**Statistical analyses**

The reproducibility of psychological science was evaluated using significance and *p*-values, effect sizes, subjective 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, using Fisher’s (1925) method we tested the hypothesis that these studies had ‘no evidential value’, i.e. the null-hypothesis of zero-effect holds for all these studies. The hypothesis that the proportions of statistically significant results are equal was tested using the McNemar test for paired nominal data, and a confidence interval of the reproducibility parameter was calculated. Second, we compared the central tendency of the distribution of *p*-values of original and replication studies using the Wilcoxon signed-rank test and the *t*-test for dependent samples. 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 whenever possible. 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 [A3]). To be able to compare and analyze correlations across study-pairs, the original study’s effect size was coded as positive; the replication study’s effect size was coded as negative if and only if the replication study’s effect was opposite to that of the original study.

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). Fourth, 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. We carried out this test on the subset, further called MA, of 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*). 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 [A4] for computational details on the test).

**Meta-analysis**. Meta-analyses were conducted on Fisher-transformed correlations for all study-pairs in subset MA, and computed the number of times the CI of the meta-analysis contained 0, which amounts to the number of times the null hypothesis of no effect is rejected according to the meta-analysis.

**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, both of studies in set MA. We ran three models; one without any predictor, one with studies’ standard error as predictor, and one with standard error and journal-field as predictor. Standard error was added to study small-study effects. A positive effect of standard error on effect size indicates that studies’ effect sizes are positively associated with their sample sizes. The results of this test, also known as Eggers’test, is often used as test of publication bias. Journal-field is a categorical variable with categories JPSP-social (= reference category), JEP:LMC-cognitive, PSCI-social, PSCI-cognitive, and PSCI-other.

**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).

**Analysis of moderators**. We correlated six moderators (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) with correlations of original studies, replication studies, and meta-analysis across study-pairs.

**Results**

*Preliminary analyses*. The input of our analyses were the *p*-values and effect sizes of both original and replication study, direction of the test (column BU)[[1]](#footnote-1), and whether the sign of both studies’ effects was the same or opposite (column BT). 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. The *p*-values used for our analyses can be found in columns DH (original study) and DR (replication). We ended up with 100 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 | 2 | 1 |
|  | Significant | 62 | 35 |

*Statistical significance and p-values (see [A2] for details)*. Table 1 shows the statistical significance of original and replication studies. Of the original studies, 92.30769% was statistically significant, as opposed to 35.35354% (CI = [25%, 45%]) of replication studies, which corresponds to a significant change (McNemar test, *χ2*(1) = 59.06349, *p* < .001). Proportions of statistical significance of original and replication studies for the three journals JPSP, JEP, PS (Cognitive), PS (Social), PS (other) were .96875 and .21875, .9643 and .4643, 1 and .7, .9583 and .333, 1 and .1667 respectively. Of 97 significant original studies, 36.08% was statistically significant in the replication study. The hypothesis that all 64 statistically non-significant replication studies came from a population of true negatives can be rejected (*χ2*(128) = 155.8262, *p* = 0.04765856). The density and cumulative *p*-value distributions of original and replication studies are presented in Figure 1. The means of the two *p*-valuedistributions (.02828 and .3023) were different from each other (*t*(98) = -8.2165, *p* < .001; W = 2406, *p* < .001). Quantiles are .0004249, .006891, .02333 for the original, and .007754, .1982, .5365 for the replication studies.

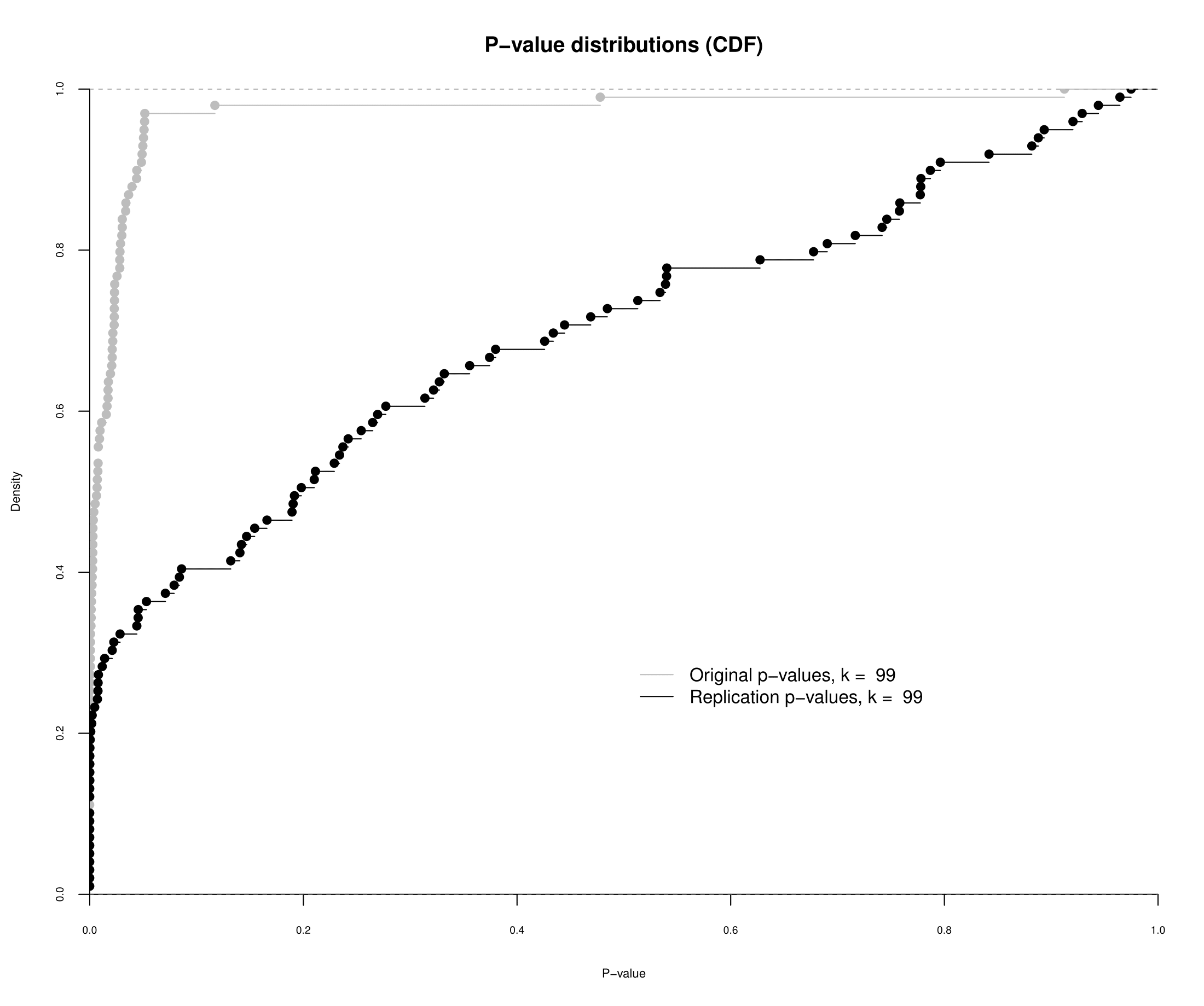


Figure 1a: Cumulative *p*-value distributions of original and replication studies.

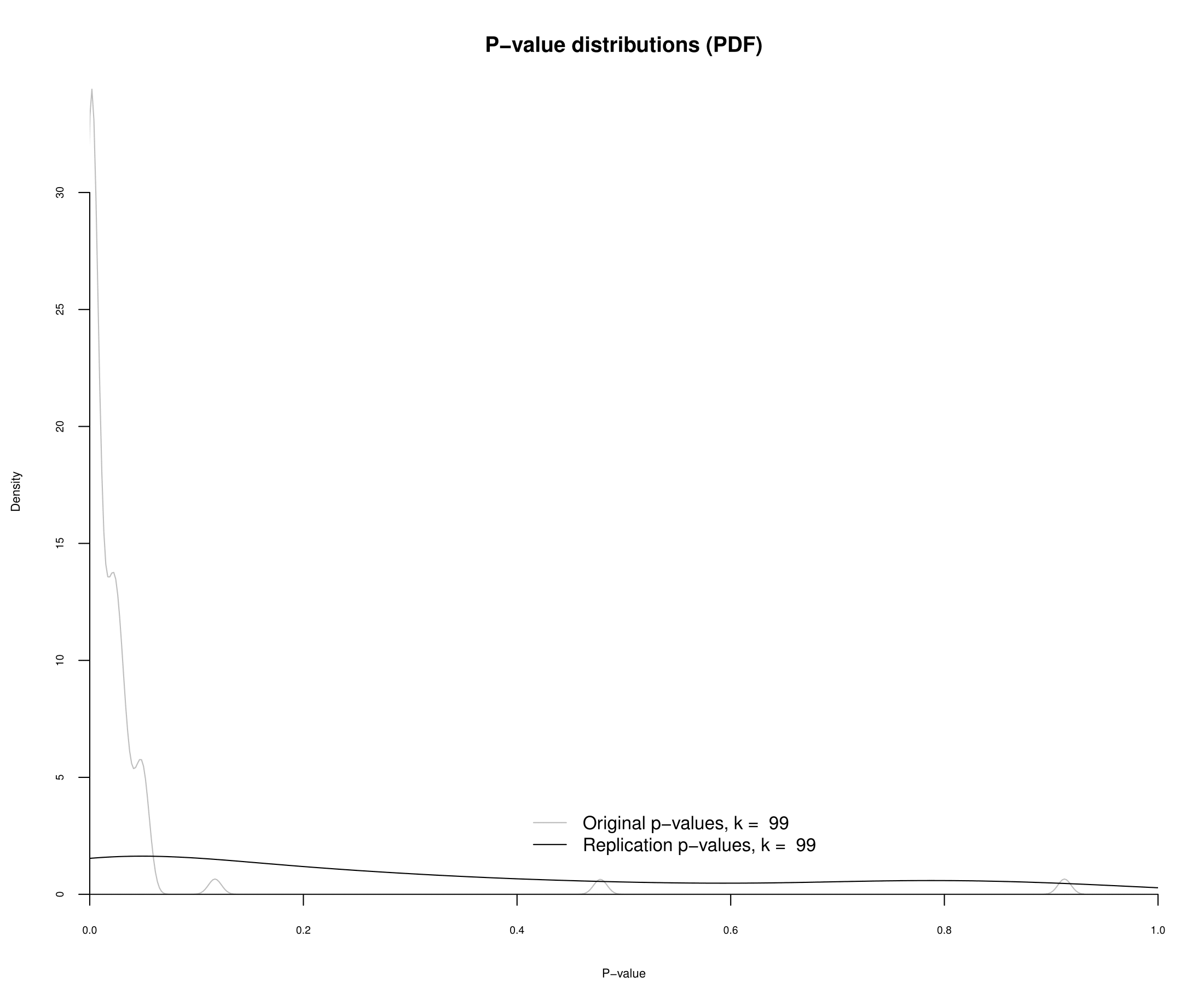


Figure 1a: Density *p*-value distributions of original and replication studies.

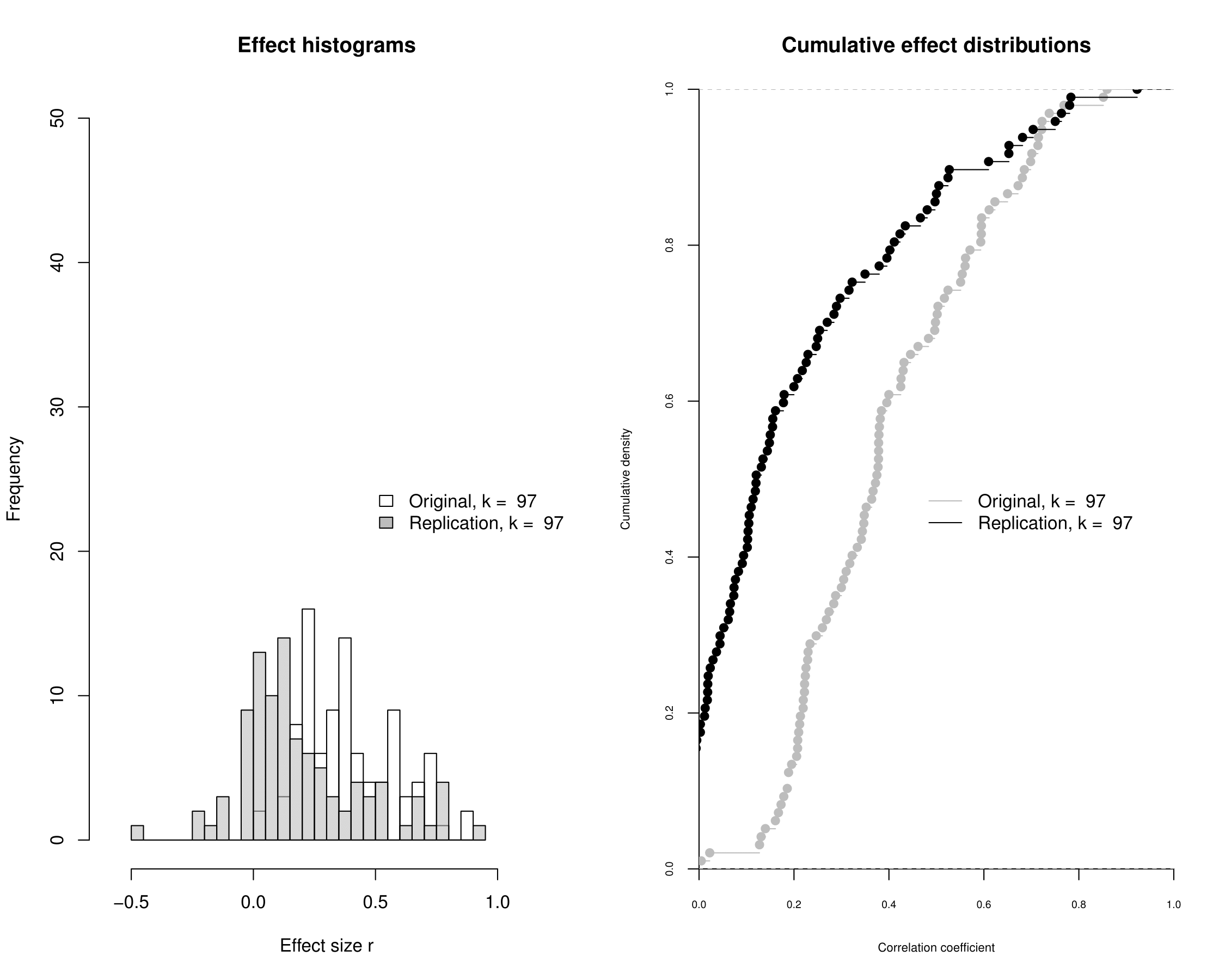


Figure 2: Histograms (left) and cumulative distribution functions of effect sizes of original and replication studies.

*Effect sizes (see [A5] for details)*. For 97 study pairs effect size correlations could be computed. Figure 2 (left) shows the distribution of effect sizes of original and replication studies, and the corresponding cumulative distribution functions (right). Spearman’s correlation between the effect sizes equals .51067. The mean effect sizes of both distributions (*M* = .3962 [*SD* = .1928]; *M* = .1979 [*SD* =.254985]) were different from each other (*t*(96) = 9.3317, *p* < .001; *W* = 7132, *p* < .001). Of those 99 studies that reported an(y) effect size in both original and replication study, 82 reported a stronger effect size in the original study (.8283%; *p* < .001, binomial test). For the subset of 69 studies where the standard error of the correlation could be computed, it was expected that 78.3% of CIs of the replication study contained the effect size of the original study; however, only 42.0% (29 out of 69) of CIs contained the original effect size (*p* < .001). Figure 3 depicts effect sizes of study-pairs of which correlations could be calculated, and codes significance of effect sizes as well.

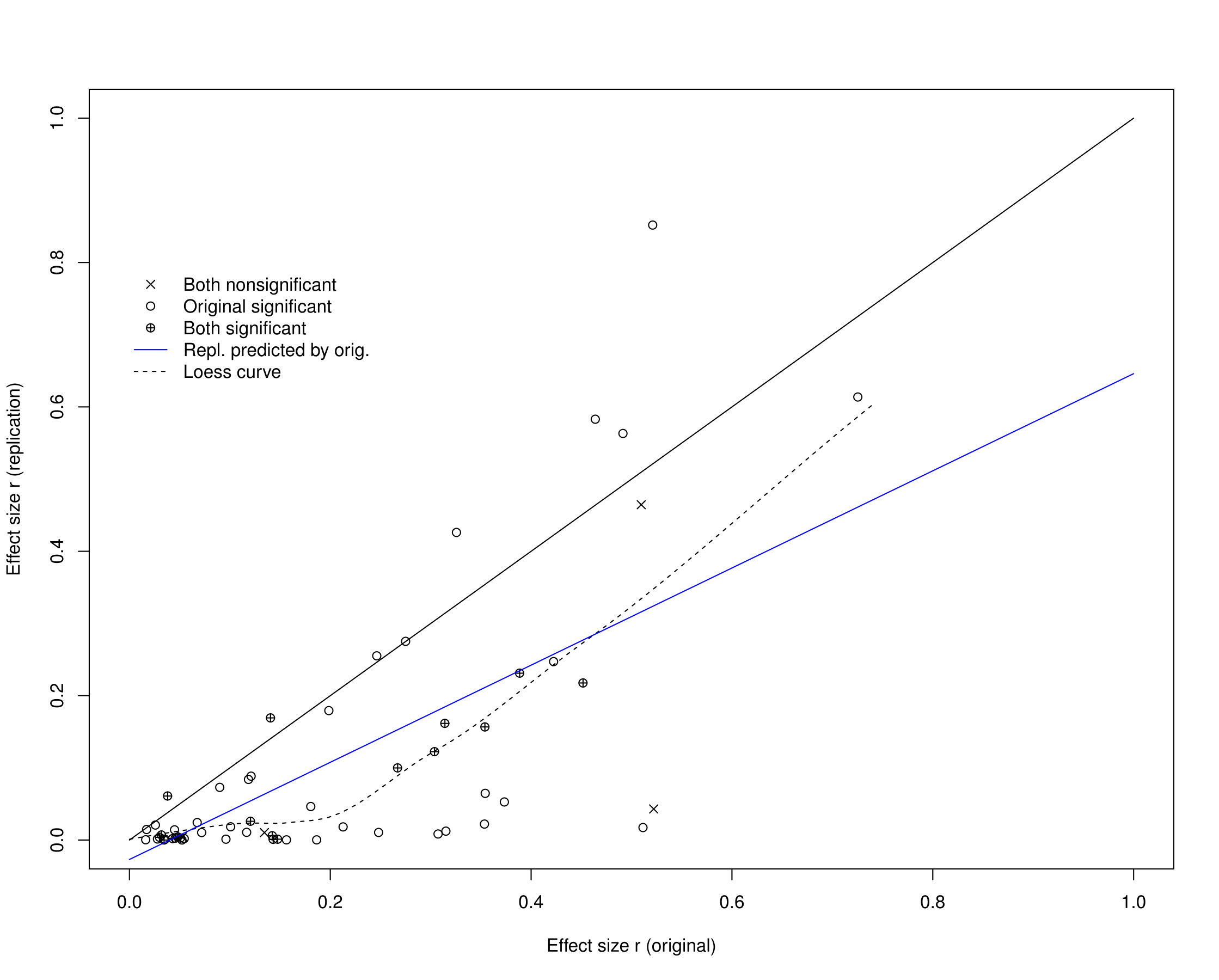
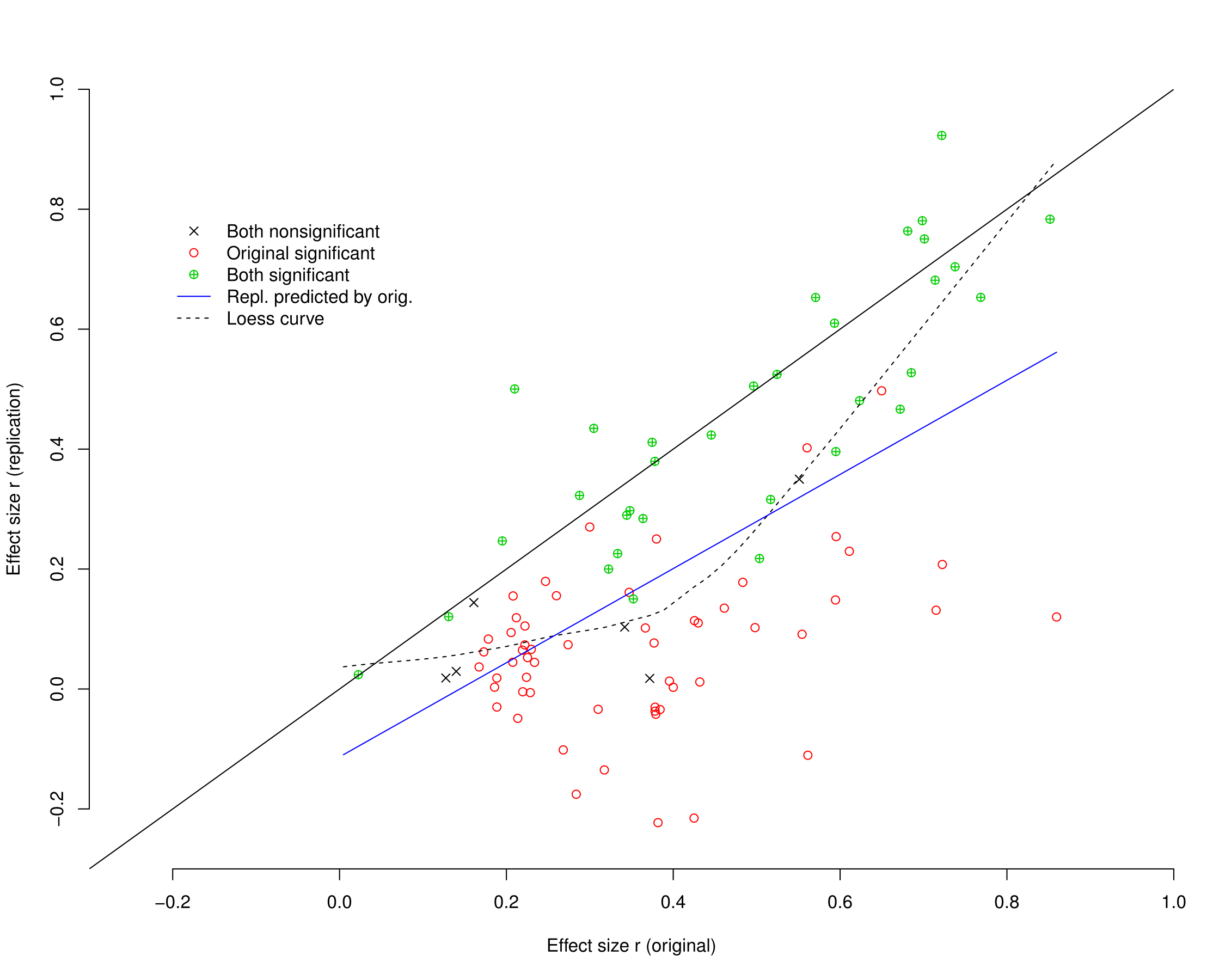


Figure 3: Explained variances (squared correlations) of both original and replication study, coded by statistical significance. Identical values are indicated by the black diagonal line.

*Meta-analysis (see [A6] for details)*. For 69 study-pairs a meta-analysis could be conducted. In 50 out of 69 pairs the null-hypothesis of no effect was rejected (72.5%).

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

*Meta-analysis of all original study effects, and of all replication study effects (see [A7] for details).*

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.

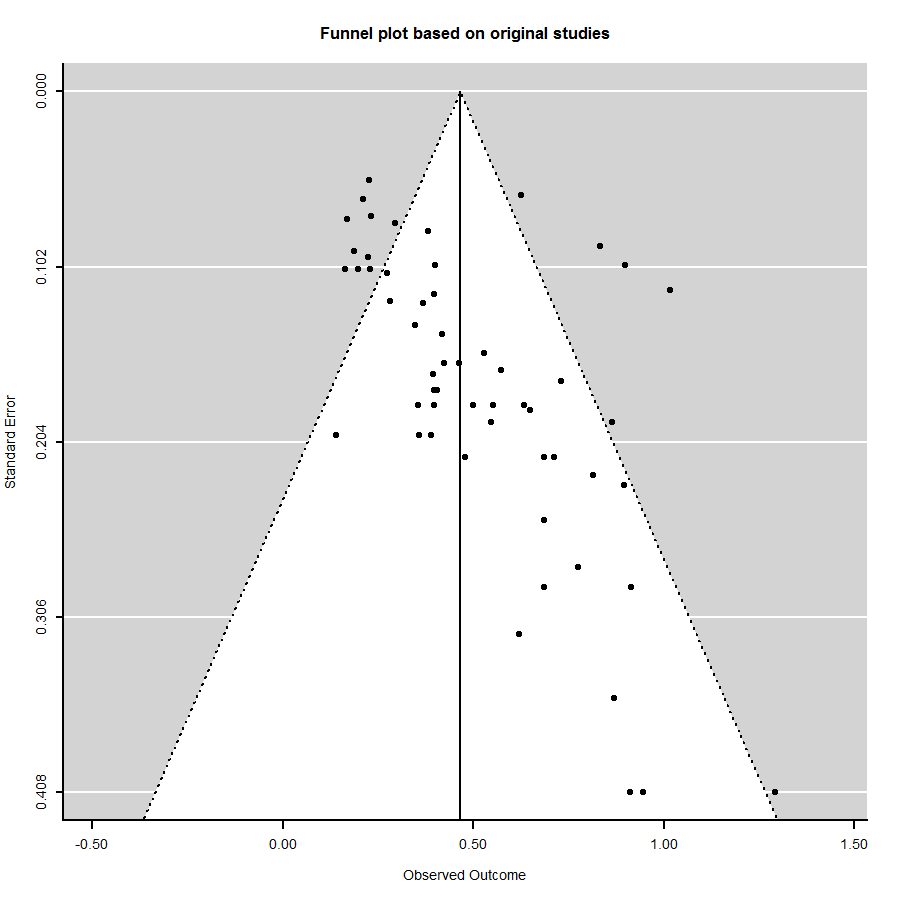


Figure 4: Funnel plot of the meta-analysis on the original study’s effect size.

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.

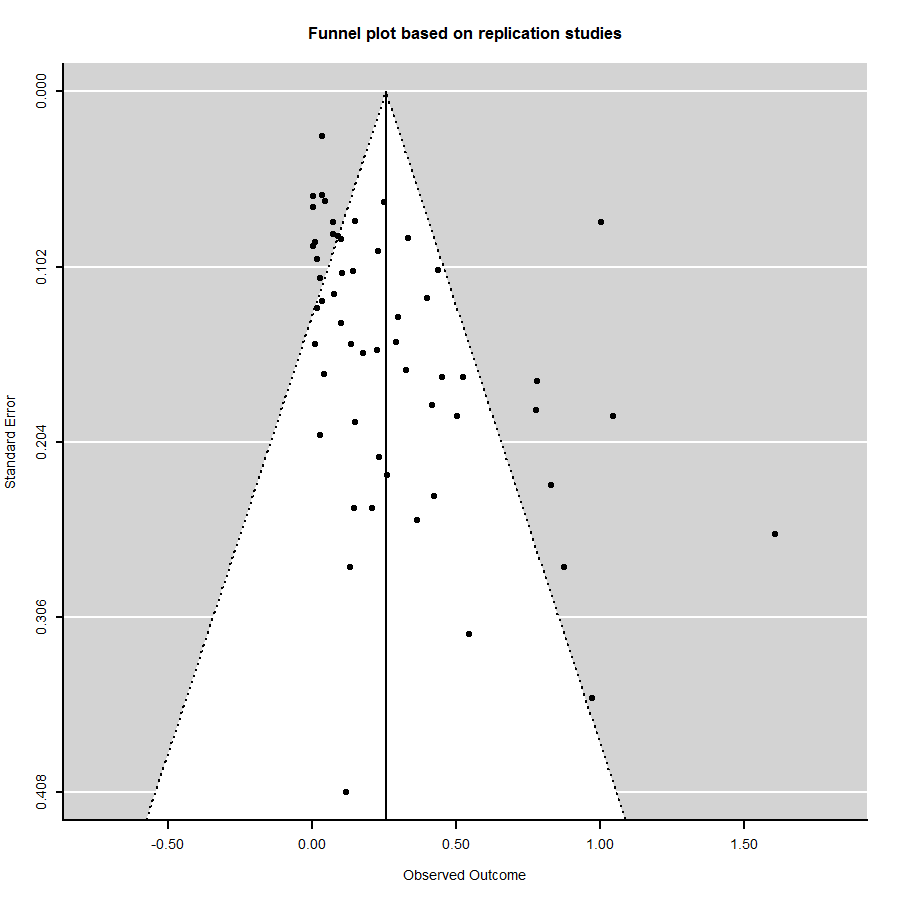


Figure 5: Funnel plot of the meta-analysis on the replication study’s effect size.

*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 \*\*\*.

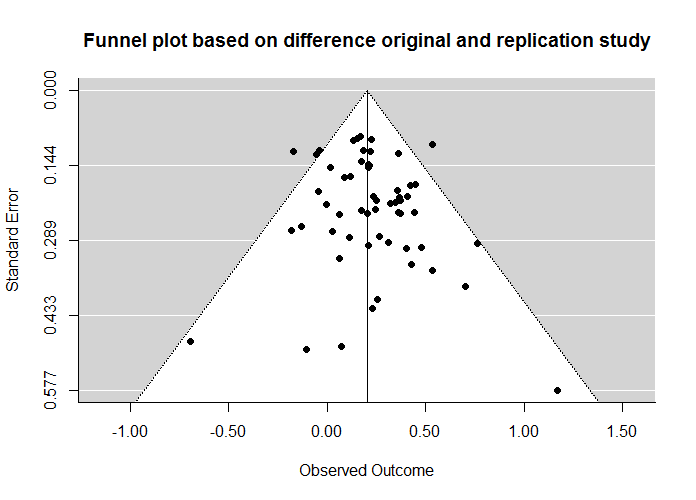


Figure 6: Funnel plot of meta-analysis on difference in effect size (original – replication).

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

**Supplementary materials**

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

*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, RCM van Aert, MALM van Assen

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)

}

*Remarks p-values*

The *p*-values of study 59 could not be recovered.

**Effect sizes**. We transformed all effect sizes into correlation coefficients whenever possible. 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 [A3]). To be able to compare and analyze correlations across study-pairs, the original study’s effect size was coded as positive; the replication study’s effect size was coded as negative if and only if the replication study’s effect was opposite to that of the original study.

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 [A4] for computational details on the test).

**[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**

Whenever 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.

The correlation effect sizes can be found in columns DJ and DV of the master data file. No correlation could be computed for study 69.

**[A4] Calculation of expected coverage of original effect size by replication CI**

One statistic to evaluate reproducibility is the probability that the original study’s effect size is covered by the replication study’s confidence interval. If *α* = .05, and we assume that both studies are sampled from a population with the same true effect size, then this probability is a function of both study’s effect size. When both studies have equal sample size, this probability equals 83.4% (Cumming, 2013). However, this probability can be any number between 0 (if the replication study has a much larger sample size) and 1 (if the original study has a much larger sample size).

The program below calculates the expected proportion of coverage across study pairs, by summing the study pairs’ probabilities. For each study, the probability of overlap is calculated using the Fisher transformed effect size and its standard error. Since the standard error can only be calculated for test statistics *t*, *F*(1,df), and *r*, we can only use this statistic for study pairs who used these tests.

overlap <- numeric()

points <- 1000000

p <- 1:points/(points+1) # uniform probability density based on equally distributed points

for (i in 1:length(final$N.r)) {

zu <- qnorm(p,0,1/sqrt(final$N.r[i]-3)) + qnorm(.975)/sqrt(final$N.r[i]-3)

# zu gives upper bound of Fisher transformed effect size for each possible point in the

# probability density

zl <- zu - 2\*qnorm(.975)/sqrt(final$N.r[i]-3)

# zl gives lower bound of Fisher transformed effect size for each possible point in the

# probability density

overlap[i] <- mean(pnorm(zu,0,1/sqrt(final$N.o[i]-3))) - mean(pnorm(zl,0,1/sqrt(final$N.o[i]-3)))

# overlap gives the probability of coverage as the average proportion that the original effect

# size is lower than the upper bound minus the average proportion that the original effect

# is larger than the lower bound

}

overlap

mean(overlap)

**[A5] Analyses of effect sizes**

The code for the first two 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 third test comparing effect sizes (‘which is stronger?’) was carried out in three steps. First, the two columns BJ and BV in the master data file containing correlations were compared using the IF function in excel. Second, for those studies where no correlation effect size could be computed, columns BG and BZ with the ‘raw’ effect sizes were compared manually. Finally, the frequency of studies where the original effect size exceeded the replication effect size (*f*) and the total number of comparisons (*n*) were entered in the binomial test:

binom.test(f, n, 0.5, "two.sided", 0.95)

The fourth and last test compared the observed proportion of study-pairs in which the effect of the original study was in the confidence interval of the effect of the replication study with the expected proportion using a goodness-of-fit *χ2­*-test. Supplement [A4] provides the code for calculating the expected proportion. The code for calculating the observed proportion can be found in supplement [A6]. The observed frequency *f*, expected proportion *p*, and number of comparisons *n* was entered in the binomial test:

binom.test(f, n, p, "two.sided", 0.95)

The number of comparisons *n* equals the number of studies in which the effect was tested using *r*, *t*, or *F*(1,df).

**[A6] Meta-analyses on effect sizes of each study-pair**

The meta-analyses were conducted on Fisher-transformed correlations for all study-pairs in subset MA, i.e. 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 of Fisher-transformed correlations were computed using , which assumes tests of one correlation or an independent sample *t*-test (but not a dependent sample *t*-test).

The results of all individual meta-analyses are reported after the code.

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

### Meta-analyses per pair ###

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

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

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 }

}

sum(in.ci)/length(in.ci) # Proportion of times the null hypothesis of no effect is rejected

### Create data frame

tab <- data.frame(ID = final$ID, fis.o = final$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)

### 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] <- TRUE

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

}

sum(in.ci.meta)/length(in.ci.meta) # Proportion of times the original study is within the CI of meta-analysis

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

### How often is original study within CI of replication ###

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

### Create confidence interval for replications

ci.lb <- final$fis.r-qnorm(.975)\*final$sei.r

ci.ub <- final$fis.r+qnorm(.975)\*final$sei.r

in.ci <- numeric()

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

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

in.ci[i] <- TRUE

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

}

sum(in.ci)/length(in.ci) # Proportion of times the original study is within the CI of the replication

Table with results of meta-analyses on study pairs:

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

1. Columns refer to the Master Data file; \*\*\*LINK. [↑](#footnote-ref-1)