THE REPLICATION PARADOX
Combining Studies Can Decrease Accuracy of Effect Size Estimates
Michèle B. Nuijten, Marcel A. L. M. van Assen, Coosje L. S. Veldkamp, & Jelte M. Wicherts, Tilburg University, m.b.nuijten@uvt.nl

INTRODUCTION
Imagine you want to estimate a certain effect size. You search the literature and find two identical studies that only differ in size.

In a survey, we gave 106 psychology students, 350 social scientists, and 31 quantitative psychologists 4 study combinations and asked: Which situation yields the most accurate estimate of the effect size in the population? The answers are summarized below:

- The intuition is the more information, the better.
- If a replication is available, researchers would include it in their effect size estimate.
- This seems very logical: Statistics teach us “the higher the N, the higher your precision” and the literature is filled with the advantages of replication.

However, the intuition is false.

PROBLEM

- PUBLICATION BIAS inflates published effect sizes.
- Seems to hold for REPLICATIONS as well?
- How does this affect effect size estimates when studies are COMBINED?

METHOD

We analytically derived several scenarios in which we investigated the influence of publication bias on the effect size estimate in simple two-independent-sample designs. We varied N (20 or 35 subjects per group), the population effect size (Cohen’s d = 0–1), and publication bias (proportion of non-significant results published: pub = 0, 0.5, 25, 50, 1).

In each scenario, we measured the bias in published effect size (bias = published effect size / population effect size). We used the weighted averages of the bias in two single studies to calculate the bias in a replication scenario.

BIAS IN SINGLE STUDIES

Researchers preferred two large studies over one. However, because both studies contain the same amount of bias, the weighted average of the two effect sizes will contain the same amount of bias as a single large study. Furthermore, adding a small replication to a large study increases bias, since a small study contains more bias in the first place. Finally, the bias in two small studies is equal to the bias in one small study, which is higher than the bias in a single large study. It turns out that in all scenarios of the questionnaire, a single large study renders the most accurate effect size estimate.

BIAS IN REPLICATIONS

There are several solutions:

- Eliminate publication bias. There are some good initiatives already, but it will take time. And what to do with the studies that are already published?
- Correct for publication bias. However, existing tests have low power or strong assumptions.
- Increase POWER. Only evaluate (and only perform) studies with high power.

Based on our results, we present the correct answers to this survey, in which we asked: Which situation yields the most accurate estimate of the effect size in the population? Remember that “A” is a study with N=20, and “B” with N=40.

SUMMARY

- A replication will increase PRECISION but...
- A replication adds BIAS if N is smaller than original study.
- But ONLY when there is PUBLICATION BIAS and POWER is not high enough.

CONCLUSION

So when you perform a literature search and want to get the most accurate effect size estimate, discard all underpowered studies.

Scan the QR code with your smartphone to download a copy of this poster including more detailed information about the calculations and an explanation of the “bumper” in the graphs.
EQUATIONS TO CALCULATE BIAS

\[ cv = 1.645 \cdot \sqrt{\frac{2}{N}} \]

\[ z_1 = (cv - d) \cdot \sqrt{\frac{N}{2}} \]

\[ D_1 = \frac{f(z_1)}{(1 - \beta) \cdot \sqrt{N/2}} + d \]

\[ D_0 = d - \frac{f(z_1)}{\beta \cdot \sqrt{N/2}} \]

\[ D = \frac{pub \beta D_0 + (1 - \beta)D_1}{pub \beta + (1 - \beta)} \]

\[ D_{total} = \frac{N_a D_0 + N_b D_1}{N_a + N_b} \]

Schematic representation of the effect of publication bias on the published effect size estimate. "H0" and "H1" are the regions of accepting and rejecting H0, respectively. Hβ represents power, α is the type I error, cv is the critical value of the test, and d is the true effect size. D0 and D1 are the expected effect sizes conditional on the acceptance or rejection of H0, respectively, and D is the expected value of the published effect size.

THE "BUMP" IN THE GRAPH

The bump in the graph can be explained as follows.

The total bias consists of two opposing forces:

1. Bias when \( pub = 0 \) decreases as \( d \) increases
2. Relative bias increases when \( d \) increases

Total bias = 1 * 2

When \( d \) is low:
- Increase in relative bias overrules decrease in bias for \( pub = 0 \)

When \( d \) is high:
- Relative bias increases relatively less than bias for \( pub = 0 \)