**Statistical analyses**

The reproducibility of psychological science was evaluated using significance and *p*-values, effect sizes, qualitative assessments of replication teams, and meta-analysis of the difference of effect size between original and replication study.

**Significance and *p*-values.** Assuming a two-tailed test and significance or alpha level of .05, all test results of original and replication studies were classified as statistically significant (*p*-value ≤ 0.05) and insignificant (*p* > .05). Using the non-significant *p*-values of the replication studies only, we tested the hypothesis that these studies had ‘no evidential value’, i.e. the null-hypothesis of zero-effect holds for all these studies (Simonsohn, Nelson, & Simmons, 2014; Van Assen, Van Aert, & Wicherts, 2015). The hypothesis that the proportions of statistically significant results are equal was tested using the McNemar test for paired nominal data. Second, we compared the central tendency of the distribution of *p*-values of original and replication studies using the Wilcoxon signed-rank test. For both tests we only used complete data, i.e. study-pairs for which both *p*-values were available.

**Effect sizes**. We transformed all effect sizes into correlation coefficients. Correlation coefficients have several advantages over other effect size measures, such as, e.g. Cohen’s *d*. Correlation coefficients are bounded, and well-known and therefore more readily interpretable. Most importantly for our purposes, analysis of correlation coefficients is rather straightforward because, after applying the Fisher transformation, their standard error is only a function of sample size. Formulas and code for converting test statistics *z*, *F*, *t*, and *χ2* into correlation coefficients are provided in supplementary materials (see [A4]). To be able to compare and analyze correlations across study-pairs, negative correlations were transformed into positive correlations of the same value, (only) if negative associations were expected.

Effect sizes were compared using four tests. The central tendency of the effect size distributions of original and replication studies were compared using both a paired two-sample *t*-test and the Wilcoxon signed-rank test. Third, we computed the proportion of study-pairs in which the effect of the original study was stronger than in the replication study, and tested the hypothesis that this proportion is .5. For this test only, we also used the data for which effect size measures were available but no correlation coefficient could be computed (e.g. if a regression coefficient was reported, but not its test statistics). Finally, we calculated the proportion of study-pairs in which the effect of the original study was in the confidence interval of the effect of the replication study, and compared this with the expected proportion using a goodness-of-fit *χ2­*-test. The expected proportion is the sum over expected probabilities across study-pairs. The test assumes the same population effect size for original and replication study in the same study-pair. See the supplementary materials for computational details on the test.

**Meta-analysis**. Meta-analyses were conducted on Fisher-transformed correlations for all study-pairs where both the correlation coefficient and its standard error could be computed. Standard errors could only be computed if test statistics were *r*, *t*, or *F*(1,*df2*)

**Qualitative assessment of “Did it replicate?”** In addition to the quantitative assessments of replication and effect estimation, we conducted a qualitative assessment of whether the replication provided evidence of replicating the original result. In some cases, the quantitative data anticipates a straightforward qualitative assessment of replication. But for more complex designs, such as multivariate interaction effects, the quantitative analysis may not provide a simple interpretation. For qualitative assessment, replication teams answered the following question: “Did the replication study show evidence for the key effect consistent with the original study?” with a 4-point response scale: No, Slightly, Mostly, Yes. Raters could also respond “Not possible to determine.”

**Meta-analysis of all original study effects, and of all replication study effects**. Two random-effects meta-analyses were run, one on effect sizes of original and one on effect sizes of replication studies. The only predictor in each analysis was the study’s standard error, which provides a test of publication bias (ref to Eggers test).

**Meta-analysis of difference of effect size between original and replication study.** The dependent variable was the difference of Fisher-transformed correlations (original – replication), with variance equal to the sum of variances of the correlation of the original and of the replication study (ref). Several random-effect meta-analyses were run using R-package metaphor (ref).

First, the intercept-only model was estimated; the intercept denotes the average difference effect size between original and replication study. Second, to test and control for publication bias, we added the standard error of the original study as a predictor, akin to Eggers test of publication bias; a positive effect signifies publication bias (refs). Our third model tested the effect of study type, with five categories (JPSP-social = reference category, JEP:LMC-cognitive, PSCI-social, PSCI-cognitive, PSCI-other). Then we ran six meta-analyses with standard error of the original study and one of the following six predictors: importance of the effect, surprising effect, experience and expertise of original team, challenge of conducting replication, experience and expertise of replication team, self-assessed quality of replication.

**Results**

*Preliminary analyses*. The input of our analyses were the *p*-values and effect sizes as described in the article of the original study, and determined by the replication team for the replication study. First, we checked the consistency of *p*-value and test statistics whenever possible (i.e., when all were provided), by recalculating the *p­*-value using the test statistics. We used the recalculated *p*-values in our analysis, with one exception where *p* = .05 was reported when the *p*-value was actually .0509; in that case we used .05, since it was interpreted as being significant. We ended up with 89 study-pairs with complete data on *p*-values. See our supplementary materials [A1] for details on the recalculation of *p*-values.

Table 1. Statistical results (statistically significant or not) of original and replication studies.

Results

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Replication | |
|  |  | Nonsignificant | Significant |
| Original | Nonsignificant | 5 | 1 |
|  | Significant | 57 | 32 |

*Significance and p-values (see [A2] for details)*. Table 1 shows the statistical significance of original and replication studies. 93% of the original studies was statistically significant, as opposed to 34.7% of replication studies, which corresponds to a significant change (McNemar test, *χ2*(1) = 54.06897, *p* < .001). Proportions statistical significance of original and replication studies for the three journals JPSP, JEP, PS were .957 and .2, .865 and .462, .949 and .385, respectively. Of 89 significant original studies, 35.96% was statistically significant in the replication study. The hypothesis that all 57 statistically non-significant replication studies come from a population of true negatives can be rejected (*χ2*(124) = 153.9089, *p* = 0.03542). The cumulative *p*-value distributions of original and replication studies are presented in Figure 1. The means of the two *p*-valuedistributions (.029 and .301) were different from each other (*t*(94) = -7.979, *p* < .001; W = 2250.5, *p* < .001).

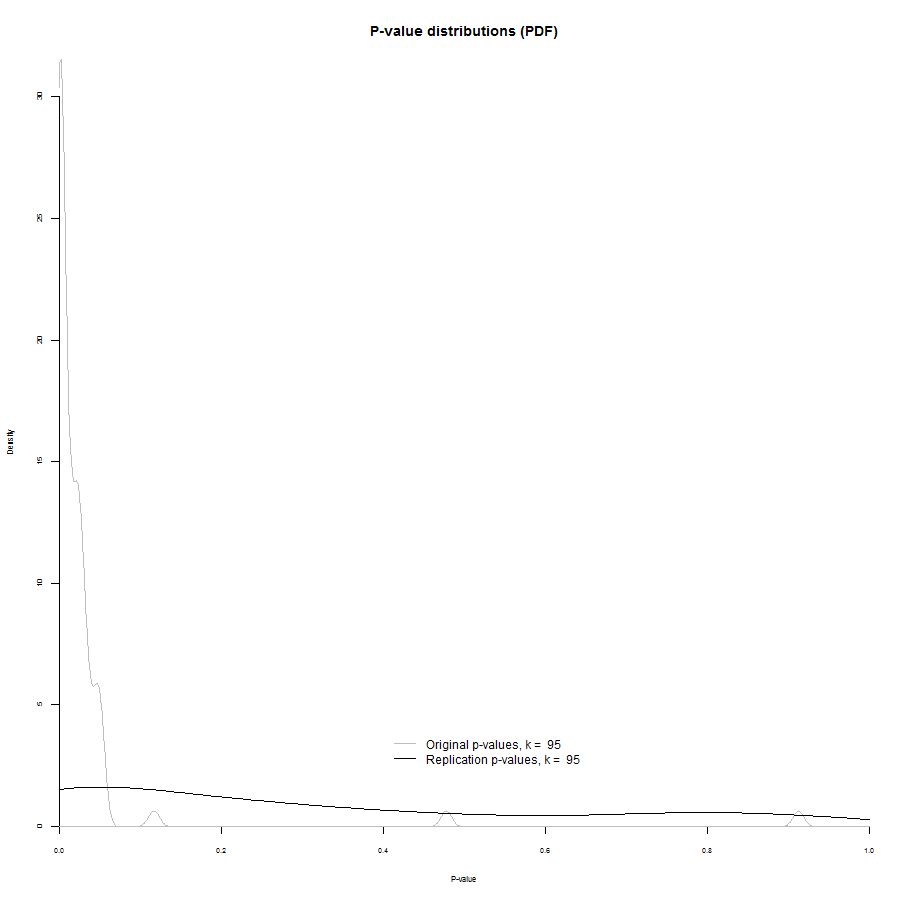
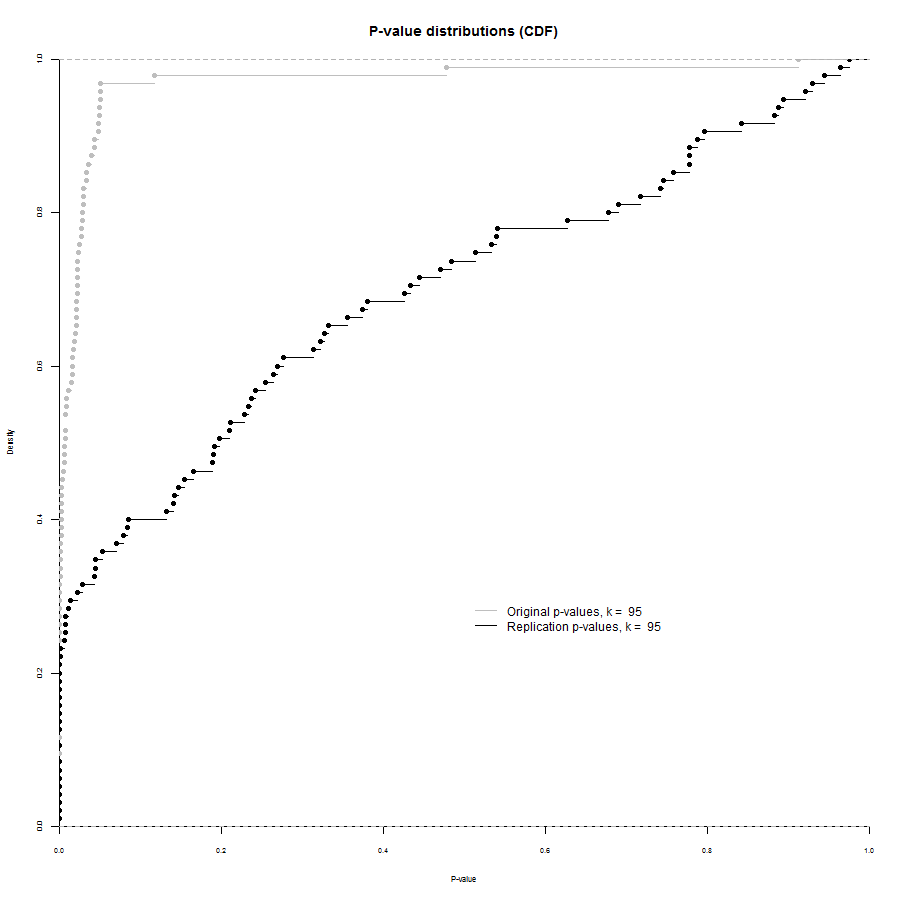


Figure 1: Cumulative *p*-value distributions of original and replication studies.

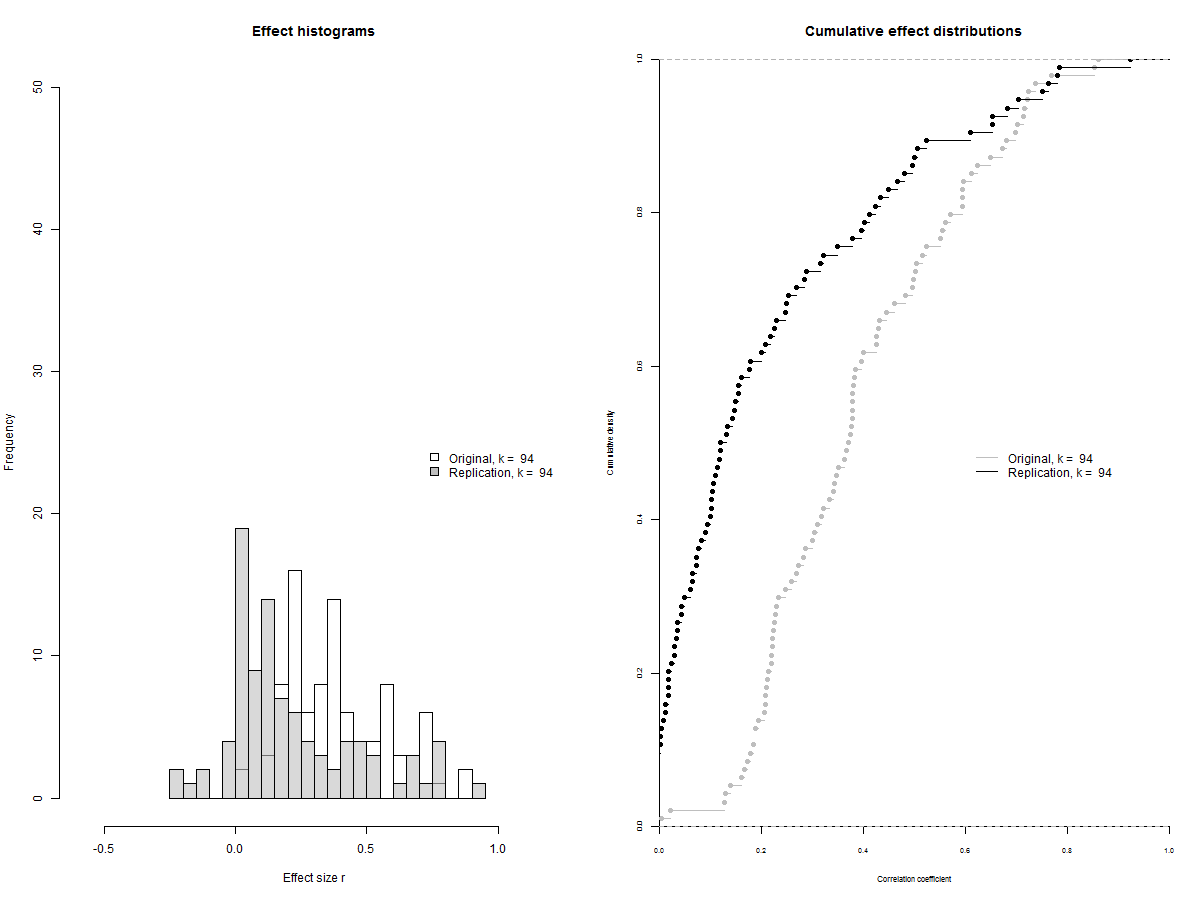


Figure 2: Distributions (left) and cumulative distribution functions of effect sizes of original and replication studies.

*Effect sizes (see [A4] for details)*. Figure 2 (left) shows the distribution of effect sizes (correlations) of original and replication studies, and the corresponding cumulative distribution functions (right). The mean effect sizes of both distributions (*M* = .392 [*SD* = .193]; *M* = .209 [*SD* =.244]) were different from each other (*t*(93) = 9.3625, *p* < .001; *W* = 6633, *p* < .001). Of those 95 studies that reported an effect size in both original and replication study, 42 reported a stronger effect size in the replication study (44.2%; *p* = .3049, binomial test). Figure 3 depicts effect sizes of study-pairs of which correlations could be calculated, and codes significance of effect sizes as well.

\*\*\* no result yet on percentage of original effect sizes in CI replication studies: see [A4] \*\*\*

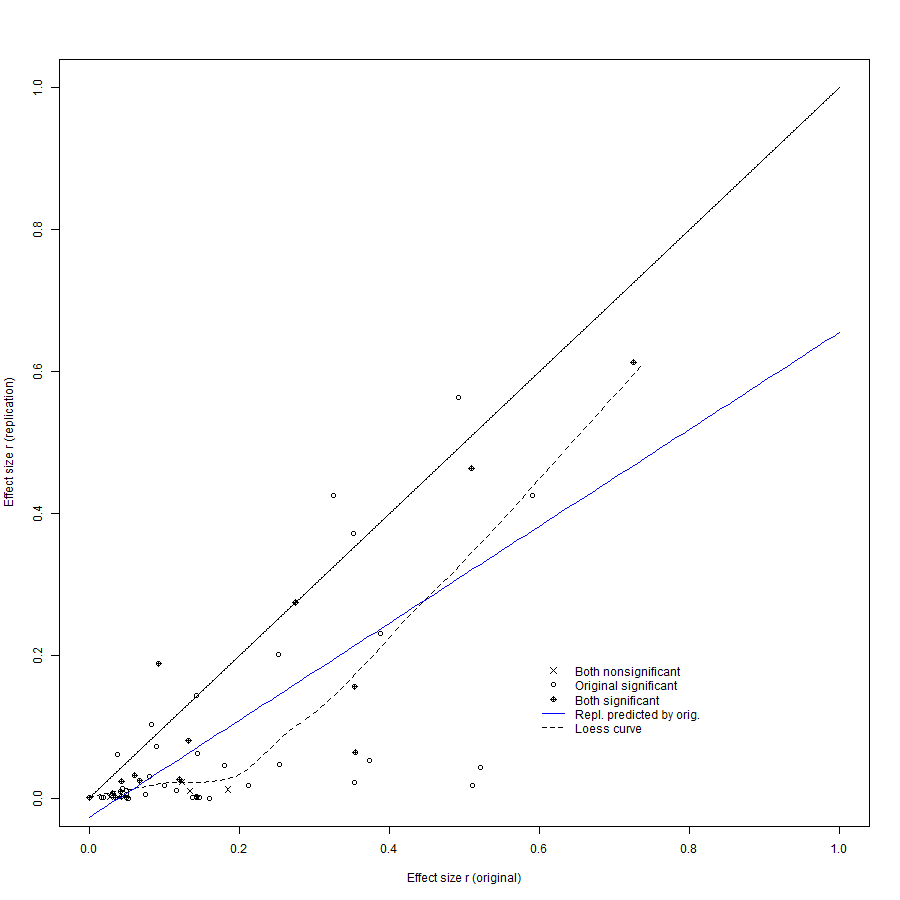


Figure 3: Correlation effect sizes measures of both original and replication study, coded by statistical significance. Identical values are indicated by the diagonal line.

*Meta-analysis (see [A5] for details)*. For 57 study-pairs a meta-analysis could be conducted. In 45 out of 57 pairs the null-hypothesis of no effect was rejected (78.9%).

*Qualitative assessment of “Did it replicate?”* [add results]

*Meta-analysis of all original study effects, and of all replication study effects.*

The meta-analysis on all original study effect sizes showed significant (*Q*(55) = 160.29, *p* < .001) and moderate to large heterogeneity (=.17, *I2* = 64.0%). The effect of the original study’s standard error on effect size was large and highly significant (*b* = 1.88, *z* = 4.16, *p* < .001). Figure 4 shows the funnel plot of the meta-analysis without predictors.

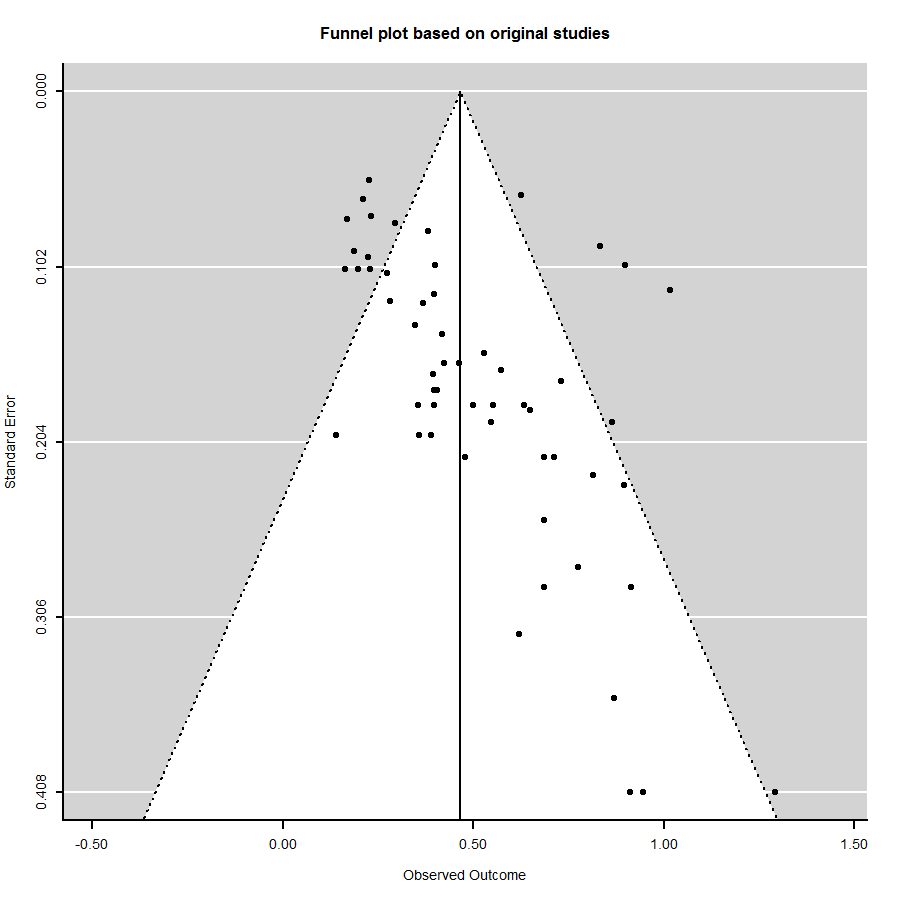


Figure 4: Funnel plot of the meta-analysis on the original study’s effect size.

The same meta-analysis on replication studies’ effect sizes significant (*Q*(55) = 255.89, *p* < .001) and large heterogeneity (=.21, *I2* = 80.7%). The effect of the standard error of the replication study was large and highly significant (*b* = 2.10, *z* = 4.09, *p* = .001), comparable to its effects for the original studies. Figure 5 shows the corresponding funnel plot.

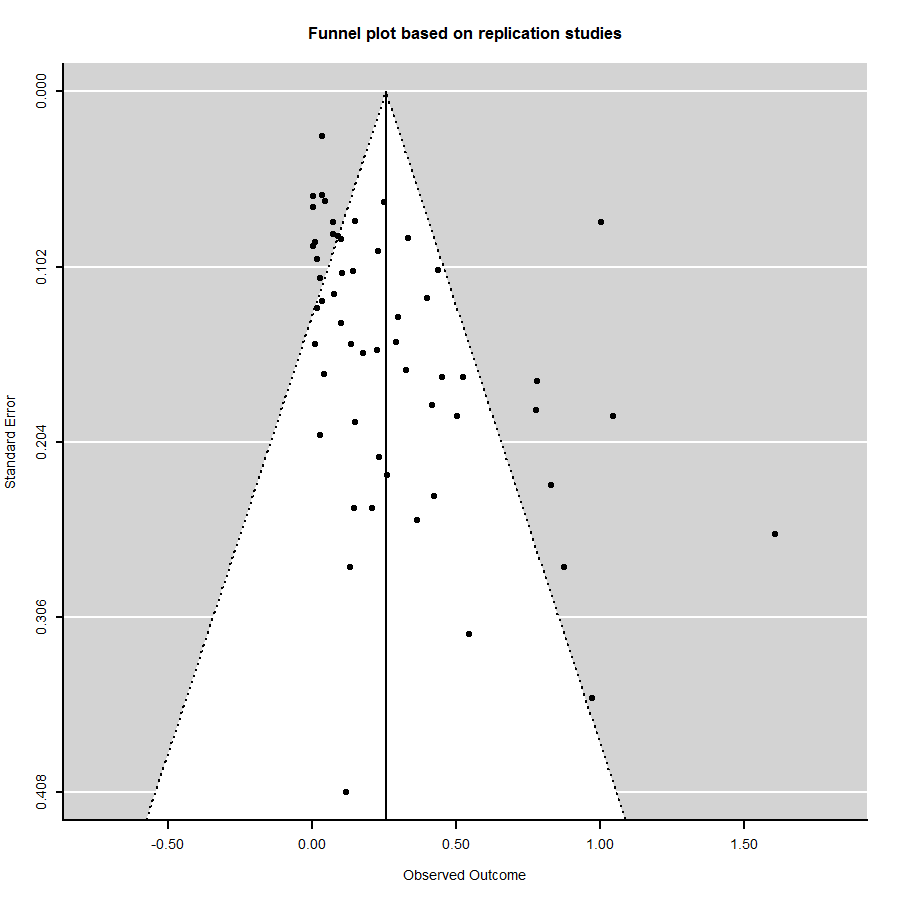


Figure 5: Funnel plot of the meta-analysis on the replication study’s effect size.

*Meta-analysis of difference of effect size between original and replication study.*

The null-model without predictors yielded an average estimated difference in effect size equal to .20 (*z* = 7.25, *p* < .001). The null-hypothesis of homogenous difference in effect sizes was not rejected (*Q*(56) = 65.7, *p* = .18), with small observed heterogeneity (=.095, I2 = 23%). Precision of the original study was not associated to the difference in effect size (*b* = .39, *z* = .86*,* one-tailed *p* = .2), hence imprecise studies (large standard error) did not yield larger effect size differences. This is confirmed by the funnel plot in Figure 6. Study type was also not associated to the difference in effect size (χ2(4) = 2.15, *p* = .71), i.e. the average difference in effect size was equal for JPSP, JEP, PS-soc, PS-cogn, and PS-other. \*\*\* Results of meta-analyses on other moderators \*\*\*.

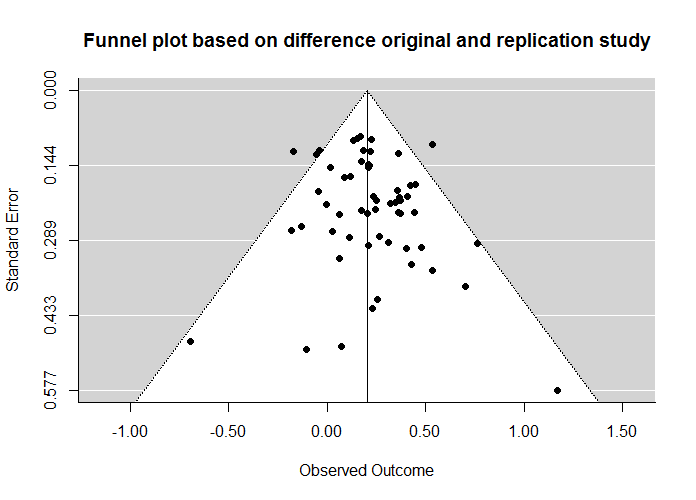


Figure 6: Funnel plot of meta-analysis on difference in effect size (original – replication).

Option 1 for standardizing: (x-mean)/sd where x is a particular score on a variable

fis.o fis.r es.meta diff importance surprising experience.O challenge experience.R quality

fis.o 1.000 0.612 0.841 0.210 -0.269 -0.172 -0.083 -0.167 0.159 -0.020

fis.r 0.612 1.000 0.912 -0.644 -0.130 -0.188 -0.022 -0.235 -0.148 -0.027

es.meta 0.841 0.912 1.000 -0.314 -0.233 -0.234 -0.071 -0.232 0.026 0.007

diff 0.210 -0.644 -0.314 1.000 -0.100 0.066 -0.052 0.130 0.338 0.014

importance -0.269 -0.130 -0.233 -0.100 1.000 0.435 0.434 0.336 0.090 0.072

surprising -0.172 -0.188 -0.234 0.066 0.435 1.000 0.027 0.228 0.080 -0.066

experience.O -0.083 -0.022 -0.071 -0.052 0.434 0.027 1.000 0.131 -0.072 0.233

challenge -0.167 -0.235 -0.232 0.130 0.336 0.228 0.131 1.000 -0.076 -0.001

experience.R 0.159 -0.148 0.026 0.338 0.090 0.080 -0.072 -0.076 1.000 -0.104

quality -0.020 -0.027 0.007 0.014 0.072 -0.066 0.233 -0.001 -0.104 1.000

Option 2 for standardizing: (x-min)/(max-min) where x is a particular score on a variable, min is the lowest possible score on the scale and max is the highest possible score on the scale

fis.o fis.r es.meta diff importance surprising experience.O challenge experience.R quality

fis.o 1.000 0.612 0.841 0.210 -0.254 -0.172 -0.089 -0.294 0.145 -0.032

fis.r 0.612 1.000 0.912 -0.644 -0.120 -0.188 -0.014 -0.317 -0.142 -0.037

es.meta 0.841 0.912 1.000 -0.314 -0.222 -0.234 -0.075 -0.363 0.011 -0.020

diff 0.210 -0.644 -0.314 1.000 -0.098 0.066 -0.069 0.107 0.316 0.015

importance -0.254 -0.120 -0.222 -0.098 1.000 0.401 0.452 0.244 0.101 0.041

surprising -0.172 -0.188 -0.234 0.066 0.401 1.000 0.043 0.107 0.092 -0.069

experience.O -0.089 -0.014 -0.075 -0.069 0.452 0.043 1.000 0.047 -0.139 0.224

challenge -0.294 -0.317 -0.363 0.107 0.244 0.107 0.047 1.000 0.018 -0.089

experience.R 0.145 -0.142 0.011 0.316 0.101 0.092 -0.139 0.018 1.000 -0.117

quality -0.032 -0.037 -0.020 0.015 0.041 -0.069 0.224 -0.089 -0.117 1.000

Table with results of meta-analyses

ID fis.o sei.o pval.o fis.r sei.r pval.r diff es.meta se.meta ci.lb.meta ci.ub.meta pval.meta

1 1 0.685 0.289 0.009 0.149 0.192 0.219 0.535 0.314 0.160 0.000 0.628 0.050

2 2 0.711 0.213 0.000 0.234 0.213 0.136 0.477 0.472 0.151 0.177 0.768 0.002

3 3 0.454 0.209 0.015 -0.219 0.183 0.884 0.672 0.073 0.137 -0.196 0.342 0.594

4 4 0.233 0.073 0.001 -0.006 0.061 0.540 0.239 0.093 0.047 0.001 0.185 0.047

5 5 0.499 0.183 0.003 0.136 0.147 0.179 0.363 0.279 0.115 0.054 0.504 0.015

6 6 0.685 0.213 0.001 0.419 0.183 0.011 0.267 0.532 0.139 0.260 0.803 0.000

7 7 0.898 0.101 0.000 0.132 0.277 0.317 0.765 0.808 0.095 0.622 0.994 0.000

8 10 0.865 0.192 0.000 1.047 0.189 0.000 -0.183 0.958 0.135 0.693 1.222 0.000

9 11 0.815 0.224 0.000 0.506 0.189 0.004 0.309 0.634 0.144 0.351 0.917 0.000

10 15 0.198 0.104 0.028 0.252 0.065 0.000 -0.054 0.237 0.055 0.129 0.344 0.000

11 19 0.634 0.183 0.000 0.426 0.236 0.035 0.208 0.556 0.144 0.273 0.839 0.000

12 20 0.228 0.104 0.014 0.019 0.098 0.421 0.209 0.117 0.071 -0.022 0.257 0.099

13 24 0.381 0.081 0.000 0.292 0.146 0.023 0.089 0.360 0.071 0.221 0.499 0.000

14 26 0.162 0.104 0.059 0.145 0.105 0.083 0.017 0.154 0.074 0.009 0.298 0.037

15 27 0.398 0.183 0.015 0.400 0.120 0.000 -0.002 0.399 0.101 0.202 0.596 0.000

16 28 0.356 0.183 0.026 0.104 0.106 0.164 0.252 0.167 0.092 -0.012 0.347 0.068

17 29 0.946 0.408 0.010 0.875 0.277 0.001 0.071 0.897 0.229 0.448 1.347 0.000

18 32 0.730 0.169 0.000 0.524 0.167 0.001 0.207 0.626 0.119 0.393 0.858 0.000

19 33 0.572 0.162 0.000 0.327 0.162 0.022 0.245 0.450 0.115 0.225 0.674 0.000

20 36 0.895 0.229 0.000 0.832 0.229 0.000 0.063 0.864 0.162 0.546 1.182 0.000

21 37 0.620 0.316 0.025 0.365 0.250 0.072 0.255 0.463 0.196 0.079 0.848 0.018

22 44 0.368 0.123 0.001 0.151 0.076 0.023 0.217 0.211 0.064 0.084 0.337 0.001

23 48 0.229 0.105 0.014 0.009 0.072 0.450 0.220 0.080 0.060 -0.036 0.197 0.178

24 49 0.398 0.174 0.011 -0.030 0.108 0.611 0.429 0.089 0.092 -0.091 0.270 0.332

25 52 0.209 0.088 0.009 0.094 0.095 0.161 0.114 0.156 0.065 0.030 0.283 0.015

26 53 0.397 0.183 0.015 0.077 0.118 0.257 0.320 0.171 0.099 -0.023 0.365 0.084

27 56 0.399 0.101 0.000 -0.042 0.164 0.601 0.441 0.278 0.086 0.109 0.447 0.001

28 58 0.169 0.074 0.012 0.037 0.060 0.270 0.132 0.089 0.047 -0.003 0.181 0.057

29 61 0.223 0.097 0.010 0.005 0.068 0.472 0.219 0.076 0.055 -0.032 0.185 0.167

30 63 0.281 0.122 0.011 0.074 0.083 0.187 0.207 0.140 0.069 0.005 0.275 0.042

31 65 0.462 0.158 0.002 0.012 0.088 0.447 0.450 0.118 0.077 -0.033 0.268 0.125

32 68 0.188 0.093 0.022 0.003 0.067 0.482 0.185 0.066 0.055 -0.041 0.173 0.224

33 71 0.226 0.052 0.000 0.073 0.076 0.166 0.152 0.177 0.043 0.093 0.261 0.000

34 72 0.211 0.062 0.000 0.045 0.064 0.242 0.166 0.129 0.045 0.042 0.217 0.004

35 81 0.275 0.106 0.005 -0.102 0.086 0.883 0.377 0.047 0.067 -0.084 0.178 0.480

36 87 0.418 0.141 0.002 0.013 0.147 0.465 0.405 0.224 0.102 0.024 0.424 0.028

37 89 0.141 0.200 0.241 0.029 0.200 0.442 0.111 0.085 0.141 -0.192 0.362 0.547

38 93 0.329 0.110 0.001 -0.136 0.122 0.867 0.465 0.120 0.082 -0.041 0.280 0.144

39 94 0.359 0.200 0.036 0.298 0.131 0.012 0.061 0.317 0.110 0.101 0.532 0.004

40 97 0.398 0.118 0.000 0.037 0.026 0.077 0.361 0.054 0.025 0.004 0.103 0.034

41 106 0.402 0.174 0.010 -0.227 0.151 0.934 0.629 0.043 0.114 -0.181 0.266 0.707

42 107 0.226 0.110 0.020 0.105 0.080 0.095 0.121 0.147 0.065 0.020 0.274 0.023

43 110 0.625 0.060 0.000 0.091 0.084 0.139 0.533 0.445 0.049 0.349 0.541 0.000

44 111 0.347 0.136 0.005 0.230 0.093 0.007 0.117 0.267 0.077 0.116 0.418 0.001

45 112 0.869 0.354 0.007 0.974 0.354 0.003 -0.104 0.922 0.250 0.432 1.412 0.000

46 113 0.831 0.090 0.000 1.005 0.076 0.000 -0.173 0.933 0.058 0.819 1.047 0.000

47 114 0.649 0.186 0.000 0.780 0.186 0.000 -0.131 0.714 0.131 0.457 0.972 0.000

48 115 0.552 0.183 0.001 0.485 0.378 0.100 0.067 0.539 0.164 0.217 0.861 0.001

49 116 0.296 0.076 0.000 0.335 0.085 0.000 -0.039 0.313 0.057 0.202 0.425 0.000

50 118 0.217 0.095 0.011 0.049 0.080 0.270 0.168 0.118 0.061 -0.002 0.238 0.053

51 120 0.400 0.189 0.017 0.255 0.158 0.053 0.145 0.315 0.121 0.077 0.553 0.009

52 122 0.912 0.408 0.013 1.608 0.258 0.000 -0.697 1.409 0.218 0.982 1.837 0.000

53 124 0.405 0.174 0.010 0.034 0.122 0.389 0.371 0.157 0.100 -0.039 0.353 0.117

54 129 0.390 0.200 0.025 0.018 0.126 0.444 0.373 0.124 0.107 -0.085 0.333 0.246

55 133 0.479 0.213 0.012 0.452 0.167 0.003 0.027 0.462 0.131 0.205 0.720 0.000

56 134 0.213 0.094 0.011 0.550 0.066 0.000 -0.336 0.439 0.054 0.334 0.544 0.000

57 135 0.005 0.042 0.456 0.106 0.017 0.000 -0.102 0.092 0.016 0.062 0.123 0.000

58 136 0.547 0.192 0.002 0.103 0.135 0.223 0.444 0.249 0.110 0.033 0.465 0.024

59 145 1.017 0.115 0.000 0.780 0.169 0.000 0.237 0.942 0.095 0.755 1.129 0.000

60 146 0.775 0.277 0.003 0.545 0.316 0.042 0.230 0.675 0.209 0.267 1.084 0.001

61 148 0.191 0.072 0.004 0.030 0.062 0.314 0.161 0.099 0.047 0.007 0.191 0.036

62 150 0.913 0.289 0.001 0.211 0.243 0.193 0.702 0.501 0.186 0.137 0.865 0.007

63 151 0.424 0.158 0.004 0.003 0.090 0.487 0.421 0.106 0.078 -0.047 0.260 0.176

64 153 1.292 0.408 0.001 0.121 0.408 0.384 1.171 0.706 0.289 0.140 1.272 0.014

65 154 0.460 0.122 0.000 0.110 0.277 0.345 0.349 0.403 0.112 0.184 0.622 0.000

66 155 0.321 0.141 0.012 -0.034 0.120 0.611 0.355 0.115 0.092 -0.065 0.295 0.210

67 158 0.394 0.164 0.008 0.437 0.104 0.000 -0.043 0.425 0.088 0.252 0.597 0.000

68 161 0.527 0.152 0.000 0.180 0.152 0.119 0.348 0.354 0.108 0.142 0.565 0.001

69 167 0.686 0.250 0.003 0.260 0.224 0.123 0.426 0.449 0.167 0.122 0.776 0.007

**Supplementary materials**

**[A1] Recalculation of *p*-values**

Studies of row 23 and row 149 were not considered, because they are copies of row 24 and row 150, respectively.

[Here how and who entered p-values in master data-file: Johanna].

*Recalculation of p-values.* The *p*-values were recalculated using the test statistic and the degrees of freedom, with the following R-function:

# Recalculating p-values

# Written by CHJ Hartgerink and RCM van Aert

pvalr <- function(x, N) {

fis.r <- 0.5\*log((1 + x) / (1 - x))

se.fis.r <- sqrt(1/(N-3))

pnorm(fis.r, mean = 0, sd = se.fis.r, lower.tail = FALSE)

}

# Computes two-tailed p-value

pvalComp <- function(

x,

df1,

df2,

N,

esType){

pvalComp <- ifelse(esType=="t",

pt(abs(x), df = df2, lower.tail = FALSE) \* 2,

ifelse(

esType=="F",

pf(x, df1 = df1, df2 = df2, lower.tail = FALSE),

ifelse(

esType=="r",

pvalr(abs(x), N) \* 2,

ifelse(

esType=="Chi2",

pchisq(x, df = df1, lower.tail = FALSE),

ifelse(

esType == "z",

pnorm(abs(x), lower.tail = FALSE) \* 2,

NA

)

)

)

))

return(pvalComp)

}

We checked manually if reported *p*-values were one-tailed or two-tailed. The default for *r*, *z*, and *t* were two-tailed *p*-values. If *p*-values could not be recalculated because test statistics were not provided, we used the reported *p*-value.

*Remarks p-values original studies*

The following three *p*-values reported in the original studies were considered inconsistent:

Row 19: reported *p*-value = .01, recalculated *p*-value = .029.

Row 38: reported *p*-value = .05, recalculated *p*-value = .0509 [.05 was used for the analyses]

Row 97: reported *p*-value = .001, recalculated *p*-value = .0031

We used the recalculated *p*-values in our analysis, with the exception of *p* = .05, since this *p*-value was interpreted as being significant in the original study.

Of 91 *p*-values, three could not be recalculated but a *p*-value was reported (rows 49, 60, 70), and two *p*-values were not reported and could not be recalculated; the latter two studies had to be discarded for the analyses of *p*-values. Hence 89 original studies had data on *p*-values.

*Remarks p-values replication studies*

No inconsistencies were found.

When inspecting only pairwise available p-values across original and replications, 1 p-value was reported as *prep*, 1 imprecisely (e.g., p < .05) and 1 as p=0. We ended up with 89 *p*-values for analysis.

Of 91 *p*-values, three *p*-values could not be recalculated but a *p*-value was reported (rows (53, 60, 70), and two studies had to be discarded for analyses because no *p*-value was reported and also could not be recalculated (rows 135 and 136).

**[A2] Analyses of significance and *p*-values**

The code for the McNemar test of change in statistical significance:

# McNemar test

tab <- table(dat$sign..O.[!is.na(dat$sign..O.) & !is.na(dat$sign..R.)],

dat$sign..R.[!is.na(dat$sign..O.) & !is.na(dat$sign..R.)])

mcnemarchi <- (tab[1,2]-tab[2,1])^2/(tab[1,2]+tab[2,1])

mcnemarp <- pchisq(q = mcnemarchi, df = 1, lower.tail = FALSE)

The code for the (Fisher, *p*-curve, *p*-uniform) test of no evidential value in the non-significant replication studies:

# Written by CHJ Hartgerink

# The Fisher method applied to test for deviation from uniformity

# In NONSIGNIFICANT P-values

FisherMethod <- function(# Compute Fisher's exact test for non-significant p-values.

### This function computes paper level Fisher test statistics, testing whether the distribution of non-significant p-values is uniform. Significant values indicate deviation from uniformity.

### Returns both the normal Fisher test, as well as the complement test.

### Computations are done for p\*=log(p), where p is all non-significant p-values for each identifier.

x,

### Vector of p-values.

id,

### Vector giving paper identifiers.

alpha = .05

### Indicate what alpha level is being maintained for the study results, which serves as a cut-off for selecting the non-significant p-values.

){

Res <- NULL

for(i in 1:length(unique(id)))

{

selP <- x[id==unique(id)[i]]

nSigP <- (na.omit(selP[selP>alpha])-alpha)/(1-alpha)

SigP <- na.omit(selP[selP<=alpha])

if(!length(nSigP)==0){

# Compute the Fisher test statistic

FMeth <- -2\*sum(log(nSigP))

# Compute p-values analytically

pFMeth <- pchisq(q=FMeth, df=2\*length(nSigP), lower.tail=F)

} else {

FMeth <- NA

pFMeth <- NA

}

Res <- rbind(Res, data.frame(

Fish = FMeth,

PFish = pFMeth,

CountNSig = length(nSigP),

CountSig = length(SigP),

PercentNonSig = length(nSigP)/length(selP)))

}

return(Res)

}

The code for the test comparing the means of the two dependent samples:

# Dependent t-test p-values

t.test(x = dat$pval\_USE..O.[!is.na(dat$pval\_USE..O.) & !is.na(dat$pval\_USE..R.)],

y = dat$pval\_USE..R.[!is.na(dat$pval\_USE..O.) & !is.na(dat$pval\_USE..R.)],

paired = TRUE)

# Wilcoxon signed-rank test p-values

wilcox.test(dat$pval\_USE..O.[!is.na(dat$pval\_USE..O.) & !is.na(dat$pval\_USE..R.)],

dat$pval\_USE..R.[!is.na(dat$pval\_USE..O.) & !is.na(dat$pval\_USE..R.)],

alternative="two.sided")

sd(dat$pval\_USE..O.[!is.na(dat$pval\_USE..O.) & !is.na(dat$pval\_USE..R.)])

summary(dat$pval\_USE..O.[!is.na(dat$pval\_USE..O.) & !is.na(dat$pval\_USE..R.)])

sd(dat$pval\_USE..R.[!is.na(dat$pval\_USE..O.) & !is.na(dat$pval\_USE..R.)])

summary(dat$pval\_USE..R.[!is.na(dat$pval\_USE..O.) & !is.na(dat$pval\_USE..R.)])

**[A3] Calculation of effect sizes**

If possible, we calculated the “correlation coefficient per df” as effect size measure based on the reported test statistics. This was possible for the *z*, χ2, *t*, and *F* statistic. The code for the calculation is:

esComp <- function(

x,

df1,

df2,

N,

esType){

esComp <- ifelse(esType=="t",

sqrt((x^2\*(1 / df2)) / (((x^2\*1) / df2) + 1)),

ifelse(

esType=="F",

sqrt((x\*(df1 / df2)) / (((x\*df1) / df2) + 1))\*sqrt(1/df1),

ifelse(

esType=="r",

x,

ifelse(

esType=="Chi2",

sqrt(x/N),

ifelse(

esType == "z",

tanh(x \* sqrt(1/(N-3))),

NA

)

)

)

))

return(esComp)

The *z* statistic is transformed into a correlation using sample size *N* with , with *rf* the Fisher-transformed correlation. The χ2 is transformed into the or correlation coefficient with . The *t* and *F* statistic are transformed into a “correlation per *df*” using

, where *F*= *t*2. The expression in the first square-root equals the proportion of variance explained by the *df1* predictors of the variance not yet explained by these same predictors. To take into account that more predictors can explain more variance, we divided this number by *df1* to obtain the “explained variance by predictor”. Taking the square root gives the correlation, or more precisely, it gives the correlation of each predictor assuming that all *df1* predictors contribute equally to the explained variance of the dependent variable.

For 91 replication studies an effect size was reported, of which 85 could be transformed into a correlation. These studies (rows 53, 60, 70, 135, and 136 were discarded from analyses on effect sizes). The correlation could not be calculated for four original studies (rows 49, 60, 70, 135), which resulted in complete data for 82 study-pairs.

To be able to compare and analyze correlations across study-pairs, correlations were multiplied with -1 if the original correlation was negative. The effect sizes used for our analyses can be found in columns DH and DS of the excel-file.

**[A4] Analyses of effect sizes**

The code for the tests comparing means of dependent samples:

# Dependent t-test effects (r values)

t.test(x = dat$r..O.[!is.na(dat$r..O.) & !is.na(dat$r..R.)],

y = dat$r..R.[!is.na(dat$r..O.) & !is.na(dat$r..R.)],

paired = TRUE)

# Wilcox test effects (r values)

wilcox.test(dat$r..O.[!is.na(dat$r..O.) & !is.na(dat$r..R.)],

dat$r..R.[!is.na(dat$r..O.) & !is.na(dat$r..R.)],

alternative="two.sided")

summary(dat$r..O.[!is.na(dat$r..O.) & !is.na(dat$r..R.)])

sd(dat$r..O.[!is.na(dat$r..O.) & !is.na(dat$r..R.)])

summary(dat$r..R.[!is.na(dat$r..O.) & !is.na(dat$r..R.)])

sd(dat$r..R.[!is.na(dat$r..O.) & !is.na(dat$r..R.)])

mean(dat$r..O.[!is.na(dat$r..O.) & !is.na(dat$r..R.)])-mean(dat$r..R.[!is.na(dat$r..O.) & !is.na(dat$r..R.)])

The test comparing effect sizes (‘which is stronger?’) is done manually, because of many variations and incompatible reported effect sizes in some study-pairs.

*\*\*\* no result yet on percentage of original effect sizes in CI replication studies \*\*\**

*We can do this for effect sizes that include statistics r, t, and F. For z-tests we need means and standard deviations of both original and replication study, and for χ2-tests we need frequencies of two-by-two tables (in case of a test of statistical independence) or other details about the statistics. Could you provide these details for studies 40, 60, 70, 74, 85, 105, 122, 135, 136, 149, 166?*

*\*\*\* \*\*\**

**[A5] Meta-analyses on effect sizes of each study-pair**

We present all code for all meta-analyses in supplementary materials [A6]. The code corresponding to meta-analyses on effect sizes of each pair is part g. of the code. The meta-analyses were conducted on Fisher-transformed for all study-pairs where both the correlation coefficient and its standard error could be computed. Standard errors could only be computed if test statistics were *r*, *t*, or *F*(1,*df2*), which was for 57 study-pairs. Standard errors were computed using , which assumes tests of one correlation or an independent sample *t*-test (but not a dependent sample *t*-test).

**[A6] Meta-analyses on all study pairs**

The code for all meta-analyses is:

################################

##### META-ANALYSES #####

##### R.C.M. van Aert #####

################################

### Load data

data <- read.csv("D:/Dropbox/reproducibility project/robbie meta.csv", sep = ";")[1:167, ]

### Recode "Publishing journal and subdiscipline"

jour <- numeric()

for(i in 1:nrow(data)) {

if(as.character(data$Journal..O.[i]) == "JEPLMC") {

jour[i] <- 1

} else if(as.character(data$Journal..O.[i]) == "JPSP") {

jour[i] <- 2

} else if(as.character(data$Journal..O.[i]) == "PS") {

if(as.character(data$Discipline..O.[i]) == "Cognitive") {

jour[i] <- 3

} else if(as.character(data$Discipline..O.[i]) == "Social") {

jour[i] <- 4

} else { jour[i] <- 5 }

}

else { jour[i] <- NA }

}

########################################

### Create right order of categories ###

########################################

### Create the right order of the variable "Exciting result"

fac.exci <- factor(data$Exciting.result..O., levels = c("Not at all exciting and important", "Slightly exciting and important",

"Somewhat exciting and important", "Moderately exciting and important",

"Very exciting and important", "Extremely exciting and important"))

### Create the right order of the variable "Surprising result"

### Recoded misspelled name

levels(data$Surprising.result..O.)[levels(data$Surprising.result..O.) == "Extremely surpising"] <- "Extremely surprising"

fac.surp <- factor(data$Surprising.result..O., levels = c("Not at all surprising", "Slightly surprising", "Somewhat surprising",

"Moderately surprising", "Very surprising", "Extremely surprising"))

### Create the right order of the variable "Perceived expertise required"

fac.expe <- factor(data$Methodology.expertise.required..O., levels = c("No expertise required", "Slight expertise required",

"Moderate expertise required", "Strong expertise required",

"Extreme expertise required"))

### Create the right order of the variable "Opportunity for expectancy biases"

fac.oppo.expe <- factor(data$Opportunity.for.expectancy.bias..O., levels = c("No opportunity for researcher expectations to influence results",

"Slight opportunity for researcher expectations to influence results",

"Moderate opportunity for researcher expectations to influence results",

"Strong opportunity for researcher expectations to influence results",

"Extreme opportunity for researcher expectations to influence results"))

### Create the right order of the variable "Opportunity for lack of diligence"

fac.oppo.dili <- factor(data$Opportunity.for.lack.of.diligence..O., levels = c("No opportunity for lack of diligence to affect the results",

"Slight opportunity for lack of diligence to affect the results",

"Moderate opportunity for lack of diligence to affect the results",

"Strong opportunity for lack of diligence to affect the results",

"Extreme opportunity for lack of diligence to affect the results"))

### Create the right order of the variable "Current position"

fac.posi <- factor(data$Current.position..R., levels = c("Professor (or equivalent)", "Associate Professor (or equivalent)", "Assistant Professor (or equivalent)",

"Post-doc or Research Scientist", "PhD student", "Master's student", "Undergraduate student", "Faculty (non-tenure track) full-time lecturer"))

### Change "Faculty (non-tenure track) full-time lecturer" into "Assistant Professor (or equivalent)"

tmp <- numeric()

for(i in 1:length(fac.posi)) {

if(is.na(fac.posi[i] == TRUE)) { tmp[i] <- NA }

else if(as.character(fac.posi[i]) == "Faculty (non-tenure track) full-time lecturer") { tmp[i] <- "Assistant Professor (or equivalent)"

} else { tmp[i] <- as.character(fac.posi[i]) }

}

fac.posi <- as.factor(tmp)

### Create the right order of the variable "Degree"

fac.degr <- factor(data$Degree..R., levels = c("PhD or equivalent", "Master's degree or equivalent", "Some graduate school", "Bachelor's degree or equivalent", "some college/university",

"high school/equivalent"))

### Create the right order of the variable "Domain expertise"

fac.doma <- factor(data$Domain.expertise..R., levels = c("No Expertise", "Slight Expertise", "Some Expertise", "Moderate Expertise", "High Expertise"))

### Create the right order of the variable "Method expertise"

fac.meth <- factor(data$Method.expertise..R., levels = c("No Expertise", "Slight Expertise", "Some Expertise", "Moderate Expertise", "High Expertise"))

### Create the right order of the variable "Implementation quality"

fac.impl <- factor(data$Implementation.quality..R., levels = c("was of much higher quality than the original study", "was of moderately higher quality than the original study",

"was of slightly higher quality than the original study", "was about the same quality as the original study",

"was of slightly lower quality than the original study", "was of moderately lower quality than the original study"))

### Create the right order of the variable "Data collection quality"

fac.data <- factor(data$Data.collection.quality..R., levels = c("was much better than the average study", "was better than the average study", "was slightly better than the average study",

"was about the same as the average study", "was slightly worse than the average study", "was worse than the average study",

"was much worse than the average study"))

### Create the right order of the variable "Replication similarity"

fac.repl <- factor(data$Replication.similarity..R., levels = c("Not at all similar", "Slightly similar", "Somewhat similar", "Moderately similar", "Very similar", "Extremely similar",

"Virtually identical"))

### Create the right order of the variable "Difficulty of implementation"

fac.diff <- factor(data$Difficulty.of.implimentation..R., levels = c("Extremely challenging", "Very challenging", "Moderately challenging", "Somewhat challenging", "Slightly challenging",

"Not at all challenging"))

###########################################

### Prepare variables for meta-analyses ###

###########################################

### Function for standardizing variables

stand <- function(x, max, min, option) {

if(option == 1) {

res <- (x-mean(x, na.rm = TRUE))/sd(x, na.rm = TRUE)

}

if(option == 2) {

res <- (x - min)/(max-min)

}

return(res)

}

### OPTION 1 FOR STANDARDIZING ###

option <- 1

# a. PUBLISHING JOURNAL AND SUBDISCIPLINE

### Create dummy variables for "Publishing journal and subdiscipline"

### JPSP is reference category because it has the most cases

d.JEP <- ifelse(jour == 1, 1, 0)

d.PSCog <- ifelse(jour == 3, 1, 0)

d.PSSoc <- ifelse(jour == 4, 1, 0)

d.PSOth <- ifelse(jour == 5, 1, 0)

# b. IMPORTANCE OF THE EFFECT

### Standardizing "Citation impact of original article"

### NOG CHECKEN!

# st.impa <- stand(data$Citation.impact..paper..O., option = option)

### Standardizing "Exciting/important effect"

st.exci <- stand(as.numeric(fac.exci), option = option)

### Creating scale

### NOG CHECKEN!

# sc.impo <- (st.impa + st.exci)/2

# c. SURPRISING EFFECT

### Standardizing "Surprising effect" and creating scale

sc.surp <- stand(as.numeric(fac.surp), option = option)

# d. EXPERIENCE AND EXPERTISE OF ORIGINAL TEAM

### Taking the average and then standardizing "Institution prestige of 1st author and senior author"

### NOG CHECKEN!

### VRAGEN: WAT ALS EERSTE AUTEUR OOK SENIOR AUTEUR IS?

# ave.pres <- (data$Institution.prestige..1st.author..O.+data$Institution.prestige..senior.author..O.)/2

# st.pres <- stand(ave.pres, option = option)

### Standardizing "Citation impact of 1st author"

### NOG CHECKEN!

# st.impa.1st <- stand(data$Citation.Impact..1st.author..O., option = option)

### Standardizing "Citation impact of senior author"

### NOG CHECKEN!

# st.impa.sen <- stand(data$Citation.Impact..senior.author..O., option = option)

### Creating scale

# sc.expe1 <- (st.pres + st.impa.1st + st.impa.sen)/3

# e. CHALLENGE OF CONDUCTING REPLICATION

### Standardizing "Perceived expertise required"

st.expe <- stand(as.numeric(fac.expe), option = option)

### Standardizing "Perceived opportunity for expectancy biases"

st.oppo.expe <- stand(as.numeric(fac.oppo.expe), option = option)

### Standardizing "Perceived opportunity for impact of lack of diligence"

st.oppo.dili <- stand(as.numeric(fac.oppo.dili), option = option)

### Create scale

sc.chal <- (st.expe + st.oppo.expe + st.oppo.dili)/3

# f. EXPERIENCE AND EXPERTISE OF REPLICATION TEAM

### Standardizing "Position of senior member of replication team"

st.posi <- stand(as.numeric(fac.posi), option = option)

### Standardizing "Highest degree of replication team's senior member"

st.degr <- stand(as.numeric(fac.degr), option = option)

### Standardizing "Replication team domain expertise"

st.doma <- stand(as.numeric(fac.doma), option = option)

### Standardizing "Replication team method expertise"

st.meth <- stand(as.numeric(fac.meth), option = option)

### Standardizing "Replication team senior member's total publications and total number of peer-reviewed articles"

### WE HAVE TO CHECK WHETHER THIS IS CORRECT

# st.publ <- stand(as.numeric(data$Total.publications..R.), option = option)

### Standardizing "Replication team senior member's total citations"

### WE HAVE TO CHECK WHETHER THIS IS CORRECT

# st.cita <- stand(as.numeric(data$Citations..R.), option = option)

### Create scale

# sc.expe2 <- (st.posi + st.degr + st.doma + st.meth + st.publ + st.cita)/6

# g. SELF-ASSESSED QUALITY OF REPLICATION

### Standardizing "Self-assessed quality of replication"

st.impl <- stand(as.numeric(fac.impl), option = option)

### Standardizing "Self-assessed data collection quality of replication"

st.data <- stand(as.numeric(fac.data), option = option)

### Standardizing "Self-assessed replication similarity to original"

st.repl <- stand(as.numeric(fac.repl), option = option)

### Standardizing "Self-assessed difficulty of implementation"

st.diff <- stand(as.numeric(fac.diff), option = option)

### Create a scale

sc.self <- (st.impl + st.data + st.repl + st.diff)/4

###########################################

### Transforming correlations to Fisher ###

###########################################

ri.o <- data$r..O.

ri.r <- data$r..R.

N.o <- data$df2..O.+2

N.r <- data$df2..R.+2

### Transform to Fisher's z

fis.o <- 0.5\*log((1 + ri.o) / (1 - ri.o))

fis.r <- 0.5\*log((1 + ri.r) / (1 - ri.r))

### Difference in Fisher's z scores

yi <- numeric()

for(i in 1:length(fis.o)) {

if(is.na(fis.o[i]) == TRUE | is.na(fis.r[i]) == TRUE) { yi[i] <- NA }

else if(fis.o[i] < 0 & fis.r[i] < 0) { yi[i] <- fis.o[i]\*-1-fis.r[i]\*-1 }

else if(fis.o[i] < 0 & fis.r[i] > 0) { yi[i] <- fis.o[i]\*-1+fis.r[i] }

else { yi[i] <- fis.o[i]-fis.r[i] }

}

### Standard errors original and replication study

sei.o <- sqrt(1/(N.o-3))

sei.r <- sqrt(1/(N.r-3))

### p-values original and replication study

pval.o <- pnorm(fis.o, sd = sei.o, lower.tail = FALSE)

pval.r <- pnorm(fis.r, sd = sei.r, lower.tail = FALSE)

### Standard error of difference score

sei <- sqrt(1/(N.o-3) + 1/(N.r-3))

#######################################

### Select studies for the analyses ###

#######################################

# df <- data.frame(yi, sei, sei.o, d.JEP, d.PSCog, d.PSSoc, d.PSOth, sc.impo, sc.surp, sc.expe1, sc.chal, sc.expe2, sc.self)

df <- data.frame(stat = as.character(data$Test.Statistic..O.), df1 = data$df1..O., yi, sei, fis.o, sei.o, pval.o, fis.r, sei.r, pval.r, d.JEP, d.PSCog, d.PSSoc, d.PSOth, sc.surp, sc.self)

df <- df[-149, ] ### Remove duplicate

### Select: F(df1 = 1, df2), t, r, and z

sub <- subset(df, (df$stat == "F" & df$df1 == 1) | df$stat == "t")

### Remove rows when NA on yi

final <- sub[!is.na(sub$yi) & !is.na(sub$sei), ]

#######################################

### Correlations between moderators ###

#######################################

### Issue with NAs has to be solved

# cor(final$sc.imp, final$sc.surp, final$sc.expe1, final$sc.chall, final$sc.expe2, final$sc.self)

# cor(final$sc.surp, final$sc.self)

#####################

### Meta-analyses ###

#####################

### Load metafor package

library(metafor)

### Meta-analysis of null model

res <- rma(yi = final$yi, sei = final$sei, method = "REML")

res

# png("C:/Users/S787802/Desktop/Funnel RPP.png", width = 900, height = 900, res = 200, pointsize = 5)

funnel(res, main = "Funnel plot based on difference original and replication study")

# dev.off()

### Meta-analysis of only original studies

res <- rma(yi = final$fis.o, sei = final$sei.o, method = "REML")

# png("C:/Users/S787802/Desktop/Funnel original RPP.png", width = 900, height = 900, res = 200, pointsize = 5)

funnel(res, main = "Funnel plot based on original studies")

# dev.off()

### Meta-analysis of only original studies with se in original study as moderator

rma(yi = final$fis.o, sei = final$sei.o, mods = ~ final$sei.o, method = "REML")

### Meta-analysis of only replication studies

res <- rma(yi = final$fis.r, sei = final$sei.r, method = "REML")

# png("C:/Users/S787802/Desktop/Funnel replication RPP.png", width = 900, height = 900, res = 200, pointsize = 5)

funnel(res, main = "Funnel plot based on replication studies")

# dev.off()

### Meta-analysis of only replication studies with se in replication study as moderator

rma(yi = final$fis.r, sei = final$sei.r, mods = ~ final$sei.r, method = "REML")

### Meta-analysis with a. PUBLISHING JOURNAL AND SUBDISCIPLINE as moderator

rma(yi = final$yi, sei = final$sei, mods = ~ final$d.JEP + final$d.PSCog + final$d.PSSoc + final$d.PSOth, method = "REML")

### Meta-analysis with sei.o as moderator

rma(yi = final$yi, sei = final$sei, mods = ~ final$sei.o, method = "REML")

# ### Meta-analysis with a. PUBLISHING JOURNAL AND SUBDISCIPLINE and sei.o as moderators

# rma(yi = final$yi, sei = final$sei, mods = ~ final$d.JEP + final$d.PSCog + final$d.PSSoc + final$d.PSOth + final$sei.o, method = "REML")

#

# ### Meta-analysis with a. PUBLISHING JOURNAL AND SUBDISCIPLINE, sei.o, and its interaction as moderators

# rma(yi = final$yi, sei = final$sei, mods = ~ final$d.JEP + final$d.PSCog + final$d.PSSoc + final$d.PSOth + final$sei.o, final$d.JEP\*final$sei.o + final$d.PSCog\*final$sei.o + final$d.PSSoc\*final$sei.o + final$d.PSoth\*final$sei.o, method = "REML")

#

# ### Meta-analysis with b. IMPORTANCE OF THE EFFECT as moderator

# # rma(yi = yi, sei = sei, mods = ~ sc.impo, method = "REML")

#

# ### Meta-analysis with c. SURPRISING EFFECT as moderator

# rma(yi = yi, sei = sei, mods = ~ sc.surp, method = "REML")

#

# ### Meta-analysis with d. EXPERIENCE AND EXPERTISE OF ORIGINAL TEAM as moderator

# # rma(yi = yi, sei = sei, mods = ~ sc.expe1, method = "REML")

#

# ### Meta-analysis with e. CHALLENGE OF CONDUCTING REPLICATION as moderator

# # rma(yi = yi, sei = sei, mods = ~ sc.chal, method = "REML")

#

# ### Meta-analysis with f. EXPERIENCE AND EXPERTISE OF REPLICATION TEAM

# # rma(yi = yi, sei = sei, mods = ~ sc.expe2, method = "REML")

#

# ### Meta-analysis with g. SELF-ASSESSED QUALITY OF REPLICATION

# rma(yi = yi, sei = sei, mods = ~ sc.self, method = "REML")

### Meta-analyses per pair

in.ci <- es.meta <- se.meta <- ci.lb.meta <- ci.ub.meta <- pval.meta <- numeric()

for(i in 1:length(final$fis.o)) {

tmp <- rma(yi = c(final$fis.o[i], final$fis.r[i]), sei = c(final$sei.o[i], final$sei.r[i]), method = "FE")

es.meta[i] <- tmp$b[1]

se.meta[i] <- tmp$se

ci.lb.meta[i] <- tmp$ci.lb

ci.ub.meta[i] <- tmp$ci.ub

pval.meta[i] <- tmp$pval

if(tmp$pval < 0.05) { in.ci[i] <- 1

} else { in.ci[i] <- 0 }

}

### How often is the null hypotheses rejected in the meta-analysis

sum(in.ci)/length(in.ci)

### Create data frame

tab <- data.frame(rows = row.names(final), fis.o = final$fis.o, sei.o = final$sei.o, pval.o = final$pval.o, fis.r = final$fis.r, sei.r = final$sei.r,

pval.r = final$pval.r, diff = final$yi, es.meta = es.meta, se.meta = se.meta, ci.lb.meta = ci.lb.meta, ci.ub.meta = ci.ub.meta, pval.meta = pval.meta)

# library(xlsx)

# wb <- createWorkbook()

# sh1 <- createSheet(wb, 'sheet1')

# addDataFrame(tab, sh1)

# saveWorkbook(wb, "C:/Users/S787802/Desktop/tab.xls")

### Check how often effect size original study is within CI of meta-analysis

in.ci.meta <- numeric()

for(i in 1:length(final$fis.o)) {

if(final$fis.o[i] > ci.lb.meta[i] & final$fis.o[i] < ci.ub.meta[i]) {

in.ci.meta[i] <- 1

} else { in.ci.meta[i] <- 0 }

}

sum(in.ci.meta)/length(in.ci.meta)

References

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), 534.

van Assen, M. A. L. M., van Aert, R. C. M., & Wicherts, J. M. (2014, November 17). Meta-Analysis Using Effect Size Distributions of Only Statistically Significant Studies. Psychological Methods. Advance online publication. http://dx.doi.org/10.1037/met0000025