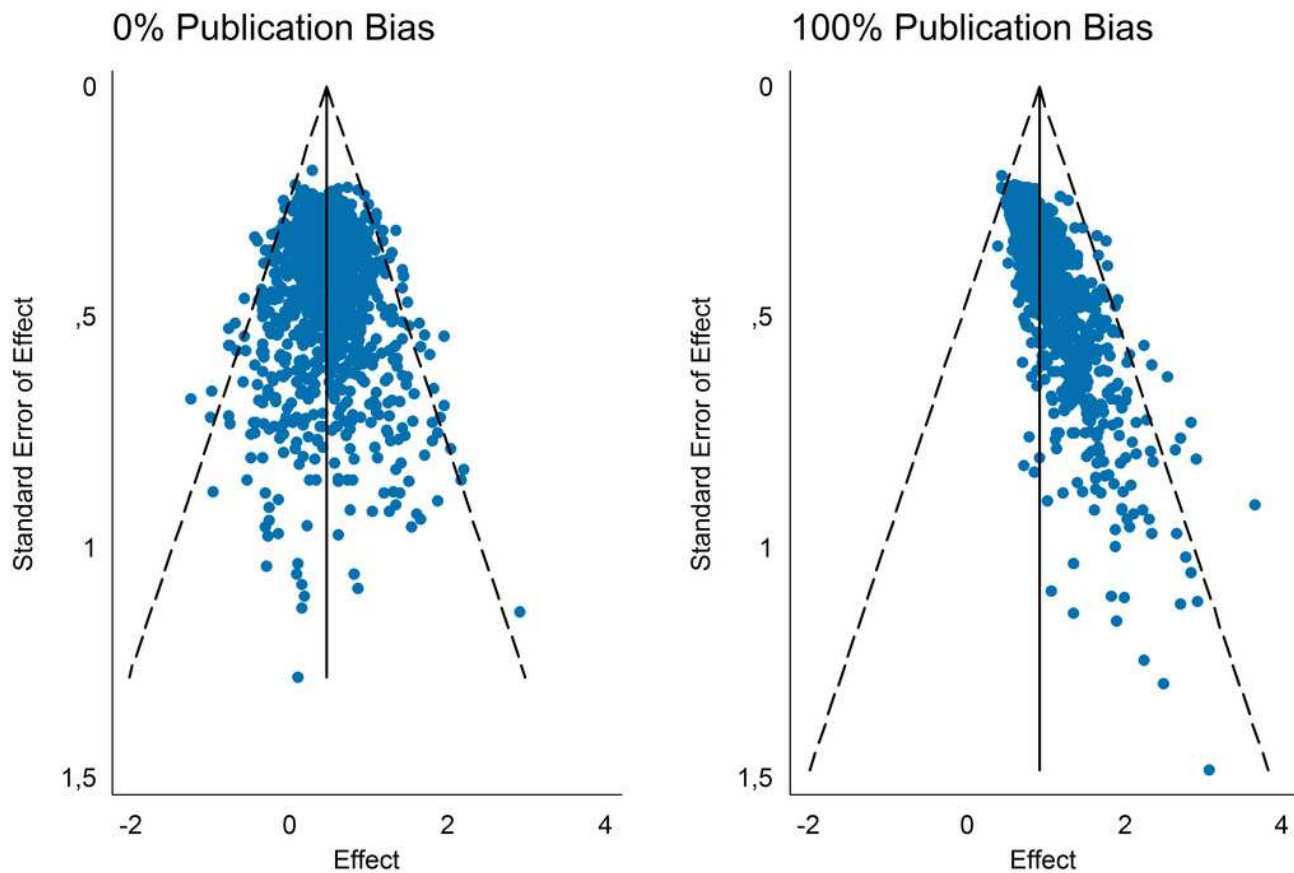
es=0.5, k=1000, n=100, file-drawer

Figure 2

Funnel asymmetry test (FAT)
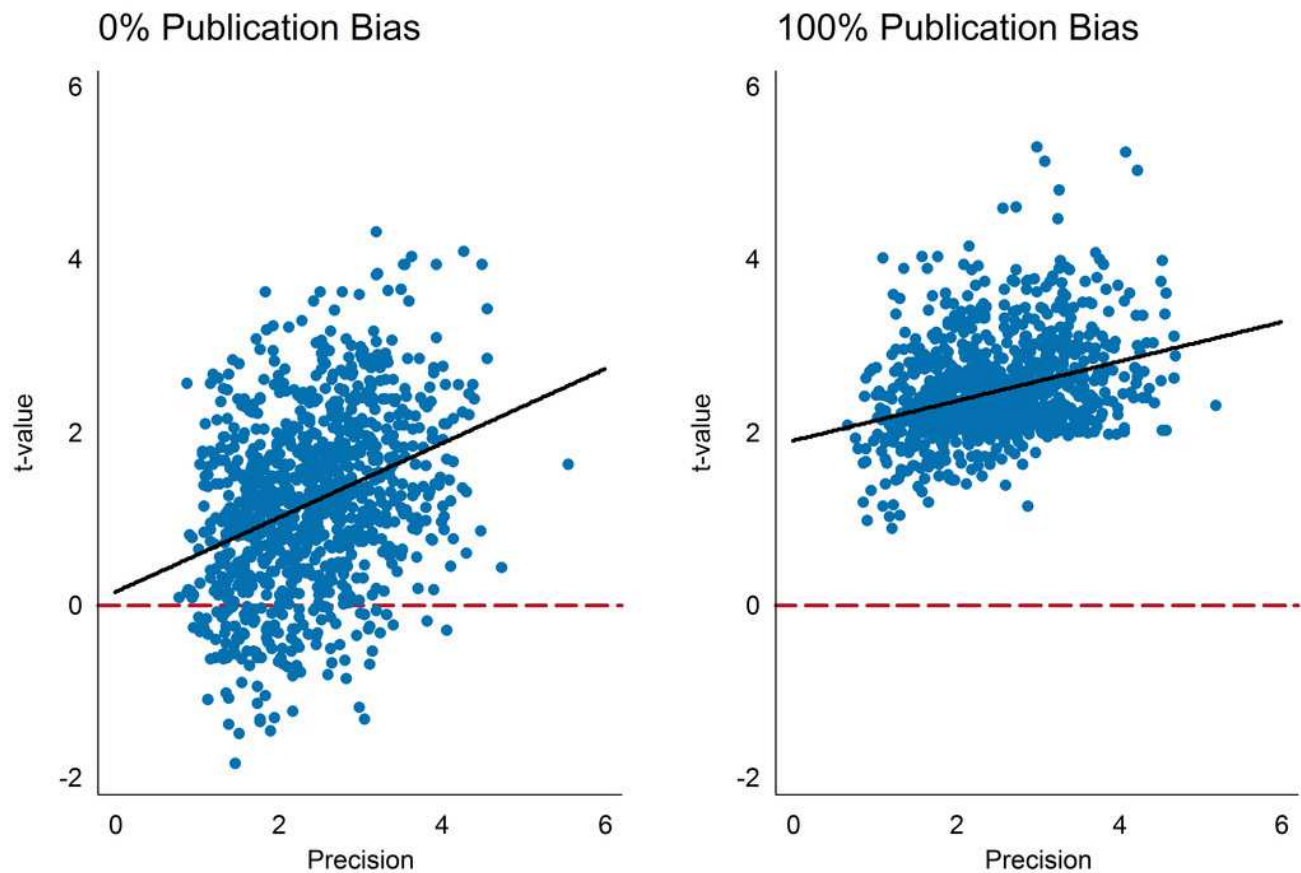

es=0.5, k=1000, n=100, file-drawer

Figure 3

p-uniform (PU)
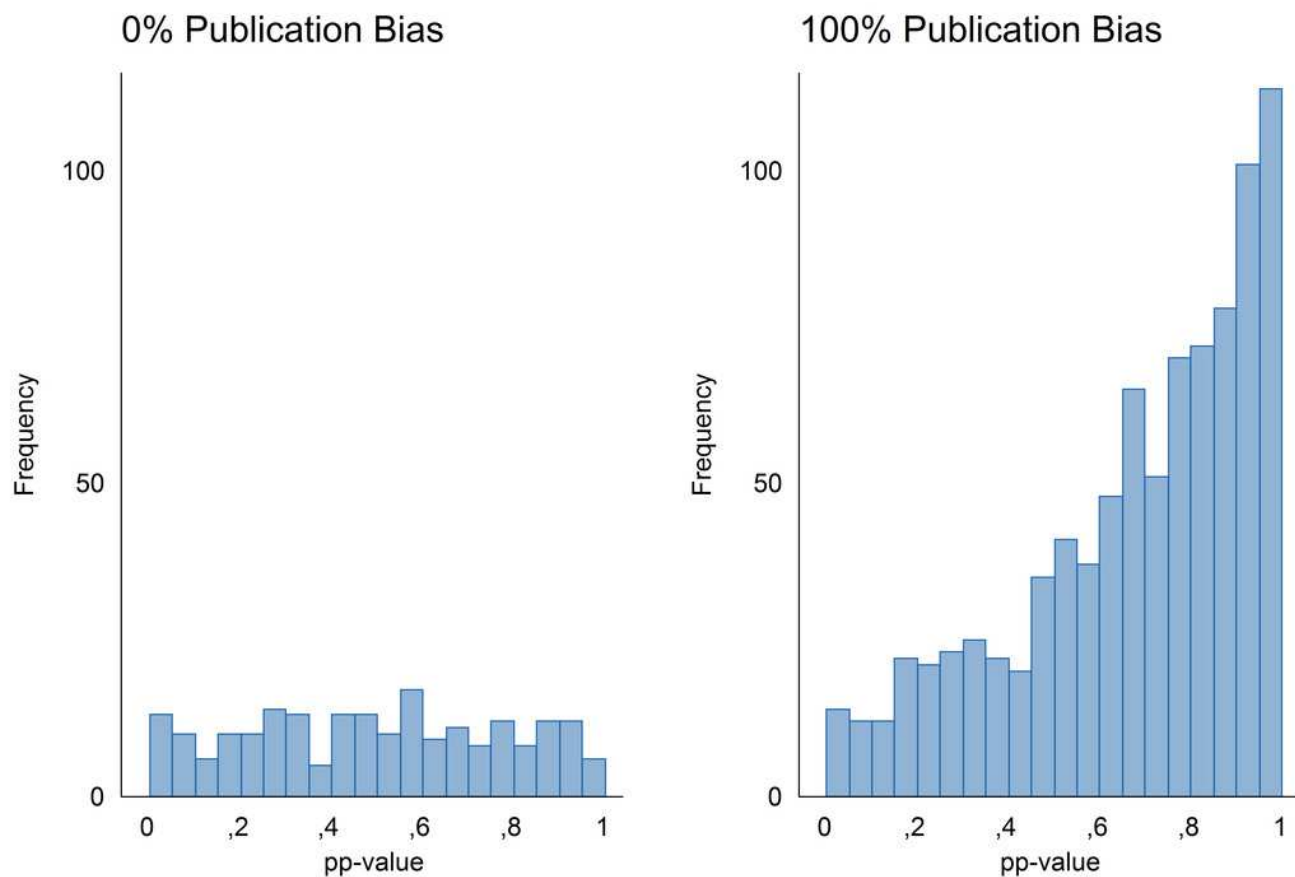
es=0.5, k=1000, n=100, file-drawer

Figure 4

Caliper test (CT with 5% caliper)
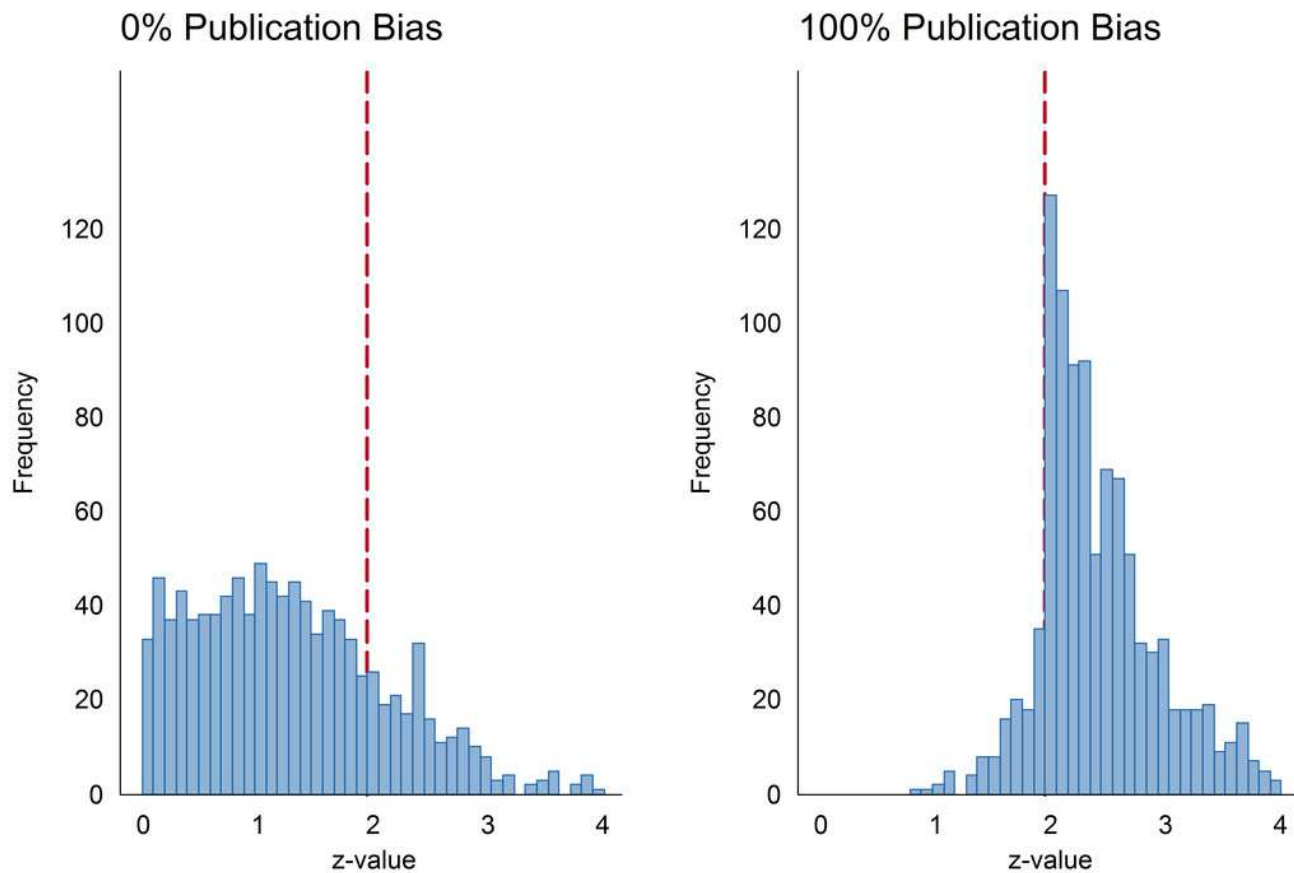
es=0.5, k=1000, n=100, file-drawer

Table 2(on next page)

Publication bias tests in comparison

Test	Measurement level	Sample	Assumption	Limitation	Test method
FAT	Continuous [-∞,∞]	All	Cov(es, se) = 0	Only directed publication bias (PB) detectable	Weighted-Least-Squares
PU	Continuous [0,1]	p < 0.05, effects of same sign	Uniform or right skewed Skewness >= 0	Only directed PB detectable Only levelled (e.g. p=0.05) PB testable Effect homogeneity (FE-MA)	Skewness test (Gamma)
TES	Dichotomous [0,1]	All	E = O	Only levelled (e.g. p=0.05) PB testable Effect homogeneity (FE-MA)	Binomial test
CT	Dichotomous [0,1]	Threshold ± caliper width	P(UC) = P(OC)	Only levelled (e.g. p=0.05) PB testable	Binomial test

1

Table 3(on next page)

DGP of Monte Carlo simulation

Conditions	Values	Functional form	N (conditions)
Data setup:			
1. True effects β:	$\boldsymbol{\beta} = \mathbf{0;0.5;1.0;1.5;}\boldsymbol{Het}$		**5**
2. Number of observations N:	$\boldsymbol{\mu_N} = \mathbf{100;500}$	$\left\lvert NV(\mu_N,150)\right\rvert\ N > 30$	**2**
3. Number of studies K:	$\boldsymbol{K} = \mathbf{100;1000}$		**2**
Behavioural setup:			
4. Publication bias:	$\boldsymbol{PB} = \mathbf{0;0.5;1}$	$\beta > 0\ \&\ p < 0.05$ Take best result after maximum runs	**1+2*2=5**
4a. File-drawer	Draw new sample size N	*(maximum 9 additional samples)*	
4b. p-hacking	Run new analyses with same dataset	$y = \beta x + \gamma_j z_j + \varepsilon$ $z = 0.5x + 0.5y + \varepsilon$ max. 3 z's = 7 combinations	
			5*2*2*5 = 100

1

Table 4(on next page)

EM of Monte Carlo simulation

Implemented tests:
- Funnel asymmetry-test (FAT)
- p-uniform (PU)
- Test of excess significance (TES)
- Caliper Test (CT)
  3%-, 5%-, 10%-, 15%-caliper
➢ Seven tests

Outcomes:

- Statistical power (1-β))

- False positive rate (α)

$s = 0$ if p $\geq$ 0.05
$s = 1$ if p $<$ 0.05

$$\sum_{i=1}^{r} s_i/r \ if \ PB > 0$$

$$\sum_{i=1}^{r} s_i/r \ if \ PB = 0$$

➢ 100 different conditions for seven tests = 700 power/error estimates

Run Monte Carlo design:
- Power calculations in order to identify deviations from expected α and the statistical power for each element of the experimental design matrix (two-sided z-test of proportion)
- 3729 runs for each 20 α-error estimates (80% power, expected 0.05)
- 1545 runs for each 80 power estimates (80% power, expected 0.8)
➢ 198,080 meta-analyses containing 108,900,000 primary studies

1

Table 5(on next page)

Descriptive results

		Runs	Mean	Median	Minimum	Maximum
Publication bias committed						
	50%	61,760	0.217	0.19	0	0.66
	100%	61,760	0.434	0.387	0	1
Publication bias successful						
file-drawer	50%	30.880	0.125	0.105	0	0.48
	100%	30.880	0.250	0.209	0	0.83
p-hacking	50%	30.880	0.125	0.112	0	0.49
	100%	30.880	0.251	0.225	0	0.84
Heterogeneity (I^2)						
	0%	158,464	0.043	0	0	0.542
	100%	39,616	0.733	0.766	0.066	0.908

1

Table 6(on next page)

False positive rate

0% FD/PH	PU	FAT	TES	3% CT	5% CT	10% CT	15% CT
N100/K100							
0.0	0.045	0.043	0.024	0.001	0.001	0.003	0.002
0.5	0.039	0.045	0.004	0.004	0.011	0.013	0.012
1.0	0.014	0.056	0.005	0.005	0.017	0.033	0.040
1.5	0.001	0.047	0.010	0.000	0.005	0.026	0.041
Het	0.000	0.042	0.001	0.002	0.012	0.021	0.025
N100/K1000							
0.0	0.032	0.051	0.036	0.020	0.012	0.000	0.000
0.5	0.020	0.046	0.005	0.031	0.023	0.007	0.001
1.0	0.008	0.048	0.003	0.043	0.049	**0.067**	**0.092**
1.5	0.002	0.046	0.013	0.040	0.049	**0.101**	**0.204**
Het	0.000	0.047	0.000	0.032	0.032	0.028	0.030
N500/K100							
0.0	0.051	0.051	0.024	0.000	0.002	0.003	0.002
0.5	0.025	0.050	0.002	0.010	0.019	0.039	0.043
1.0	0.000	0.045	0.007	0.000	0.000	0.001	0.010
1.5	0.000	0.047	0.000	0.000	0.000	0.000	0.000
Het	0.000	0.037	0.000	0.000	0.002	0.011	0.018
N500/K1000							
0.0	0.043	0.052	0.037	0.019	0.009	0.001	0.000
0.5	0.024	0.042	0.004	0.039	0.045	**0.070**	**0.104**
1.0	0.000	0.054	0.033	0.018	0.043	**0.108**	**0.244**
1.5	0.000	0.048	0.003	0.000	0.000	0.002	0.007
Het	0.000	0.035	0.000	0.031	0.036	0.038	0.035

Bold numbers > 0.05 at $p < 0.05$

1

Table 7(on next page)

Statistical power for 50% file-drawer

50% FD	PU	FAT	TES	3% CT	5% CT	10% CT	15% CT
N100/K100							
0.0	0.179	_0.662_	0.148	0.007	0.013	0.015	0.005
0.5	0.691	**0.822**	**_0.912_**	0.108	0.220	0.416	0.563
1.0	0.348	**0.823**	**_0.881_**	0.052	0.149	0.415	0.594
1.5	0.034	0.457	_0.537_	0.007	0.032	0.164	0.285
Het	0.000	0.370	0.042	0.032	0.082	0.220	0.321
N100/K1000							
0.0	0.720	**_1.000_**	0.737	0.029	0.019	0.001	0.000
0.5	**_1.000_**	**_1.000_**	**_1.000_**	**0.894**	**0.981**	**_1.000_**	**_1.000_**
1.0	**_1.000_**	**_1.000_**	**_1.000_**	**0.859**	**0.976**	0.999	1.000
1.5	0.521	**_0.999_**	**_1.000_**	0.530	0.763	0.981	0.999
Het	0.000	**_0.997_**	0.125	0.639	**0.839**	**0.977**	**0.996**
N500/K100							
0.0	0.238	0.245	0.104	0.007	0.010	0.013	0.003
0.5	0.580	0.499	0.736	0.080	0.201	0.442	0.671
1.0	0.001	0.110	0.039	0.000	0.001	0.003	0.021
1.5	0.000	0.056	0.000	0.000	0.000	0.000	0.000
Het	0.000	0.058	0.000	0.005	0.029	0.095	_0.166_
N500/K1000							
0.0	**0.905**	**0.950**	0.544	0.043	0.028	0.001	0.000
0.5	**_1.000_**	**0.999**	**_1.000_**	**0.911**	**0.987**	**_1.000_**	**_1.000_**
1.0	0.004	0.396	**_0.874_**	0.068	0.165	0.529	**0.826**
1.5	0.001	0.064	0.019	0.000	0.000	0.005	0.019
Het	0.000	0.214	0.000	0.373	0.569	**0.855**	**0.950**
Best / Satisfactory	3 / 4	_14_ / 6	8 / _7_	0 / 3	0 / 4	2 / 6	3 / 7

Bold numbers: > 0.8 $p < 0.05$. Underlined: best estimator. Grey shaded: inflated false positive rate, cf. Table 6

1

Table 8(on next page)

Statistical power for 100% file-drawer

100% FD	PU	FAT	TES	3% CT	5% CT	10% CT	15% CT
N100/K100							
0.0	0.756	**1.000**	0.000	0.013	0.016	0.012	0.005
0.5	**1.000**	**1.000**	**1.000**	0.328	0.569	**0.891**	**0.981**
1.0	**0.958**	**1.000**	**1.000**	0.222	0.618	**0.962**	**0.999**
1.5	0.177	**0.975**	**1.000**	0.028	0.138	0.621	**0.898**
Het	0.000	**0.962**	**0.882**	0.124	0.278	0.595	0.790
N100/K1000							
0.0	**1.000**	**1.000**	0.000	0.047	0.021	0.002	0.000
0.5	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
1.0	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
1.5	**0.999**	**1.000**	**1.000**	**0.999**	**1.000**	**1.000**	**1.000**
Het	0.000	**1.000**	**1.000**	**0.981**	**0.999**	**1.000**	**1.000**
N500/K100							
0.0	**0.888**	**0.990**	0.000	0.011	0.012	0.015	0.003
0.5	**1.000**	**0.999**	**1.000**	0.351	0.755	**0.992**	**1.000**
1.0	0.001	0.221	**0.235**	0.000	0.000	0.006	0.047
1.5	0.001	**0.059**	0.000	0.000	0.000	0.000	0.000
Het	0.000	0.129	0.000	0.026	0.092	0.290	**0.473**
N500/K1000							
0.0	**1.000**	**1.000**	0.000	0.039	0.021	0.003	0.000
0.5	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
1.0	0.093	**0.898**	**1.000**	0.233	0.669	**0.995**	**1.000**
1.5	0.000	0.108	**0.145**	0.000	0.000	0.009	0.041
Het	0.000	0.628	0.000	**0.829**	**0.957**	**1.000**	**1.000**
Best / Satisfactory	7 / 10	**13** / **15**	12 / 11	3 / 6	4 / 6	6 / 10	9 / 11

Bold numbers: > 0.8 p < 0.05. Underlined: best estimator. Grey shaded: inflated false positive rate, cf. Table 6

1

Table 9(on next page)

Statistical power for 50% p-hacking

50% PH	PU	FAT	TES	3% CT	5% CT	10% CT	15% CT
N100/K100							
0.0	<u>0.764</u>	0.006	0.598	0.077	0.139	0.196	0.166
0.5	**<u>0.870</u>**	0.371	0.490	0.152	0.288	0.465	0.527
1.0	0.321	0.395	0.422	0.079	0.168	0.396	<u>0.528</u>
1.5	0.018	0.129	0.166	0.015	0.058	0.154	<u>0.259</u>
Het	0.000	0.369	0.020	0.068	0.139	0.292	<u>0.380</u>
N100/K1000							
0.0	**<u>1.000</u>**	0.000	**<u>1.000</u>**	0.767	**0.874**	**0.929**	**0.846**
0.5	**<u>1.000</u>**	**0.992**	**<u>1.000</u>**	**0.973**	**0.995**	**<u>1.000</u>**	**<u>1.000</u>**
1.0	**0.997**	**0.997**	**<u>1.000</u>**	**0.879**	**0.968**	**0.999**	**<u>1.000</u>**
1.5	0.175	0.503	**0.962**	0.505	0.733	**0.958**	**<u>0.994</u>**
Het	0.000	**0.994**	0.007	0.797	**0.942**	**0.997**	**<u>0.999</u>**
N500/K100							
0.0	**0.958**	0.000	**<u>1.000</u>**	0.211	0.394	0.659	0.769
0.5	0.806	0.437	**<u>0.843</u>**	0.112	0.285	0.602	0.784
1.0	0.000	<u>0.066</u>	0.028	0.000	0.001	0.013	0.046
1.5	0.001	<u>0.058</u>	0.000	0.000	0.000	0.000	0.000
Het	0.000	<u>0.408</u>	0.000	0.015	0.067	0.233	0.383
N500/K1000							
0.0	**<u>1.000</u>**	0.000	**<u>1.000</u>**	**0.995**	**<u>1.000</u>**	**<u>1.000</u>**	**<u>1.000</u>**
0.5	**<u>1.000</u>**	**0.999**	**<u>1.000</u>**	**0.966**	**0.999**	**<u>1.000</u>**	**<u>1.000</u>**
1.0	0.004	0.159	0.775	0.116	0.271	0.676	**<u>0.908</u>**
1.5	0.000	<u>0.046</u>	0.012	0.002	0.002	0.007	0.026
Het	0.000	**0.997**	0.000	0.772	**0.935**	**0.998**	**<u>1.000</u>**
Best / Satisfactory	6 / 7	4 / 5	7 / <u>8</u>	0 / 4	1 / 7	3 / <u>8</u>	<u>11</u> / <u>8</u>

Bold numbers: > 0.8 p < 0.05. Underlined: best estimator. Grey shaded: inflated false positive rate, cf. Table 6

1

Table 10(on next page)

Statistical power for 100% p-hacking

100% PH	PU	FAT	TES	3% CT	5% CT	10% CT	15% CT
N100/K100							
0.0	**0.999**	0.808	0.212	0.203	0.331	0.477	0.497
0.5	**1.000**	**0.997**	**0.992**	0.481	0.727	**0.918**	**0.964**
1.0	**0.854**	**0.916**	**0.985**	0.286	0.518	**0.835**	**0.947**
1.5	0.089	0.293	0.679	0.051	0.165	0.443	0.648
Het	0.000	**0.903**	0.390	0.235	0.436	0.724	**0.847**
N100/K1000							
0.0	**1.000**	**1.000**	**0.999**	**0.984**	**0.999**	**1.000**	**1.000**
0.5	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
1.0	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
1.5	**0.887**	0.976	**1.000**	0.957	0.997	**1.000**	**1.000**
Het	0.000	**1.000**	**0.999**	**0.997**	**1.000**	**1.000**	**1.000**
N500/K100							
0.0	**1.000**	0.979	**1.000**	0.561	0.791	**0.977**	**0.994**
0.5	**0.999**	**0.997**	**1.000**	0.525	**0.847**	**0.995**	**1.000**
1.0	0.001	0.106	0.138	0.000	0.002	0.036	0.119
1.5	0.000	0.051	0.000	0.000	0.000	0.000	0.000
Het	0.000	**0.916**	0.003	0.099	0.328	0.758	**0.917**
N500/K1000							
0.0	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
0.5	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
1.0	0.028	0.405	**1.000**	0.438	0.804	**0.996**	**1.000**
1.5	0.000	0.061	0.028	0.000	0.002	0.022	0.072
Het	0.000	**1.000**	0.000	**0.998**	**1.000**	**1.000**	**1.000**
Best / Satisfactory	8 / 11	7 / 14	9 / 12	4 / 8	6 / 9	8 / 13	12 / 15

1

Table 11(on next page)

Regression analysis false positive rate

	PU	FAT	TES	3% CT	5% CT	10% CT	15% CT
K = 1.000	-0.005***	0.000	0.006***	0.026***	0.023***	0.026***	0.050***
(ref. K = 100)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
I^2 [+10 percentage points]	-0.003***	-0.001***	-0.001***	0.000	0.000	-0.001***	-0.003***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Constant [a]	0.023***	0.049	0.010***	0.003***	0.008***	0.019***	0.028***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	73,960	74,560	74,560	62,644	66,546	69,718	70,936
R^2	0.005	0.000	0.002	0.010	0.007	0.006	0.017

[a] Test H0: constant = 0.05; Standard errors in parentheses; *** $p<0.01$. ** $p<0.05$. * $p<0.1$

1

Table 12(on next page)

Regression analysis statistical power

	PU	FAT	TES	3% CT	5% CT	10% CT	15% CT
K = 1.000	0.165***	0.244***	0.238***	0.573***	0.513***	0.382***	0.307***
(ref. K = 100)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
I² [+10 percentage points]	-0.065***	0.001**	-0.064***	0.006***	0.005***	0.008***	0.010***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
p-hacking	0.048***	-0.110***	0.075***	0.179***	0.187***	0.186***	0.177***
(ref. file-drawer)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Comitted PB [+10ppts]	0.051***	0.030***	-0.065***	-0.035***	-0.053***	-0.073***	-0.084***
(ref. mean = 32.5%)	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Successful PB [+10ppts]	0.103***	0.099***	0.221***	0.162***	0.193***	0.224***	0.234***
(ref. mean = 18.8%)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Constant [a]	0.483***	0.569***	0.515***	-0.002***	0.125***	0.300***	0.386***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Observations	123,520	123,520	123,520	107,736	111,315	115,243	117,207
R²	0.572	0.306	0.473	0.497	0.483	0.457	0.446

[a] Test H0: constant = 0.05; Standard errors in parentheses; *** $p<0.01$. ** $p<0.05$. * $p<0.1$

1

Table 13(on next page)

Recommended tests for different conditions

	Directional		Non-directional	
	Small K	Large K	Small K	Large K
Homogenous	FAT	FAT / TES	TES	TES
Heterogenous	FAT	FAT	CT	CT

1