**RPP: Frequentist Committee Intermediate Report**

**Members:** Marcel van Assen (MA), Denny Borsboom (DB), Deborah Mayo (DM), Frank Renkewitz (FR), Jelte Wicherts (JW), Fred Hasselman (FH, chair)

**Activities:**

***Startup***

The committee started with an invitation to members in which they were encouraged to provide ideas for analysis strategies, point out potential problems, and/or comment on text provided by other members.

Not every member responded to the request, but among those who did there was a majority in favor of a strategy that involved study-by-study evaluation of replication success, versus analyses on effect sizes, taking the studies as a sample.

The document is attached as *Appendix 1*

***Intermediate Report:***

After the request was made to draft an intermediate report, several members took up the invitation to exchange ideas / comments via e-mail.

At least two members expressed the wish to include the concept of Severity (Mayo & Spanos, 2011) in the frequentist evaluation.

It soon became clear that the available data needed closer scrutiny than expected in advance; this was judged a priority over developing sophisticated modeling strategies.

A report was drafted s (this document) based on contributions by FH, DB, MA and great efforts to get the correct data available by other teams.

Part 1 concerns preliminary analyses, part 2 contains recommendations, suggestions on how to proceed and issues that need to be addressed.

1. **Some preliminary ‘Frequentist’ answers**
   1. Descriptives
   2. It takes 2 (Effect Sizes) to tango: 92% studies correct by QDA
   3. Exploratory Meta-Analyses
   4. Towards Severity Assessment
2. **Recommendations, Suggestions**

**1. Some preliminary ‘Frequentist’ answers**

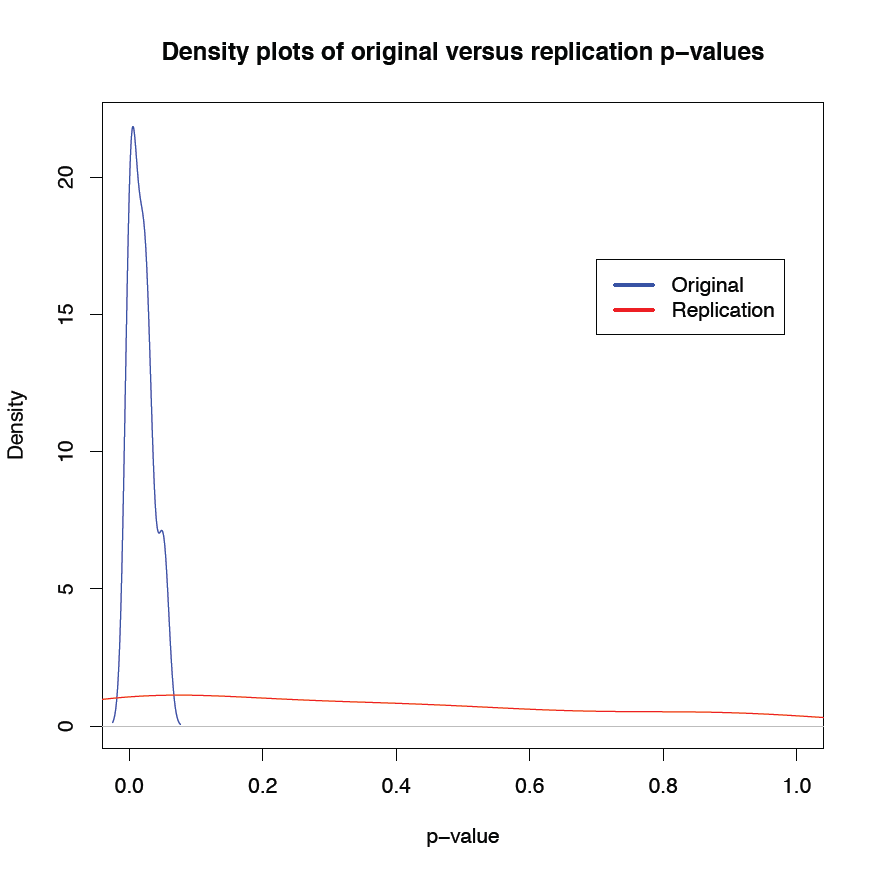
*1.1. Descriptive information about preliminary findings in the Reproducibility Project*

*Data*

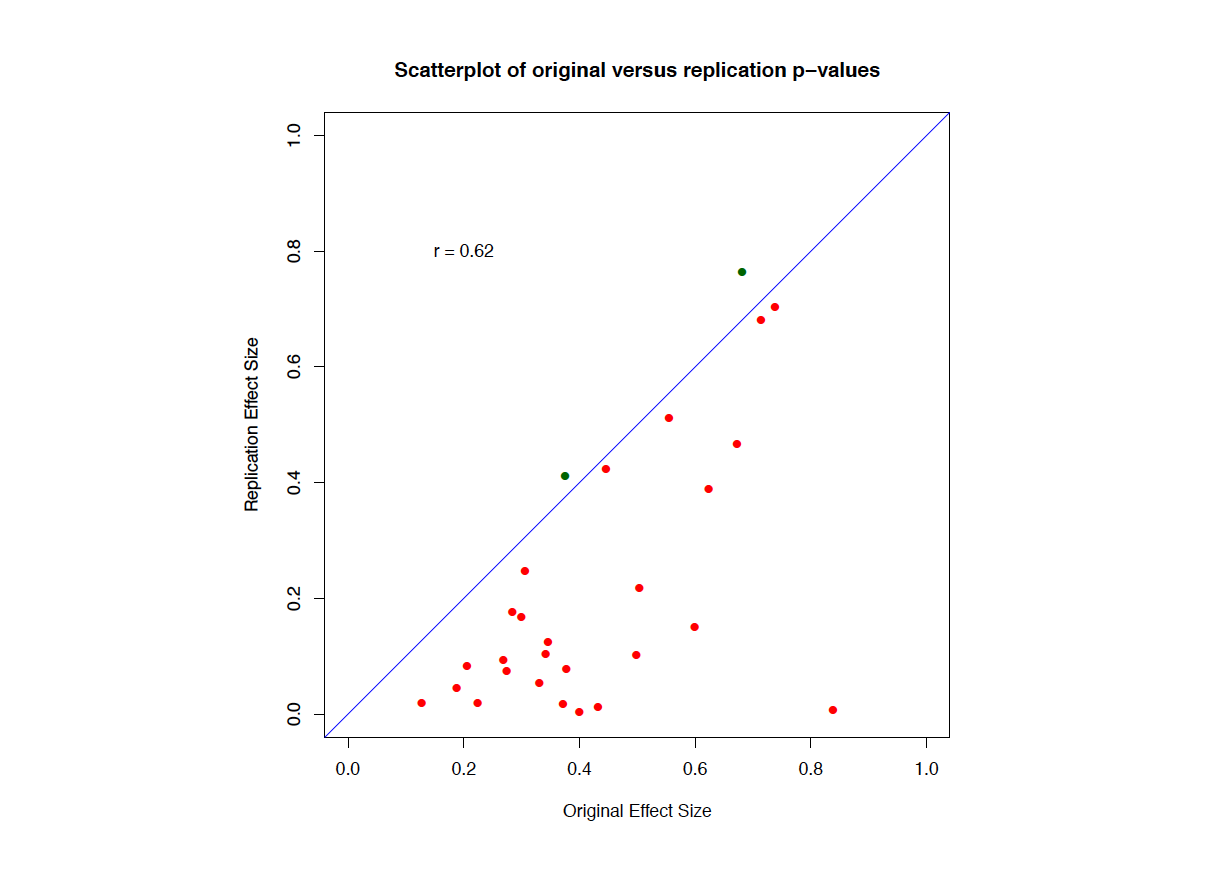
For this report, we used the 28-30 studies for which complete data were available (i.e., test statistics and effect sizes for both the original study and the replication were available and information could be verified). All effect sizes measures were converted to r’s. The data on which the below analyses were carried out were archived in the file *RPPdata\_cast.dat[[1]](#footnote-1)*. The R-script *RPP\_CastData.R* produces the data below and stores figures in a separate document entitled *RPP\_Figures\_30.pdf*.

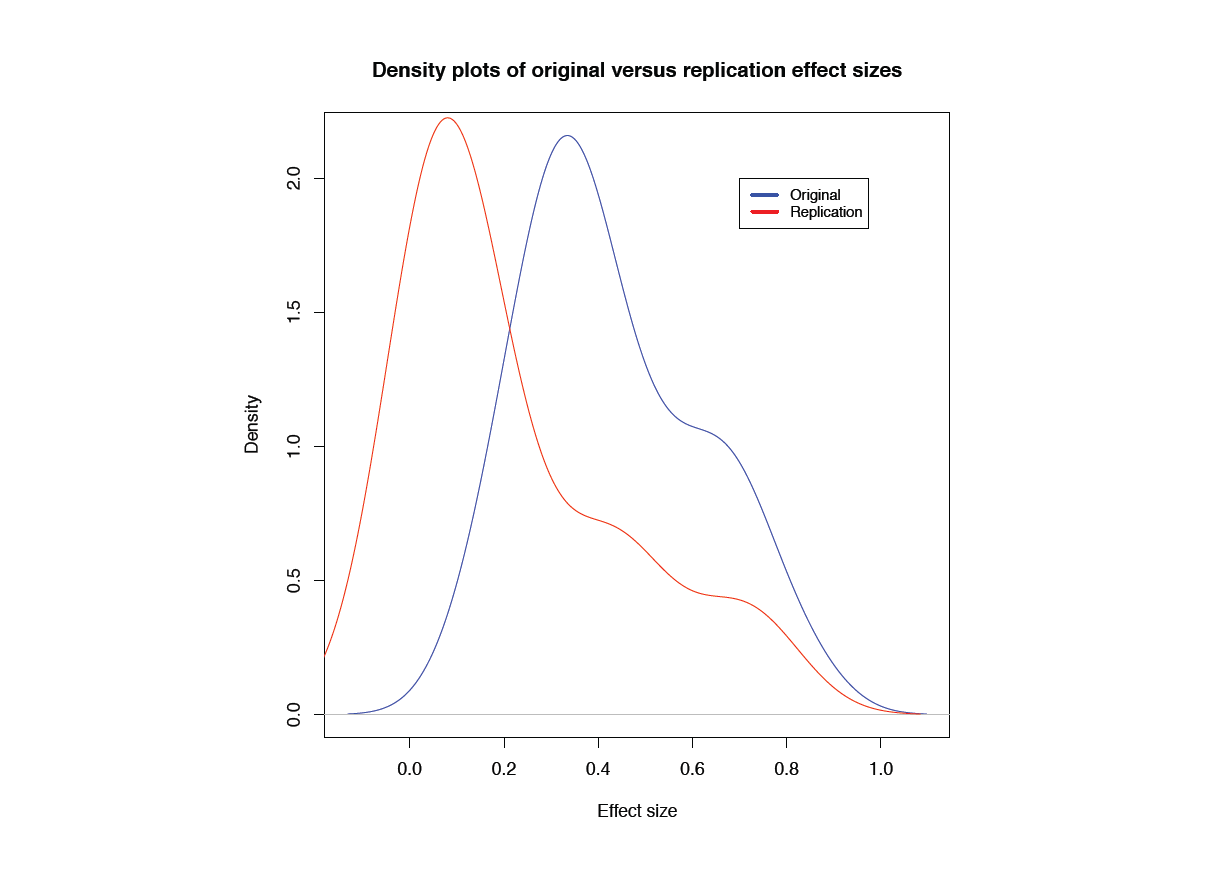
*Findings*

Eleven out of 28 studies (39.3%) replicate according to the criterion of statistical significance at the alpha=0.05 level and magnitude (assuming that direction was not important for the prediction). The distribution of p-values in the original studies is centered entirely below .05 (M=.02; SD=.02), with 3 non-significant “findings” (after all, the studies were selected for replication); the distribution of p-values in the replication studies is shifted to the right and has a much larger variance (M=.35; SD=.34). This is shown in the below density plots:



The scatterplot shows the correlation between Original and Replicated effect size is high: .62



However, even though the original and replication effect sizes are highly correlated, the replication effect sizes are almost invariably smaller than the original effect sizes (barring two studies, indicated in green in the above figure). The mean original effect size equals r=.43 (SD=.19). The mean replication effect size equals r=.22 (SD=.23) So when an effect is larger in the original study, it is also likely to be larger in the replication, but the average effect size in a replication study is likely to be smaller than the original effect. The below density plots illustrate this phenomenon:

*1.2 It Takes Two (Effect Sizes) to Tango: 92% studies correct by QDA*

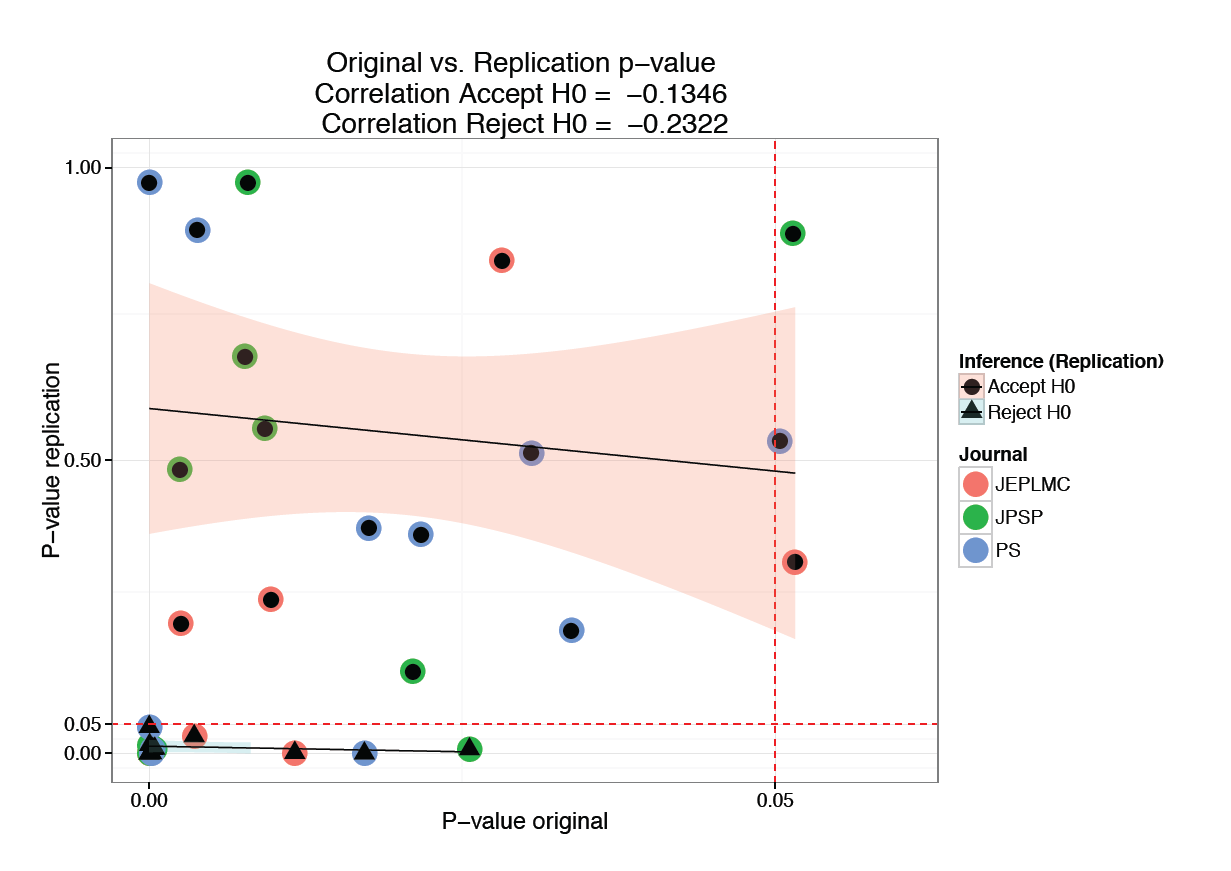
The descriptive data appear to show a clear pattern: Phenomena that were originally evidenced by large effects and/or low p-values, successfully replicate. Other phenomena do not replicate.

To provide some more details, the figures below show a breakdown of the descriptive results with respect to the journal in which the original study was published and the inference about the original claim based on the replication.

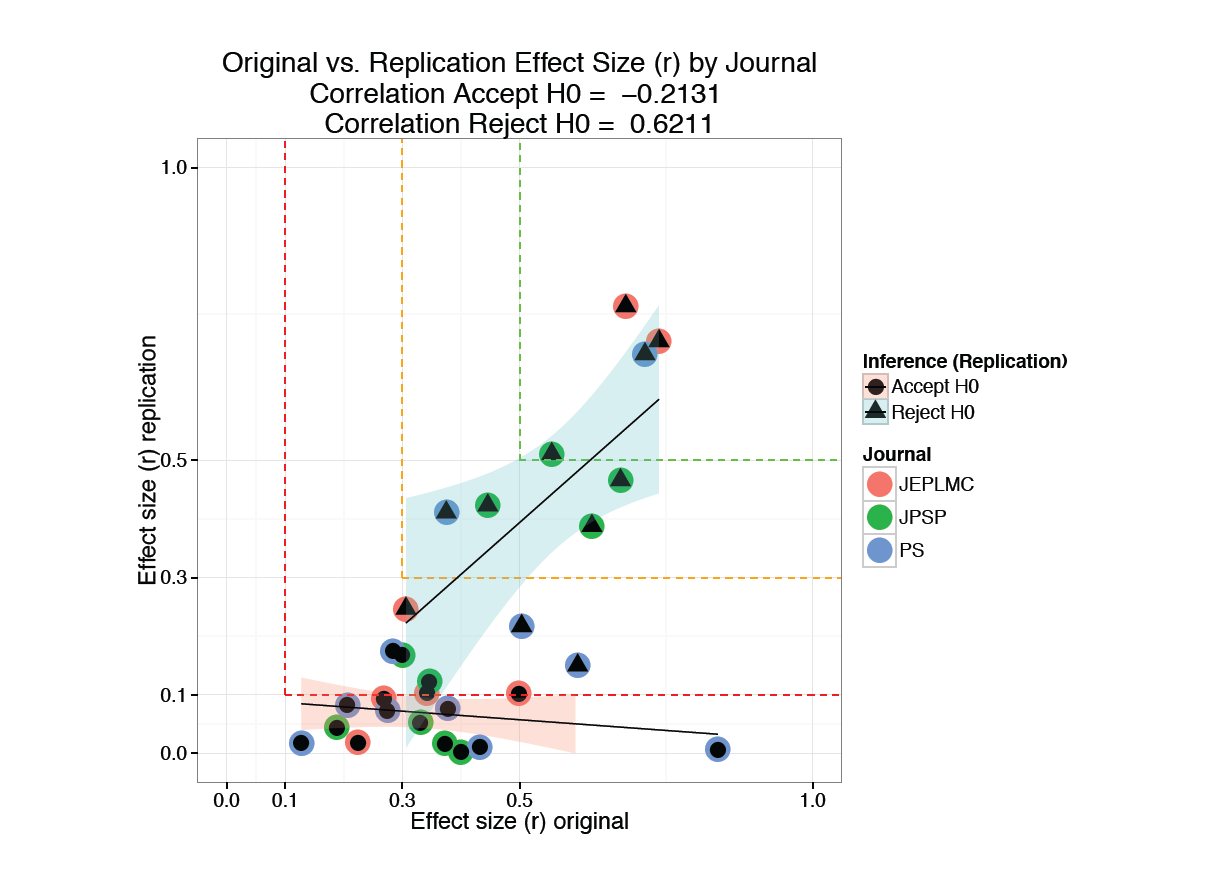
From the perspective of 28 replication studies, the ‘reject H0’ inferences, or, the replication success, broken down by Peer-Reviewed Journal is listed in Table 1.1.

|  |  |  |  |
| --- | --- | --- | --- |
| *Table 1.1*  *Replication ‘succes’ by Peer-Reviewed Journal* | | | |
| **Journal** | **Accept H0** | **Reject H0** | **Replication %** |
| *JEPLMC* | 4 | 3 | 43% |
| *JPSP* | 6 | 4 | 40% |
| *PS* | 7 | 4 | 36% |

Whether this division holds when the number of replicated studies increases remains to be seen.



**Figure 1.1.** Original versus replicated p-values associated with the statistical test of the effect of interest. The trend line was fitted on the 28 studies in the figure. Two studies with incomplete pairwise information were omitted.

**

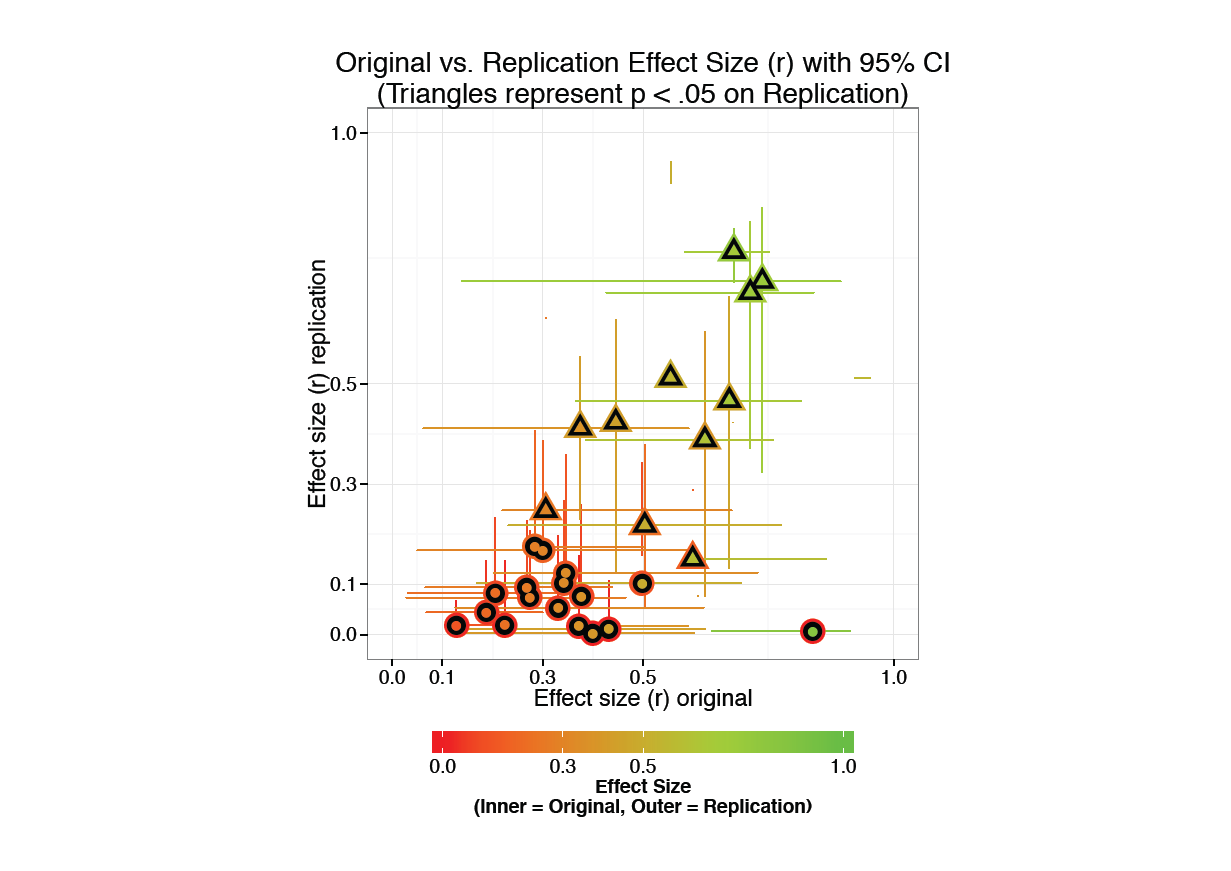
**Figure 1.2.**Original versus replicated magnitude (effect size *r*) associated with the effect of interest. The trend line was fitted on the 28 studies in the figure; two studies with incomplete pairwise information were omitted.

When the original and replicated effect sizes (transformed to *r*) are plotted against each other (Figure 1.2) conjecture could be posited about studies that will successfully replicate:

**Conjecture 1:** *If the original is effect size was r > .4 then a sufficiently powered replication study will evidence the original phenomenon by rejecting H0 in an appropriate statistical test, adjusting the estimate of the magnitude of the effect to 50% of the original estimate.*

A further comparison of the effect sizes concerns the Effect Size Confidence Interval. The CIs in Figure 1.3 were calculated using R package MBESS: The CI was constructed around the reported test statistic, subsequently converted to r. The details of the conversion can be found in *RPP\_DataCast.R* and *scicuRe\_source.R[[2]](#footnote-2)*

Is the association between original and replicated effect sizes and/or p-values indicative of replication success? A simple way to test this is to see whether a simple classifier is able to separate replicating from non-replicating studies.

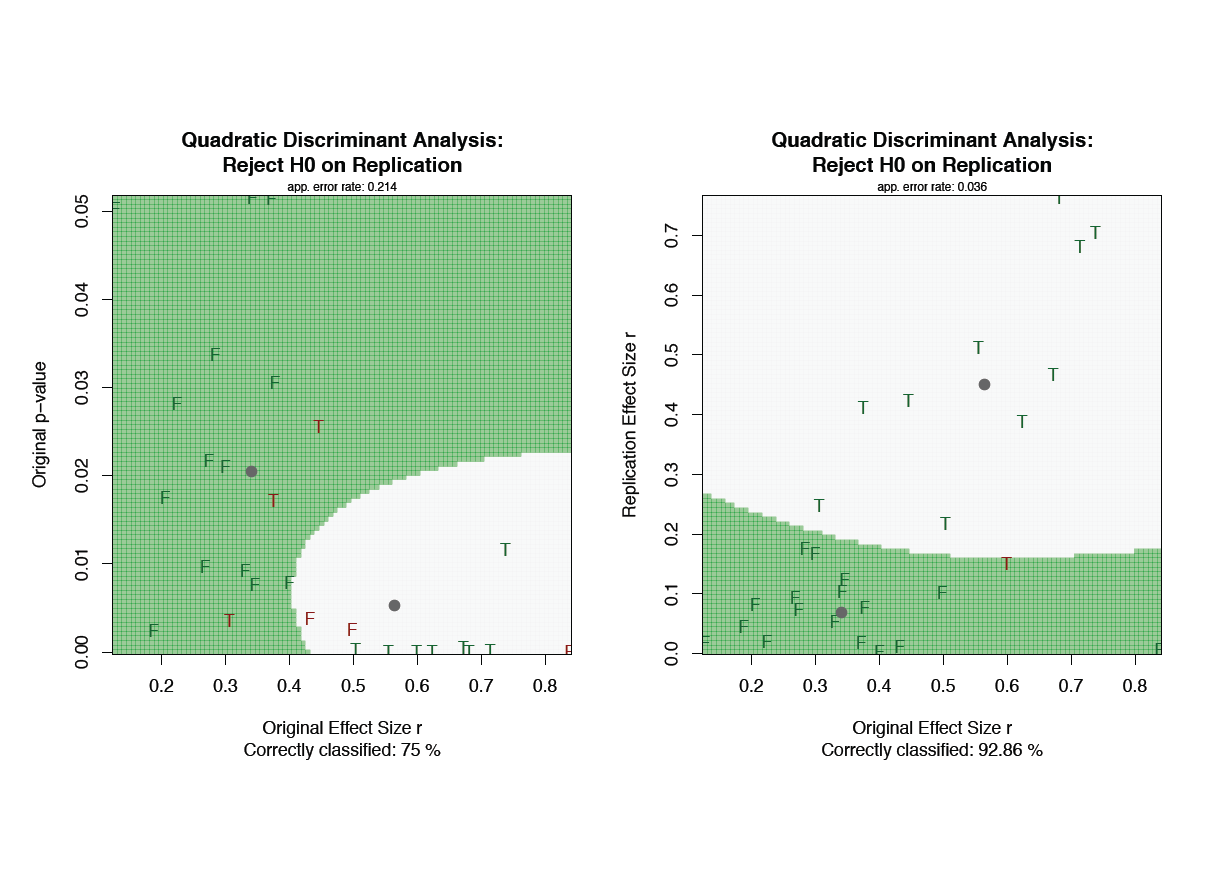


**Figure 1.3.**Original versus replicated magnitude (effect size *r*) with 95% CI. The colour code of the inner part of each marker reflects the original magnitude of r, the outer part the replication magnitude. Likewise, the horizontal CIs belong to the original effect, the vertical to the replication.

Figure 1.4 shows the results from a Quadratic Discriminant Analysis. The graphical representation of the QDA results on left, used the original p-value and effect size to classify replication success, the right used the original and replicated effect size.

The winner is quite clearly the second analysis with 92% correct classification; just two studies are misclassified. This implies all we need to know in order to evaluate the credibility of an effect, is an original, and a replicated effect size magnitude.

**Recommendation 1:** If the relation in Figure 1.4 holds across the larger sample of replications, an effect r <. 4 needs a high powered literal/operational replication in order to be credible (for publication)



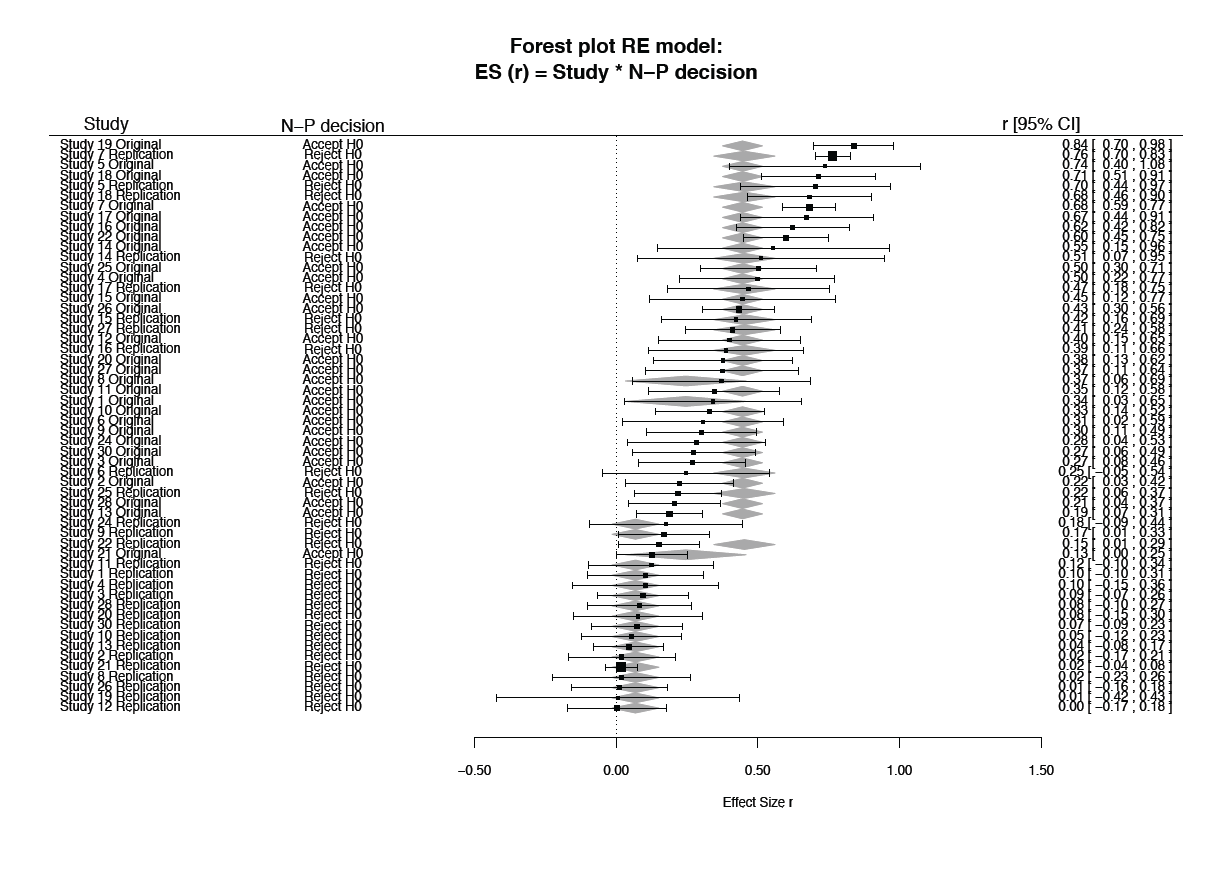
**Figure 1.4.**Two Quadratic Discriminant Analyses in which the studies are classified according to replication success as indicated by statistical significance. See text for details.

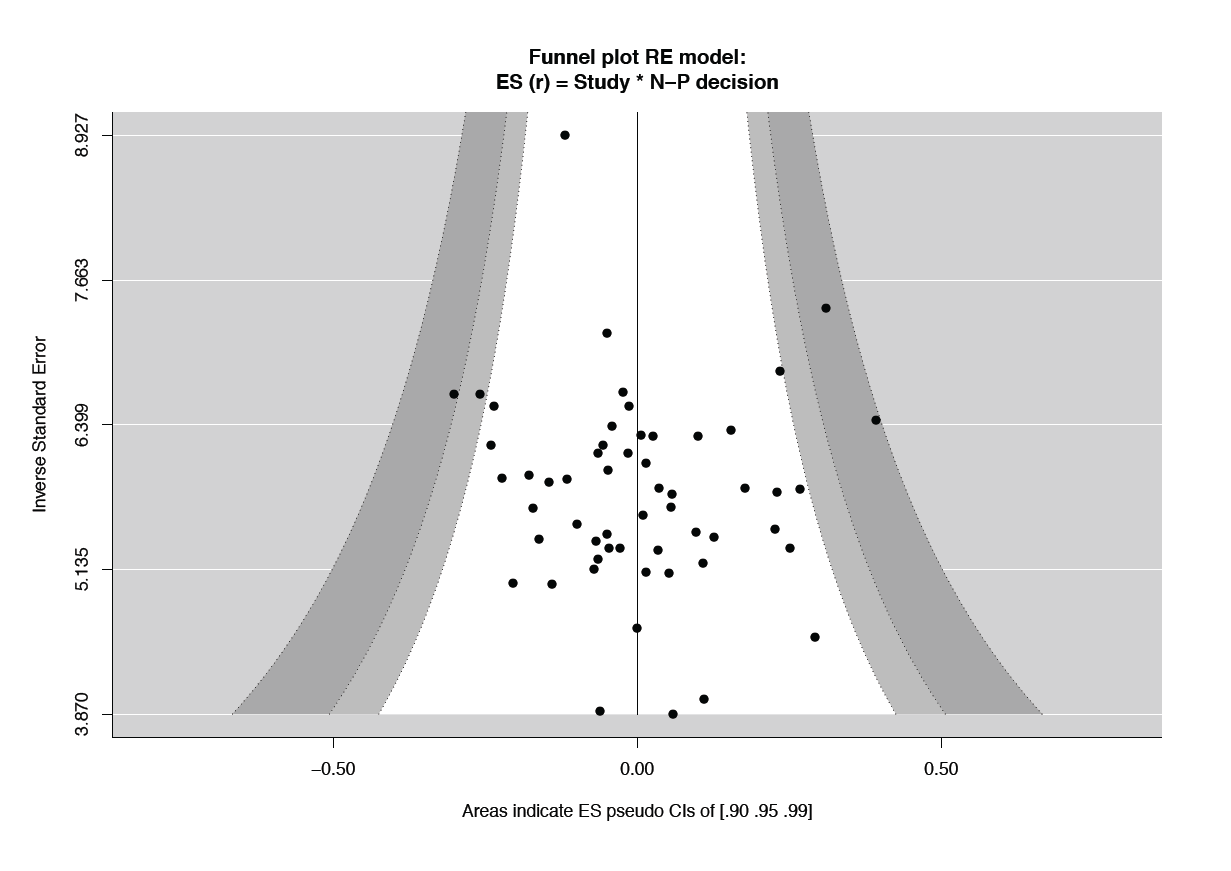
The question is… is this a surprising result? Probably not… This result can be expected in a scientific discipline with a huge bias for publishing positive results only. Most of the published effects in such a discipline will be exaggerated. Only the “True” effects greater than about *r* = .2 (exaggerated = .4) will prevail the scrutiny of more rigorous testing that as is prescribed by the scientific method.

Note that sufficiently powering a replication study in order to ensure detection of the original effect size is just one of many options available to a scientist to put a scientific claim to the test. The preferred confirmatory test is strong inference (e.g., Platt 1964, Fiedler et al., 2012) in which two competing claims are tested directly in a measurement context in which they predict diverging measurement outcomes. Moreover, the assumption that power/precision of psychological measurement increases when sample size is increases, is based on assumptions that are almost certainly false (True score theory, the Ergodic condition, see e.g., Molenaar, 2009).

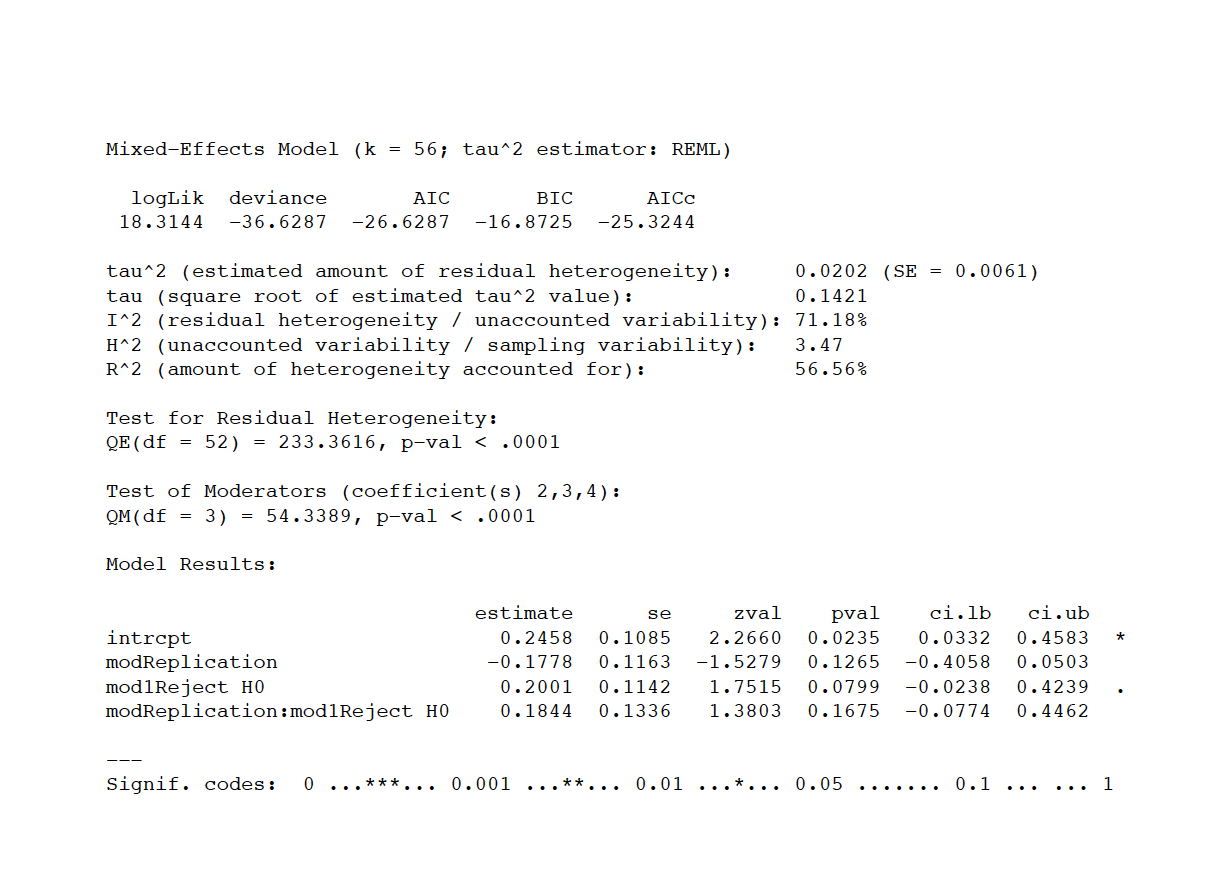
*1.3 Exploratory Meta-Analyses*

Figures in 1.5 (A Forest plot and a Funnel plot) represent a meta-analysis of the effect size, with moderators Neyman-Pearson decision (Accept/Reject H0) and Study (Original/Replication). Output of the analysis is displayed in Figure 1.6





**Figure 1.5.**Forest and Funnel plot of model ES r = Study(Ori/Rep) + N-Pdecision (Accept/Reject) + Study\*N-Pdecision



**Figure 1.6.**Output from the meta analysis.

MORE…*1.4. Evaluation in the context of the Severity principle* (Mayo & Spanos, 2011)

**Severity Principle (full)**. Data **x**0 (produced by process G) provides

good evidence for hypothesis H (just) to the extent that test T severely

passes H with **x**0.

Specifically (Mayo & Spanos, 2011, p. 164):

**Passing a Severe Test.**

We can encapsulate this as follows:

A hypothesis H passes a severe test T with data x0 if,

(S-1) x0 accords with H, (for a suitable notion of accordance) and

(S-2) with very high probability, test T would have produced a result that accords less well with H than x0 does, if H were false or incorrect.

Equivalently, (S-2) can be stated:

(S-2)\*: with very low probability, test T would have produced a result that accords as well as or better with H than x0 does, if H were false or incorrect.

Severity, in our conception, somewhat in contrast to how it is often used, is not a characteristic of a test in and of itself, but rather of the test T, a specific test result x0, and a specific inference H (not necessarily predesignated) being entertained. That is, the severity function has three arguments. We use the notation: SEV (T, x0, H), or even SEV (H), to abbreviate:

“The severity with which claim H passes test T with outcome x0”.

In terms of a replication attempt we can evaluate:

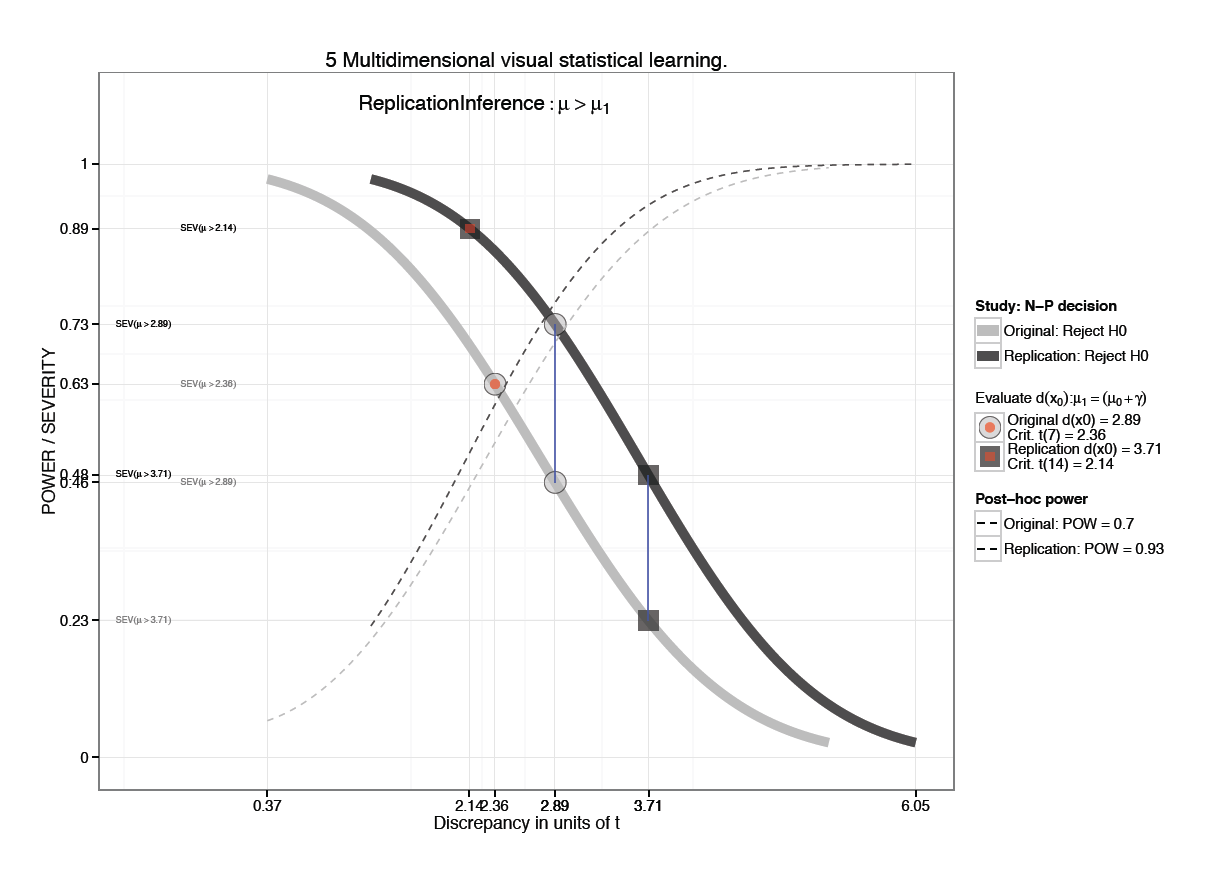
1. The Severity with which the original claim Hori (in all cases the inference of ‘an effect’) passed test Tori with data xori.
2. The severity with which the replication claim Hrep (in many cases the inference of ‘no effect’) passed test Trep with outcome xrep
3. From the perspective of test Trep and data xrep, what would have been the Severity with which claim Hrep passed the test with xori?

Example:

**A. Original = significant AND Replication = Significant**

*Original test gives: SEV(μ > Xt(7)=2.89) = 0.46*

The measurement outcome d(x0) = (μ0 + γ) associated with Μt(7)=2.89,is not a good warrant for discrepancies γ that are more extreme than d(x0). More than 46% of the time a significant test result greater than Μt(7)=2.89 will occur, even if μ ≤ Xt(7)=2.89.



**Figure 1.7.** Replication success for study 5. Severity assessment.

*Replication test gives: SEV(μ > Xt(14)=2.14) = 0.48*

This leads to the same assessment: The measurement outcome d(x0) associated with Xt(14)=2.14,is not a good warrant for discrepancies γ that are larger than d(x0). More than 48% of the time, a significant test result with discrepancy greater than Xt(14)=2.14 will occur, even if in reality μ ≤ Xt(14)=2.14.

*Compare d(x0) of the Original Study to Replication:*

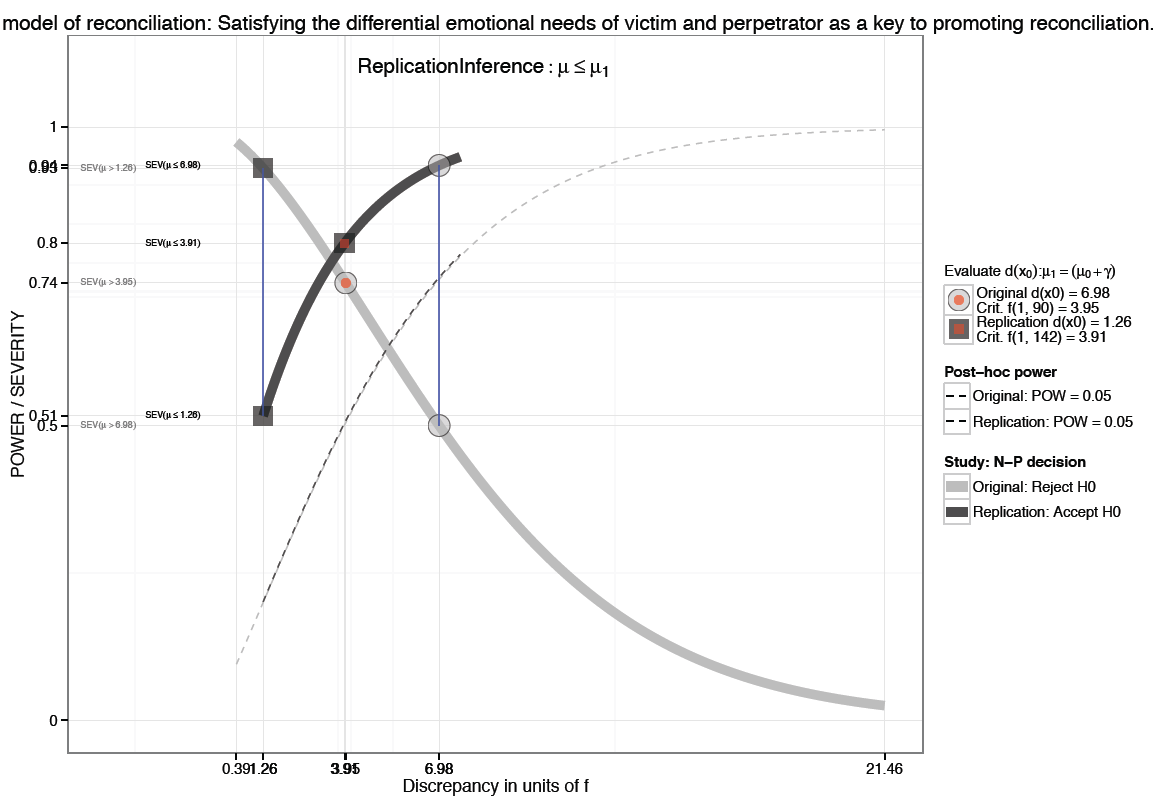
Calculating t(7) under t(14), with ncp = 3.7 (replication) gives SEV = .78.

In R:

pt(2.8920,df=14,ncp=3.7080,F)

[1] 0.7775638

This means that, given the replication data, the observation of the original discrepancy from μ0 would have passed a more severe test of the claim μ > Xt(14)=2.89, but still not ideal (SEV ~ .95)

**

**Figure 1.8.** Replication failure for study 3. Severity assessment.

**B. Original = significant AND Replication = Not Significant**

*Original test gives: SEV(μ > XF(1, 90)=6.98) = 0.50*

The measurement outcome d(x0) = (μ0 + γ) associated with XF(1, 90)=6.98, is not a good warrant for discrepancies γ that are more extreme than d(x0). More than 51% of the time a significant test result greater than XF(1, 90)=6.98 will occur, even if μ ≤ Xf(1, 90)=6.98

*Replication test gives: SEV(μ ≤ XF(1, 142)=1.26) = 0.51*

The measurement outcome d(x0) = (μ0 + γ) associated with XF(1, 142)=1.26,is not a good warrant for inferring μ ≤ XF(1, 142)=1.26. More than 51% of the time an insignificant test result smaller than XF(1, 142)=1.26 will occur, even if μ > XF(1, 142)=1.26

*Compare d(x0) of the Original Study to Replication:*

Calculating F(1, 90) = 6.98 under F(1, 142), with ncp = 1.26 (replication) gives SEV = .93

In R:

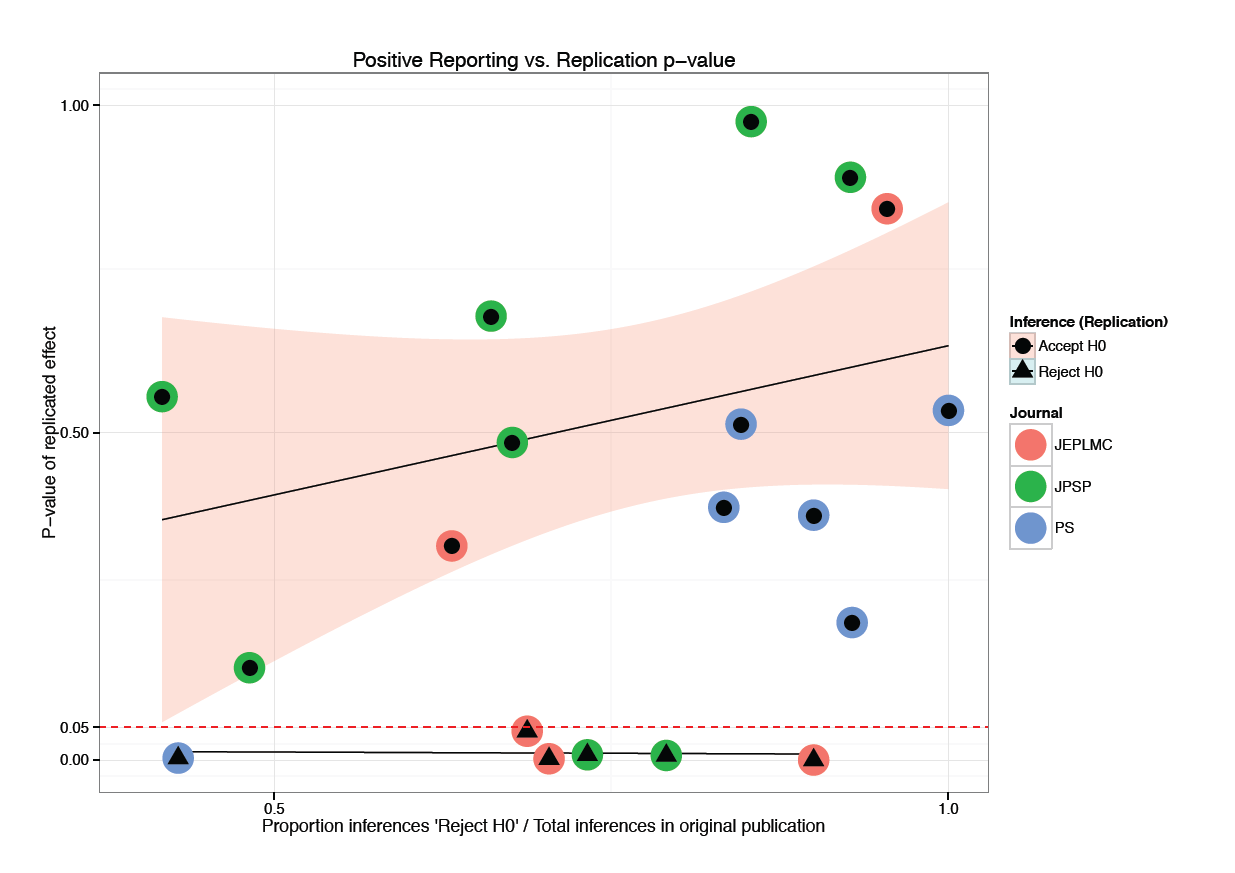
> pf(6.980,df1=1,df2=147,ncp=1.26,T)

[1] 0.9327525

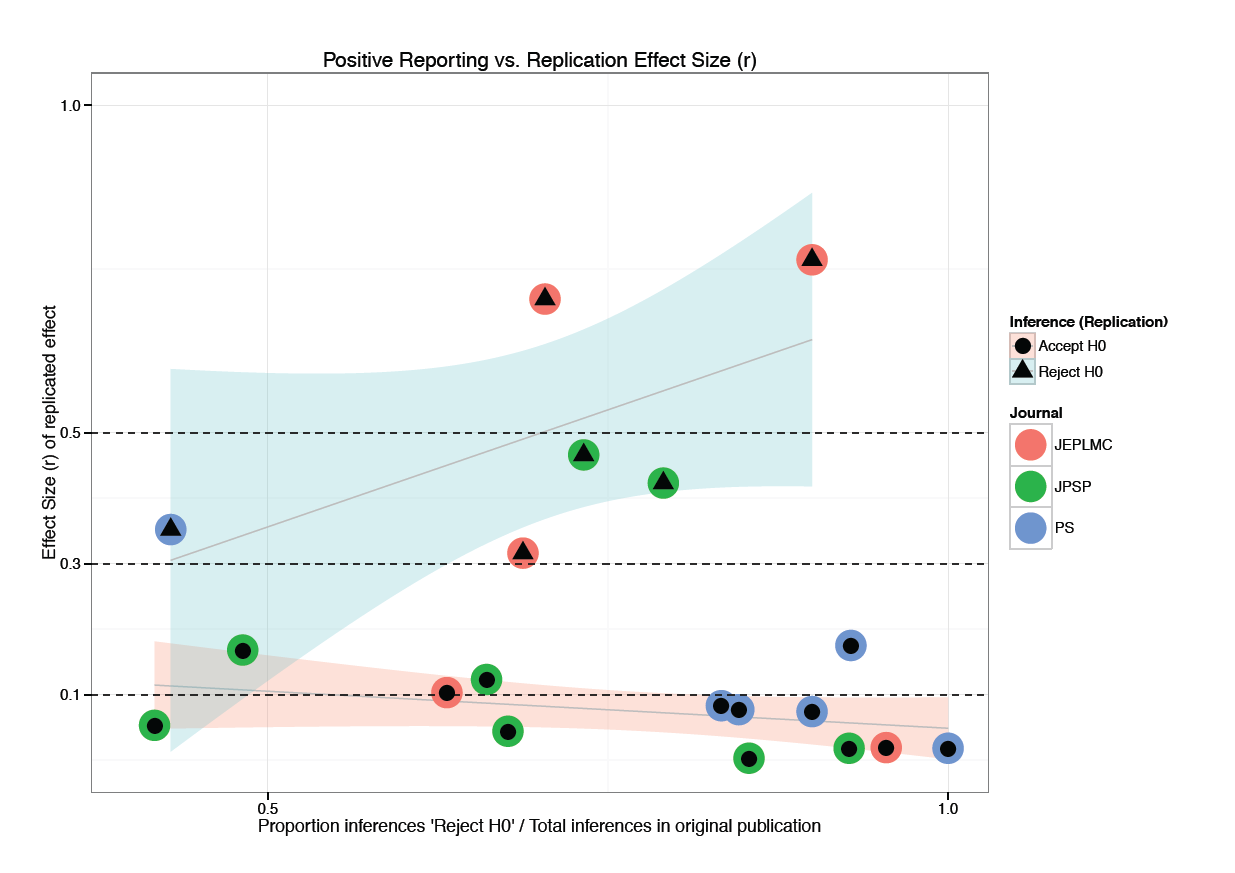
This means that, given the replication data, the observation of the original discrepancy from μ0 would have passed as a severe test for the claim that μ ≤ XF(1, 142)=. 6.98

*1.X Original versus replicated “evidence” in relation to reporting bias (also see section 2.1)*

Figures 1.1 and 1.2 plot the p-value and magnitude associated with the attempt to replicate the effect of interest against the proportion positive / reported statistics in the published in the original article.[[3]](#footnote-3) The number significant test results (*p* < .05, or *prep* > .917) were divided on the total number of test statistics extracted for each article using R scripts (based on package *scicuRe[[4]](#footnote-4)*). Exact accuracy is unknown, but an estimated 90+% of all statistics reported in the set of 19 publications were extracted (see *extracted.xls*). A cause of error is the fact that non-significant results are sometimes reported as “all *Fs* > 1”. These cases are counted as 1 non-significant result. Non-significant results described only indirectly (e.g., “*only the 2 X 2 interaction-effect was significant, F…*”) could of course not be extracted. Assumption is that null-effects cannot be predicted and tests seek rejection rather than confirmation oh H0.



**Figure 1.X.** Replicated p-value associated with the statistical test of the effect of interest versus the proportion positive results reported in the original study. The trend lines were fitted on the “Inference (Replication)” subgroups (6 reject, 14 accept H0) of the 20 studies in the figure (data in Table 1.2).



**Figure 1.X.** Replicated ES (r) associated with the effect of interest versus the proportion positive results reported in the original study. The trend lines were fitted on the “Inference (Replication)” subgroups (6 reject, 14 accept H0) of the 20 studies in the figure (data in Tale1.3).

It appears to be the case that a positive linear relation exists between:

1. The replication p-value of effects of interest that do *NOT* replicate, with the proportion positive results reported in the original study. That is, in hindsight those additional positive results provide false credibility for the effect of interest.
2. The replication ES of effects of interest that *DO* replicate, with the proportion positive results reported in the original study. That is, in hindsight those additional positive results provide raise credibility for the effect of interest.

More data is needed to confirm these preliminary descriptive trends. A relation between smaller replication p-values and reported positive findings (reversal) for replicated effects would be expected if these findings truly provide additional evidence for the observation of the effect.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Table 1.2  *Replication study information for the 20 articles and replicated effects used in this document.* | | | | | | | |
| **Master Row** | **Article Name**  **(truncated)** | **p-value original** | **ES(r) original** | **p-value replication** | **ES(r) replication** | **Journal** | **Inference (Replication)** |
| 1 | Multiple roles for time in short-term | 0.00050 | 0.51700 | 0.04400 | 0.31600 | JEPLMC | Reject H0 |
| 4 | Attractor dynamics and semantic | 0.00040 | 0.67200 | 0.00800 | 0.46600 | JEPLMC | Reject H0 |
| 5 | An attention-based associative account | 0.00020 | 0.71400 | 0.00050 | 0.68200 | JEPLMC | Reject H0 |
| 6 | Multidimensional visual | 0.02300 | 0.73800 | 0.00200 | 0.70400 | JEPLMC | Reject H0 |
| 7 | Intentional forgetting is easier | 0.02800 | 0.22400 | 0.84200 | 0.01900 | JEPLMC | Accept H0 |
| 8 | A needs-based model of reconciliation: | 0.01000 | 0.26800 | 0.26400 | 0.09400 | JPSP | Accept H0 |
| 9 | Why do implicit and explicit | NA | NA | 0.00300 | 0.35200 | JPSP | Reject H0 |
| 10 | 1/f noise and effort on implicit | 0.02200 | 0.27400 | 0.37400 | 0.07400 | JPSP | Accept H0 |
| 11 | Individual differences in the | 0.03100 | 0.37700 | 0.51300 | 0.07700 | JPSP | Accept H0 |
| 12 | Opening the mind to close it: | 0.00240 | 0.18800 | 0.48500 | 0.04400 | JPSP | Accept H0 |
| 15 | Self-regulation and selective exposure: | 0.02100 | 0.30000 | 0.14100 | 0.16800 | JPSP | Accept H0 |
| 16 | Not so innocent: Does seeing one's | 0.00380 | 0.43200 | 0.89300 | 0.01200 | JPSP | Accept H0 |
| 17 | Affective incoherence: | 0.01750 | 0.20600 | 0.38600 | 0.08300 | JPSP | Accept H0 |
| 19 | Conflict-triggered goal shielding: | 0.01700 | 0.37500 | 0.00004 | 0.41100 | PS | Reject H0 |
| 21 | Temporal selection is suppressed, | 0.00001 | 0.55500 | 0.00000 | 0.51100 | PS | Reject H0 |
| 22 | With a clean conscience: | 0.00790 | 0.40000 | 0.97500 | 0.00280 | PS | Accept H0 |
| 23 | Creating social connection through | 0.00770 | 0.34600 | 0.67800 | 0.12300 | PS | Accept H0 |
| 25 | Effects of fluency on psychological | 0.05030 | 0.12700 | 0.53400 | 0.01800 | PS | Accept H0 |
| 28 | The threat of appearing | 0.05140 | 0.37200 | 0.89000 | 0.01800 | PS | Accept H0 |
| 29 | Keeping one's distance: The | NA | NA | 0.55500 | 0.05300 | PS | Accept H0 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Table 1.3  *Summary information for the 20 articles used in this document: Number of extracted statistics and proportion ‘Reject H0’* | | | | | |
| **Master Row** | **Article Name (full)** | **N Stats** | **Reject H0** | **Accept H0** | **PSI** |
| 1 | Multiple roles for time in short-term memory: Evidence from serial recall of order and timing. | 19 | 12 | 7 | 0.63 |
| 4 | Attractor dynamics and semantic neighborhood density: Processing is slowed by near neighbors and speeded by distant neighbors. | 22 | 21 | 1 | 0.95 |
| 5 | An attention-based associative account of adjacent and nonadjacent dependency learning. | 27 | 19 | 8 | 0.70 |
| 6 | Multidimensional visual statistical learning. | 16 | 11 | 5 | 0.69 |
| 7 | Intentional forgetting is easier after two "shots" than one. | 10 | 9 | 1 | 0.90 |
| 8 | A needs-based model of reconciliation: Satisfying the differential emotional needs of victim and perpetrator as a key to promoting reconciliation. | 110 | 102 | 7 | 0.93 |
| 9 | Why do implicit and explicit attitude tests diverge? the role of structural fit. | 27 | 13 | 13 | 0.48 |
| 10 | 1/f noise and effort on implicit measures of bias. | 24 | 10 | 12 | 0.42 |
| 11 | Individual differences in the regulation of intergroup bias: The role of conflict monitoring and neural signals for control. | 56 | 37 | 15 | 0.66 |
| 12 | Opening the mind to close it: Considering a message in light of important values increases message processing and later resistance to change. | 82 | 70 | 12 | 0.85 |
| 15 | Self-regulation and selective exposure: The impact of depleted self-regulation resources on confirmatory information processing. | 34 | 23 | 11 | 0.68 |
| 16 | Not so innocent: Does seeing one's own capacity for wrongdoing predict forgiveness? | 11 | 4 | 7 | 0.36 |
| 17 | Affective incoherence: When affective concepts and embodied reactions clash. | 43 | 34 | 2 | 0.79 |
| 19 | Conflict-triggered goal shielding: Response conflicts attenuate background monitoring for prospective memory cues. | 56 | 41 | 14 | 0.73 |
| 21 | Temporal selection is suppressed, delayed, and diffused during the attentional blink. | 13 | 11 | 0 | 0.85 |
| 22 | With a clean conscience: Cleanliness reduces the severity of moral judgments. | 3 | 3 | 0 | 1.00 |
| 23 | Creating social connection through inferential reproduction: Loneliness and perceived agency in gadgets, gods, and greyhounds. | 7 | 3 | 4 | 0.43 |
| 25 | Effects of fluency on psychological distance and mental construal (or why new york is a large city, but new york is a civilized jungle). | 14 | 13 | 1 | 0.93 |
| 28 | The threat of appearing prejudiced and race-based attentional biases. | 6 | 5 | 1 | 0.83 |
| 29 | Keeping one's distance: The influence of spatial distance cues on affect and evaluation. | 10 | 9 | 1 | 0.90 |

**2. *Why* and *what* of judging replication attempts**

*2.1* ***Why*** *this outcome?*

Guidelines were provided for replication teams to target an effect. The effects were not selected at random from a target article. This is defendable with respect to feasibility (e.g., costs, sample size, etc.). The following guideline must be taken into account when evaluating replication efforts:

**Note that,** by default, the last study of each article is the target for replication. If it is not feasible or sensible to select the last study for replication, document the rationale for not selecting that study and then evaluate whether it is feasible to replicate the second-to-last study. If no studies are feasible to replicate by any OSC members, then the article will not be included for replication. (<https://osf.io/ezcuj/wiki/current-contributors:replicate:p2/>)

Specific selection of the last study in a multi-experiment article that can be conceived of as close replications of a single phenomenon as the target for replication may require some attention. It could result in a focus by RPP on exaggerated effect sizes. Each subsequent claim of having evidenced a phenomenon increases credibility by suggesting a Type-I error was unlikely the cause of the previous positive observation. To accommodate this drop in the effective alpha level requires an increase in power (e.g., by increasing N) of each study in order to keep the probability of committing a Type-II error the same across the experiments (see e.g., Button et al., 2013, Schimmack, 2012). Schimmack (2012) provides a table with the total power (provided by N particiants) of a multiple-experiment study necessary to maintain 80% power per study for large, moderate and small effect sizes, here's an excerpt for independent sample tests:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| N experiments | total Power needed | Large (d=.8) | Moderate (d=.5) | Small (d=.2) |
| 1 | 80.0 | 52 | 128 | 788 |
| 2 | 89.4 | 136 | 336 | 2068 |
| 5 | 95.6 | 440 | 1090 | 6750 |
| 10 | 97.8 | 1020 | 2560 | 15820 |

Many replication teams go beyond 80% power:

"(...) [T]he interaction was significant, F(1,94) = 4.97, p < .05, MSE = .029, η2 =.05, (...)". (p. 412). Power analysis (G\*Power (Version 3.1): ANOVA: Repeated measures, within-between interaction with a zero correlation between the repeated measures) indicated that sample sizes for 80%, 90% and 95% power were respectively 78, 102 and 126.

The quoted study chose 90% power as the target for the replication sample (N=102, actual N=108). The original was a 3-experiment study (effective α = .000125), the effect is between small and moderate (f = 0.23). Total power at 80%, 90%, and 95% would require 216, 258, and 294 subjects respectively. The post-hoc power for the original effect size at effective α was (1-β) = 0.22. The power of the replication study to detect the original effect size at effective α with N=108 is only (1-β) = .28 (G\*Power 3.1).

* **Recommendation:** Calculate the total power of original and replication study. Differences in replication success between different N-experiment studies may be explained this way.

*2.2* ***What*** *is this outcome?*

Some replication studies concern the sign of the difference in means between experimental conditions; some are interaction effects between multileveled factors. Should successful replication of a simple effect be treated the same way as a successful replication of a complex interaction? Why focus on the conditional effect of the factor and not the model / omnibus outcome. It is often the case that the prediction being tested by factorial models involves more than a significant interaction effect.

These intuitively different situations with respect to evaluation and interpretation of replication outcomes concern the effect of epistemic sloughing, or breaking of the epistemic linkage between theoretical prediction and statistical hypothesis that occurs when a statistical test is conducted.

**Theory Specifications (Meehl, 1990)**

I . Type of entity postulated (substance, structure, event, state, disposition, field)

II. Compositional, developmental, or efficient-causal connections between the entities in I

III . Signs of first derivatives of functional dynamic laws in II

IV . Signs of second derivatives of functional dynamic laws in II

V . Ordering relationships among the derivatives in I

VI . Signs of mixed second order partial derivatives (Fisher “interactions") in II

VII . Function forms (e. g., linear? logarithmic? exponential?) in II

VIII . Trans-situationality of parameters in b

IX . Quantitative relations among parameters in VII

X. Numerical values of parameters in II Figure

It is possible to ‘rank’ the kind of prediction that a statistical test is supposed to evidence. Borrowing from Meehl’s list of theory specification, differences between replicated studies can be shown to exist. For example, McKinstry, Dale and Spivey (2008) report a number of tests conducted on mouse movement trajectories. The tests reflect predictions that increase down the table of theory specification, e.g., from III –signs of first derivatives– where the trajectory end points are predicted to be correlated to stimulus type (displacement stimulus X > stimulus Y) down to second order derivatives (velocity, acceleration peak) and function forms (sample entropy).

* **Suggestion:** Rank the original prediction along the levels of theory specification and include in the final evaluation.

**APPENDIX 1**

**“Frequentist” analysis committee:** Marcel van Assen (MA), Denny Borsboom (DB), Fred Hasselman (FH), Deborah Mayo (DM), Frank Renkewitz (FR), Jelte Wicherts (JW)

**Goal**

The purpose of this document is to collect ideas and opinions about what it is we can / should be asking of the replication data produced by the reproducibility project, from a frequentist perspective. Topics may vary from very practical “how to”, to validity of interpretation. Remember that results from our discussion can also inform other committees e.g. concerning effect size measures or meta-meta-meta analysis strategies.

Another goal that I would like to set is: “get it as right as possible”, that is, be explicit about assumptions, use most recent recommended practices and terminology, if the right tools or terminology do not exist, we create them. Issues that need to be addressed from this perspective can be listed in the “Concerns” section

I am not an expert in these matters, perhaps none of us really are, because this will be a unique dataset. Do not hesitate to point out any errors you encounter, or comment on opinions you disagree with, this way, we will all learn.

Some guidelines:

* add *new topics* within a section using a boldface title starting with the section name (see below)
* add *new contributions* within a topic by adding a paragraph starting with your initials (see below)
* add *comments* to existing text by insert -> comment in the Googledocs menu.
* do not delete any text that is not yours.

All the best,

Fred

**Contents**

*Concerns*(about anything related to replication data analysis, “the frequentist way”)

* NHST
* “regular” meta-analysis
* Random effect models
* Effect size measures

*Proposals*(modelling strategies, tests of specific hypotheses, addressed concerns)

* “Truth indicator” - Categorical mixture model
* “Power check” - Identify anomalous results with respect to statistical power

*Meta* (check your baggage, the philosophical kind that is)

* Interpreting replication studies

**Concerns about NHST**

*Suppose we want to strictly follow the rules of inference dictated by NHST, that is, if we can decide what those rules exactly are, what problems need to be addressed?*

*DM:* I'd recommend a move away from a term that is non-standard and, given how it's been described since Meehl, may correspond to a variant on statistical significance tests that exists only as a flawed version of tests. I refer of course to NHST, and the central two flaws are its (1) identifying a rejection of the null with evidence for a substantive alternative (which might be thought to 'explain' or predict the significant result), and (2) leaving in limbo the status of non-significant results (such that people allege that the null is inferred or that nothing at all may be said from these tests in the case of non-rejection). There is no statistical significance test, not even "pure" significance tests, that licenses these moves, so the term seems to me to signify a caricature of these methods. I believe it was introduced by psychologists, perhaps as a shorthand, but does not exist in Fisher or in Neyman and Pearson. "Statistical significance tests" would be a better term.

*FR*: I am not sure if I understand where this problem is relevant for us. In general, I am very much in line with the idea to move away from NHST. However, given that probably (almost) all of the original papers used NHST to evaluate their hypotheses, it is certainly an interesting and relevant information which proportion of the replication studies again obtained a significant result. At least in a first step of this analysis, I would not even hesitate to re-run tests that seem completely flawed to me. I think that it is an interesting information in itself how often an identical analytic procedure would lead to the same conclusion (even if the logic behind this conclusion is dubious). This has the additional advantage of being an easy task we can definitely accomplish… – The PB\_QRP group seems to go for a search of statistical errors in the original papers (e.g., inconsistent dfs, p-values that are rounded off to appear significant). Of course, it may be a good idea to take into account these results in addition. – As far as tests of hypotheses that we will develop here are concerned I am inclined to evaluate these hypotheses without NHST.

**Concerns about “regular” meta-analysis**

*Suppose we want to use classical meta-analytic techniques (just to see how far we can get), what problems need to be addressed?*

*MA*: Problems are that the results from the original research are contaminated by qrps and or publication bias. The publication bias and qrp committee addresses these issues and their effects on the estimates of the meta-analysis.

**Concerns about applicability of “random effects models” for meta-analysis**

*In a previous e-mail exchange, Denny expressed concerns about the applicability of a multilevel approach / meta-SEM strategy to analyse the replication data. Add your comments.*

*DB:* So for instance one study replicates a main effect in a factorial design, another replicates an interaction in a mixed design, a third replicates a correlation in a quasi experimental design. Unless I am missing something, that precludes the application of random effects sem. I am also unsure of the applicability of general meta analytic techniques in this case. Are there precedents (i.e. where people modelled qualitatively different effects using meta analysis)?

*FH:* The applicability of using a random effects model to explain differences in some effect size measure calculated for each study is questionable because of the fact that the replicated studies are not the outcomes of a random selection of the entire sample space e.g., “studies published in 2008”. If random selection of studies had been the case, I would have had less problems assuming a standardised effect size magnitude to represent the value a random variable had taken on based on the random outcome “empirical study X conducted in 2008”.

*MA*: I assume that the most important effects are or can be determined, for each paper. Moreover, these effect sizes can be converted to one and the same metric, e.g., d or (fischer) r. Then, in my opinion, the random-effects meta-analysis model can be applied in two ways. First, the model can be estimated separately on all original studies and on all replication studies separately. The estimates of mu but also of tau2, and their dependence on e.g. sample size but also other study characteristics can be determined. Second, the model can be applied to the differences in effect sizes of replication and original studies. That is, each ‘pair’ yields one effect size (difference), which is modelled.

*FR*: From a classical meta-analytical perspective, I do not think that mu or tau^2 represent anything meaningful in our case. Obviously, there is no single true effect size and there is no variability in true effect sizes that could be considered as (purely) random. So, there is little sense in estimating a true effect size or the variability of true effect sizes. Additionally, the studies do not a represent a random sample from a known population, so we do not know to which population our results generalize. Still, I would use meta-analytic techniques. The reason simply is that I am interested in the average effect size of the first 30 (?) studies that were published in JPSP in 2008. I would like to compare this average effect size to the average effect size of direct replications of these 30 studies. The same is true for effect sizes of original studies and replications in JEP: LMC, for the correlation between effect sizes of original studies and replications, for the difference in average effect sizes in different journals and so on. I think all of this can be done in a purely descriptive way and is still more than interesting. Again, my feeling is that the 50 studies we will analyze in this first step constitute a population that is interesting in itself. Fixed and random effects models (or even an unweighted mean) then simply provide different weighting schemes for the computation of an average effect size. Assuming that we are able to convert effect sizes from all studies to the same metric, all of this is easily computed and just provides a robustness check of our results. In the end, I would probably also feel tempted to compute, say, a CI for these average effect sizes, but I assume that it will remain hard to argue that this represents the CI for the mean effect size in psychology, social psychology, JPSP, or even in JPSP in 2008.

**Concerns about Effect Size measures**

*As mentioned above, there will be great diversity in statistical hypotheses tested. I have doubts about the validity of using a common effect size measure for all replicated studies. Add your comments.*

*FH:* Some concerns about effect size measures:

* Restriction of range problems: Clinical studies are often conducted on participants that were sampled from very specific regions under the population distribution (e.g., the lowest 25%).
* Predicting an interaction between factors often concerns a prediction about a specific pattern of main effects in addition to a significant interaction. The omnibus effect size of the entire model would be preferable above the interaction effect size in such cases. This is probably not reported in many articles.
* Predicting an interaction is riskier than predicting a nonzero mean difference: One concerns a partial derivative, the other the sign of a correlation. Should this difference in prediction risky-ness be accounted for when evaluating replication results?
* There is not a lot of attention for measurement scales and effect sizes, I am not convinced it is appropriate to calculate a standardised mean difference from ordered categorical variables (rating scales), but people do.
* Response latencies are not normally distributed, moreover, the measurement scale allows for unstandardised ES. Should we look at unstandardised effect sizes if it is allowed by the measurement scale?
* Should we use Effect Size CIs based on non-central distributions or estimate CIs based on (pooled) sample SDs?

*FR*: I am less optimistic with respect to the problem of converting different effect sizes to the same metric. In the PB\_QRP group, Marcel already pointed me to a section in Borenstein et al. (2009) that gives formulas to convert d, g and r to one of these metrics. I am afraid that this will not solve the problem. The actual problem arises from different designs. Even in a simple 2 x 2 design there are at least two ways to compute a d (or g or r) for the main effects. The problem is to compute an effect size that is comparable to an effect size from a two group design (and to more complex designs, of course). I guess that the effect size that is available from the original papers in case of 2 x 2 designs will mostly be a partial eta^2. This can be converted to d, but this procedure is likely to result in effect sizes that are larger than the effect sizes from two group designs (as some portion of the error variance is partialled out). Even if we have the necessary information to take into account the complete error variance I do not think that it is clear what the appropriate way to go is (and I have some doubts whether there is a general answer to this question that is independent of the specific experiment, hypotheses and purpose of the analysis). As far as I can see, there is no agreement about the appropriate effect size for the interaction effect in a 2 x 2 design (at least, with respect to replications as the usual eta^2 may mean very different things in two (identical) studies). And the problem definitely gets more complicated with more complex designs. – From a practical perspective, I think that this is our first and most urgent problem. When this problem is solved there is a large amount of analyses that can be easily done and that seem interesting and valuable to me. As long as it is not solved, I have some difficulties to see what we can do at all (on an aggregate level of analysis). Maybe it is not necessary to convert effect sizes from all different designs to one metric. An alternative might be to define “classes” of designs that yield comparable effect sizes. However, at the moment I could not do this either. Would it be a useful first step to determine which designs were used how often in our sample of studies? At the least, this would tell us what we would like to compare and which problems we have to solve.

**Proposal: “Truth indicator” - categorical mixture:**

*Suggested as an alternative to random effect models*

*DB:* A possible alternative approach, one level higher in the aggregation process, might be to model the situation as a mixture, where each study and its replication are treated as two (equally fallible) indicators of a categorical latent variable that would code the truth (i.e. true vs false positives (or negatives), sorry andrew ;-). Then we could apply a mixture model and estimate the "prevalence" of true positives/negatives in the system. Maybe that could work if we have enough replications (these would count as cases). I am not sure but it may be worth looking into.

*MA:* Personally, I am not a fan of this approach, but I could be convinced that it is a good approach. My problem with the approach is that we make some strong assumptions, which could strongly affect the results of the analysis;are the results and conclusions based on these results robust to changes in the assumptions?

**Proposal: “Power Check” - Identify anomalous results with respect to statistical power**

*There is more to power than sample size*

*FH:* In the ManyLabs project some effects were replicated at much higher magnitudes than the effect of the original study, even though the original study had enough statistical power to detect both the original and replicated effects. The proposal is to identify such apparent anomalies and suggest explanations for their occurrence.

**Meta: Interpreting Replication Studies - 50 shades of grey areas?**

*This section is to inform other members about your opinion on the role of replication studies in advancing scientific knowledge about psychological phenomena. Purely informative, skip if you want. It may help to put into perspective why members use certain arguments to get a point across, or why they even raise a point in the first place.*

*FH:* An empirical article should be a report of rigorous and severe testing of a theoretical claim, i.e. a sincere effort to show it is false. A credible theory has a track-record of surviving such rigorous and severe tests. The lack of replication studies, p-hacking, publication bias reflect that earning credibility by testing risky predictions is not on the minds of empirical social scientists. From the perspective of rigorous testing, there are three types of replication studies: 1) Least impressive are literal replications (different sample, same population, same lab, same researchers); 2) A successful operational replication (same procedure and materials, tested in another lab by peers) should provide credibility, but only if literal replications are possible. 3) A conceptual replication (a test of the theoretical construct) can give the highest levels of credibility, but only if type 1 and 2 replications have been successful.

1. Files (scripts, data, etc.) and PDFs of Figures are available here https://osf.io/xtfwd [↑](#footnote-ref-1)
2. Functions are called: t\_r, f\_r, X\_r, etc. and can be found here: <https://github.com/FredHasselman/scicuRe>/ [↑](#footnote-ref-2)
3. [↑](#footnote-ref-3)
4. In ‘very beta’ stage of development: <https://github.com/FredHasselman/scicuRe> [↑](#footnote-ref-4)