Keanu Lim

Notes on Replicability Study

**All Claims/Subclaims**

1. “In stage 1, we find a significant effect in the same direction as the original study for 12 replications (57.1%) (Fig. 1a and 19,22–25,27,29,30,36 Supplementary Table 3). When we increase the statistical power further in stage 2 (Fig. 1b and Supplementary Table 4), two additional studies replicate based on this criterion” (2)
   1. 13 studies after stage 2
2. “The mean standardized effect size (correlation coefficient *r*) of the replications is 0.249, compared to 0.460 in the original studies (Supplementary Fig. 4). This difference is significant (Wilcoxon signed-ranks test, *z* = 3.667, *P* < 0.001, *n* = 21) and the mean relative effect size of the replications is 46.2%. For the 13 studies that replicated, the mean relative effect size is 74.5%, and for the 8 studies that did not replicate, the mean relative effect size is 0.3%.”
3. “As seen in Fig. 1c, 16 studies (76.2%) have a significant effect in the same direction as the original study in the meta-analysis”
4. A complementary replication criterion is to count how many replicated effects lie in a 95% prediction interval47, which takes into account the variability in both the original study and the replication study. Using this method, 14 effects replicated”
5. Following the small telescopes approach, 12 studies(57.1%) replicate
6. The one-sided default Bayes factor exceeds 1, providing evidence in favour of an effect in the direction of the original study for the 13 (61.9%) studies that replicated according to our primary replication indicator (Fig. 3). This evidence is strong to extreme for 9 (42.9%) studies. The default Bayes factor is below 1 for 8 (38.1%) studies, providing evidence in support of the null hypothesis; this evidence is strong to extreme for 4 (19.0%) studies.
7. “The estimated true-positive rate is 67% (Supplementary Fig. 5), which is close to the other replicability estimates. The mixture model also estimates that the average relative effect size of true positives is 71%”’
8. To summarize, we successfully replicated 13 out of 21 findings from experimental social and behavioural science studies published in *Science* or *Nature* between 2010 and 2015 based on the statistical significance criterion with very high-powered studies compared to the RPP12 and the EERP
9. First, even among successful replications, the estimated effect sizes were smaller than the original study. For the 13 studies that replicated according to the statistical significance criterion, the replication effect sizes were about 75% of the original effect size. This provides an estimate of the overestimation of effect sizes of true positives in the original studies. The Bayesian mixture model corroborates this result, yielding an estimate of the relative effect size of true positives of 71%.
10. Second, among the unsuccessful replications, there was essentially no evidence for the original finding. The average relative effect size was very close to zero for the eight findings that failed to replicate according to the statistical significance criterion.

**Main Claims**

1. To summarize, we successfully replicated 13 out of 21 findings from experimental social and behavioural science studies published in *Science* or *Nature* between 2010 and 2015 based on the statistical significance criterion with very high-powered studies compared to the RPP12 and the EERP
   1. “In stage 1, we find a significant effect in the same direction as the original study for 12 replications (57.1%) (Fig. 1a and 19,22–25,27,29,30,36 Supplementary Table 3). When we increase the statistical power further in stage 2 (Fig. 1b and Supplementary Table 4), two additional studies replicate based on this criterion” (2)
      1. 13 studies after stage 2
   2. “As seen in Fig. 1c, 16 studies (76.2%) have a significant effect in the same direction as the original study in the meta-analysis”
   3. A complementary replication criterion is to count how many replicated effects lie in a 95% prediction interval47, which takes into account the variability in both the original study and the replication study. Using this method, 14 effects replicated”
   4. Following the small telescopes approach, 12 studies(57.1%) replicate
   5. The one-sided default Bayes factor exceeds 1, providing evidence in favour of an effect in the direction of the original study for the 13 (61.9%) studies that replicated according to our primary replication indicator (Fig. 3). This evidence is strong to extreme for 9 (42.9%) studies. The default Bayes factor is below 1 for 8 (38.1%) studies, providing evidence in support of the null hypothesis; this evidence is strong to extreme for 4 (19.0%) studies.
2. “The mean standardized effect size (correlation coefficient *r*) of the replications is 0.249, compared to 0.460 in the original studies (Supplementary Fig. 4). This difference is significant (Wilcoxon signed-ranks test, *z* = 3.667, *P* < 0.001, *n* = 21) and the mean relative effect size of the replications is 46.2%. For the 13 studies that replicated, the mean relative effect size is 74.5%, and for the 8 studies that did not replicate, the mean relative effect size is 0.3%.”
   1. First, even among successful replications, the estimated effect sizes were smaller than the original study. For the 13 studies that replicated according to the statistical significance criterion, the replication effect sizes were about 75% of the original effect size. This provides an estimate of the overestimation of effect sizes of true positives in the original studies. The Bayesian mixture model corroborates this result, yielding an estimate of the relative effect size of true positives of 71%.
   2. Second, among the unsuccessful replications, there was essentially no evidence for the original finding. The average relative effect size was very close to zero for the eight findings that failed to replicate according to the statistical significance criterion.
3. “The estimated true-positive rate is 67% (Supplementary Fig. 5), which is close to the other replicability estimates. The mixture model also estimates that the average relative effect size of true positives is 71%”’