Good science, bad science: Questioning research practices in psychological research

Bakker, M.

Appendix B

Appendix to Chapter 6

Selection of published meta-analyses

To gather a representative sample of sets of psychological studies that concern the same phenomenon or at least highly similar phenomena, we retrieved from the PsycARTICLES database all 108 peer-reviewed articles published in 2011 that contained the strings "research synthesis", "systematic review", or "meta-anal*" in the title and/or abstract. We drew a random number between 1 and 108 without replacement for each of the articles and selected the articles with the lowest number. A further 21 articles were also selected but could not be included because the article did not report effect sizes of primary studies (13 meta-analyses), there were fewer than 10 primary studies (3), the review did not revolve around effect sizes (2), the paper was temporarily unavailable (1), or because the cases in the samples were dependent, leading to biased standard errors (2). The final set of eleven meta-analyses (10% of the total) is given in Table B.1.

In several cases we could not exactly replicate the meta-analytic findings (Alfieri, Card, Farber, Green, Woodin), although our results were similar to those presented in the paper. In one meta-analysis (Hallion) we came across a gross error in the computation of an effect size in which a standard error was inadvertently taken as the SD. We corrected that error and used our own computations throughout. The datasets for each meta-analysis are available as an Excel file. In light of the relatively small sample sizes, we used Hedges' g as the effect sizes for mean differences. For correlations we used Fisher-transformed correlations.

Table B.1

Overview of the eleven selected meta-analyses with the number of published and unpublished primary studies.

Reference	**Content of review**	**Subgroup**	**Published studies**	**Unpublished studies**
Alfieri et al. 2011	Effects of discovery-based instruction versus other types of instruction on learning.	Enhanced Discovery: Children	21	1
Benish et al. 2011	The efficacy of culturally adapted psychotherapy versus unadapted, bona fide psychotherapy for ethnic and racial minority clients.	All	13	6
Berry et al. 2011	A comparison of the reports of others versus self-reports of counterproductive work behavior.	Self – Other	12	8
Card et al. 2011	Associations of parental deployment during military conflicts and internalizing, externalizing and academic adjustment among children.	Externalizing	11	0
Farber & Doolin 2011	Associations between therapists' provision of positive regard and psychotherapy outcomes.	All	16	2
Green & Rosenfeld 2011	Efficacy of the Structured Interview of Reported Symptoms to identify feigning from genuine responders.	Average SIRS simulators versus nonclinical	7	5
Hallion & Ruscio 2011	Effects of Cognitive Bias Modification on Anxiety and Depression.	Posttest	37	1
Lucassen et al. 2011	Associations Between Paternal Sensitivity and Infant–Father Attachment Security.	All	13	3
Mol & Bus 2011	Associations between print exposure and components of reading (reading comprehension, technical reading and spelling) across development.	Grades 1-12 Basics	13	1
Woodin 2011	Associations between couple conflict behaviors and relationship satisfaction.	Satisfaction Hostility	40	0
Woodley 2011	Association between measures of life history speed (*K*) and *g* (IQ),	All	3	7

Simulation (Figures 6.2 and 6.4)

Figure 6.2 is generated with the R-code that is available on www.bdat.nl. For each of the underlying true effect sizes (ES) in Cohen's *d* (ranging from 0 to 1 in steps of .05), and each of the three different total sample sizes (10, 20, and 40 for the small studies and 50, 100, and 200 for the large studies), the four strategies were applied 10,000 times.

For Strategy 1 we randomly drew a (large) sample from a normal distribution with mean = 0 and SD = 1, and a second sample from a normal distribution with the underlying true ES as mean, and SD = 1 to generate a dataset that represents the two experimental conditions. We then applied an independent samples t-test (henceforth t-test) on the dataset and saved the p-value and estimated ES in Cohen's *d* from each simulation.

For Strategy 2 we applied QRPs on each dataset generated under Strategy 1. If the original t-test was significant ($p < .05$) and the estimated ES was positive (i.e., in the predicted direction), the p-value, estimated ES, and QRP category (1) were saved and no QRPs were applied. Otherwise, a t-test was performed on a secondary dependent variable that was generated together with the primary dependent variable and that correlated .5 with it. If this second t-test was significant and had a positive outcome (i.e., in the predicted direction), the p-value, estimated ES, and QRP category (2) were saved. Otherwise, 10 values were added to the dataset (for both the primary and secondary dependent variable) and both dependent variables were tested again with a t-test. If one of these tests was significant with a positively estimated ES, this p-value, estimated ES, and QRP category (3) were saved. If both tests were significant and had positive estimated ES, the results with the lowest p-value were saved. If none was significant, outliers ($|Z|>2$) were removed from both datasets (original and second dependent variable both with 10 added values per cell) and another t-test was applied. If one of these tests was significant with positively estimated ES, the p-value, estimated ES, and QRP category (4) were saved. If both tests were significant and showed a positively estimated ES, the result with the lowest p-value was saved. If no significant result could be found, we saved the best result of all performed tests, which was defined as the test with the lowest p-value and a positively estimated ES. If none of the steps resulted in a positive ES, we selected the results with the highest p-value. The p-value, estimated ES and QRP category (5) were kept and formed the basis of the figures.

Strategy 3 was the same as Strategy 1 but rather applied to 5 small and ordered datasets. The results (p-value and estimated ES) of the first of these five datasets that showed a significant p-value and positive ES were collected, together with the number of datasets needed to arrive at this result (i.e., value between 1 and 5). If none of the 5 datasets resulted in a significant t test with a positive ES, we collected the best result, which was defined as the result with the lowest

p-value and a positively estimated ES. If none of the datasets resulted in a positive ES, we selected the results with the highest p-value. The p-value, estimated ES were saved and the maximum number of datasets was set at 5.

Strategy 4 entails a combination of Strategies 2 and 3. On each of the 5 small datasets used in Strategy 3 we applied the QRPs as described under Strategy 2. The results (p-value and estimated ES) of the first dataset (with or without the use of QRPs) were collected together with the number of datasets needed and the QRP category used in the final analysis. If none of the datasets with or without QRPs resulted in a significant t-test, the best result was saved, which was defined above.

For Figure 6.2 we calculated the proportion of significant p-values and ES bias (average estimated ES minus true ES) of the 10,000 simulations for each underlying true ES value, sample size, and strategy.

We performed 16 meta-analyses for Figure 6.4. We used four levels of true ES (d = 0.0, 0.2, 0.5, or 0.8) and the four strategies as described above. Each meta-analysis is based on 100 datasets, where the total sample size of each small dataset is drawn from a negative binomial distribution (mu = 30, size = 2; 10 added). The five small datasets used in Strategies 3 and 4 all have the same sample size and only one of the five datasets was collected and used in the meta-analysis (selection is described above). Sample size of the 100 large datasets (Strategy 1 and 2) was 5 times the small study sample size.

We performed a fixed effect meta-analysis based on Cohen's d and collected the estimated ES, Q statistic with accompanying p-value, the Z and p-value of Sterne & Egger's test for funnel plot asymmetry, and the Chi-square and p-value of Ioannidis & Trikalinos' test for an excess of significant results.

We also ran the 16 meta-analyses 10,000 times to collect the expected values, and those are shown in Table B.2. These results show that with 100 studies, Sterne & Egger's test performs very well with multiple small studies with and without QRPs and reasonably well with one large study with QRPs. The proportion of significant Sterne & Egger's tests with 1 large study and QRPs and a true ES of .8 is fairly low, but in this case QRPs are often not needed to find a significant result. Ioannidis & Trikalinos's test performs well except for 1 large study or 5 small studies with QRPs and a true ES of 0, and for 5 small studies and a true ES of .2. Type 1 error rate of the test as evidenced by the number of rejections under Strategy 1 (left column) is well within the .05 range.

Table B.2

Expected values of estimated ES, Q-test, Sterne and Egger's test for funnel plot asymmetry, and Ioannnidis &Trikalinos' test for an excess of significant results on the basis of 10,000 simulations.

	Strategy 1 1 large study	Strategy 2 1 large study with QRPs	Strategy 3 5 small studies	Strategy 4 5 small studies with QRPs
ES=0	Est ES=.000	Est ES=.084	Est ES=.354	Est ES=.487
	Q=98.4 P(p<.05)=.040	Q=88.6 P(p<.05)=.004	Q=53.3 P(p<.05)=.000	Q=50.8 P(p<.05)=.000
	Bias Z=.002 P(p<.05)=.052	Bias Z=1.933 P(p<.05)=.486	Bias Z=3.548 P(p<.05)=.991	Bias Z=4.875 P(p<.05)=1.000
	I-Chi=.964 P(p<.05)=.032	I-Chi=.773 P(p<.05)=.022	I-Chi=4.664 P(p<.05)=.571	I-Chi=.567 P(p<.05)=.009
ES=.2	Est ES=.199	Est ES=.262	Est ES=.550	Est ES=.624
	Q=98.0 P(p<.05)=.038	Q=70.6 P(p<.05)=.000	Q=51.3 P(p<.05)=.000	Q=45.5 P(p<.05)=.000
	Bias Z=.253 P(p<.05)=.060	Bias Z=2.160 P(p<.05)=.598	Bias Z=3.779 P(p<.05)=.996	Bias Z=5.176 P(p<.05)=1.000
	I-Chi=.444 P(p<.05)=.004	I-Chi=2.096 P(p<.05)=.161	I-Chi=.911 P(p<.05)=.030	I-Chi=14.011 P(p<.05)=.989
ES=.5	Est ES=.497	Est ES=.511	Est ES=.780	Est ES=.763
	Q=95.7 P(p<.05)=.027	Q=77.9 P(p<.05)=.000	Q=52.0 P(p<.05)=.000	Q=52.8 P(p<.05)=.000
	Bias Z=.599 P(p<.05)=.092	Bias Z=1.770 P(p<.05)=.417	Bias Z=4.758 P(p<.05)=1.000	Bias Z=5.777 P(p<.05)=1.000
	I-Chi=.670 P(p<.05)=.016	I-Chi=5.459 P(p<.05)=.692	I-Chi=13.177 P(p<.05)=.986	I-Chi=41.467 P(p<.05)=1.000
ES=.8	Est ES=.795	Est ES=.797	Est ES=.949	Est ES=.917
	Q=91.9 P(p<.05)=.012	Q=88.5 P(p<.05)=.005	Q=61.2 P(p<.05)=.000	Q=61.5 P(p<.05)=.000
	Bias Z=-.862 P(p<.05)=.130	Bias Z=1.056 P(p<.05)=.169	Bias Z=5.099 P(p<.05)=1.000	Bias Z=5.234 P(p<.05)=1.000
	I-Chi=.921 P(p<.05)=.040	I-Chi=1.004 P(p<.05)=.000	I-Chi=21.579 P(p<.05)=1.000	I-Chi=30.150 P(p<.05)=1.000

Notes: Est ES: Estimated mean effect size; Q: test for homogeneity (*DF* = 99), Bias Z: Sterne and Egger's (2005) regression test; I-Chi: Ioannidis & Trikalinos' (2008) test for an excess of significant results (DF = 1); P(p<.05): proportion of significant outcomes in each panel.