## Supplementary materials

### Supplementary material 1

#### Replication project sampling frames, extraction, participants, and exclusion details

#### Open Science Collaboration (2015)

##### Sampling frame

Open Science Collaboration (2015) performed a total of 100 replications. They attempted to obtain a quasi-random sample by sampling from articles published in 2008 from the journals Psychological Science, Journal of Personality and Social Psychology, and Journal of Experimental Psychology: Learning, Memory, and Cognition. Replication teams were instructed to select from a subset of the sampling frame (initially just the first 20 published articles from each journal were available, a pool that was expanded as it became difficult to match teams with articles). Project coordinators also actively recruited replication teams from the community that they thought had expertise to assess particular claims. Of the 488 articles that were published in the applicable journal issues, 158 (32%) were eligible for selection. A total of 84 of the 100 replications conducted were of the final experiment in each original article. Sixteen replications deviated from this rule and used a different experiment due to a recommendation from the original authors or on the basis of feasibility. The statistical test that was chosen to reflect the replication outcome was presented to the original authors as part of the design protocol, and was changed at the request of the original authors.

Open Science Collaboration (2015) reports that 47 articles were eligible for replication and were not claimed. Six of these studies were deemed infeasible due to practical constraints such as cost, historical dependence, or difficult to recruit samples, and were not presented to potential replicators. The remaining 41 were presented to potential replicators, but were not claimed. Open Science Collaboration (2015) reports that these projects “often” required access to rare populations, were likely to be expensive, required specialized materials (e.g., MRI studies), or access to specialized knowledge.

##### Participants

Each replication study was powered on the basis of the original study's observed effect size. The average power to detect the original effect of the 97 studies with significant results was .92. Individual studies had sample sizes (at the individual level) of between nine and 455306 (mean n = 5229.11, median = 70, sd = 48250.51).

##### Extraction and inclusion rules

Three original studies which did not report that their findings were indicative of a non-zero effect were excluded from those studies extracted from Open Science Collaboration (2015). Three studies for which z transformed correlation coefficients could not be extracted due to missing data in the downloaded data set were also excluded from analysis (these included 1 study which used multiple statistical tests in the original and replication studies, one study in which the replication and original study used different statistical tests, and one study for which the effect size was reported as a beta coefficients without test statistics or degrees of freedom). Effect sizes for original and replication studies are included for 94 out of 97 studies replicated studies from Open Science Collaboration (2015) which reported having found a non-zero effect.

#### The Nature Science Replication Project (Camerer et al., 2018)

##### Sampling frame and reported deviations

Camerer et al. (2018) replicated all 21 studies that were (a) published between 2010-2015, (b) in the journals Nature or Science, (c) tested for an experimental treatment effect, (d) had “at least one clear hypothesis with a statistically significant finding”, and that (e) “were performed on students or other accessible subject pools”. In order to choose which of the original studies’ statistical tests to use as the replication outcome, they selected the statistical tests that they subjectively identified as the most important in the first reported study that reported a significant treatment effect. When they could not identify a single most important effect they randomly selected from those that they identified as potential candidates. Camerer et al. (2018) report that there were no deviations from original studies’ protocols for seven replications, “minor” deviations for 12, and an “unintended” deviation for one study.

##### Participants

Camerer et al. (2018) took a two-stage sampling procedure, where each study was initially replicated with a sample large enough to ensure 90% power to detect an effect 75% of the size of the original study’s observed effect size. If there were no significant results in this first study, a second study was performed with 90% power to detect 50% of the original effect size. For one study, there was a significant result in the same direction as the original study in the first wave of replication data collection, but a second wave was still performed due to an error.

##### Extraction and inclusion rules

Original and replication effect sizes were extracted for all 21 studies included in Camerer et al. (2018). In some cases in the Nature Science reproducibility project multiple replication studies were performed for a single effect. In each of these cases we performed a fixed effects meta-analysis using the metafor package (Viechtbauer, 2010) to estimate a meta-analytic effect size estimate. The effect size, standard errors and sample sizes used in the current study reflect this pooled estimate. This method leads to one study more “replicating” according to the ‘statistical significance in the same direction of the original study’ criterion than was originally reported in Camerer et al. (2018), where they used the p value from the largest performed study instead of a pooled estimate.

#### LOOPR (Soto, 2019)

##### Sampling frame

Soto (2019) extracted 78 trait-outcome associations (i.e., associations between The Big Five personality traits and life outcomes like volunteerism or politically conservatism) from a review article that presented a total of 86 associations between the Big Five traits and life outcomes (Ozer & Benet-Martinez, 2006). Soto and a research assistant extracted the empirical evidence behind these associations, and selected the 78 that were replicated on the basis of feasibility. In Soto (2019) the 78 trait-outcome associations were the primary units of analysis for estimating replicability, whereas in the current study we used the study level for comparability with the other replication projects. In total 121 studies were included in the current analysis.

##### Deviations from original study protocols

All replications were surveys administered using Qualtrics. The replications were all conducted in one of two large surveys. A single measure of the Big 5 personality traits was administered rather than the disparate versions used in each individual original study. The outcomes were assessed using measures that Soto reports were “selected to follow the original studies as closely as was feasible”. Soto reports that most outcome measures were administered using the same measure, but some interview questions were altered to fit a survey format and some lengthy questions or measures were shortened. As this means that some replication studies used shorter form versions of the original data collection instruments, all significance tests or effect size analyses in the current paper use the effect sizes that Soto (2019) reports have been disattenuated using the Spearman-Brown prediction formula and Spearman disattenuation formula to estimate the trait-outcome associations that would be expected if the measures had used the same number of items as the original study (Lord & Novick, 1968).

##### Participants

Two young adult (18-25 year olds) and two adult samples (18 years of age or older) were recruited using quota sampling. Quota sampling was employed in order to match the US population in terms of sex, race and ethnicity in both the adult and young adult samples, and additionally for age, educational attainment and household income in the adult samples. Samples of 1559 adults and 1550 young adults completed Survey 1 (including approximately half of the replication studies), and samples of 1512 adults and 1505 young adults completed Survey 2 (which included the remaining replication studies). Soto reports that this yielded a minimum power of 97.3% to detect a correlation of just .1 using a two-tailed test with an alpha of .05.

##### Extraction and inclusion rules

Effect sizes were extracted for original and replication studies for 101 out of 121 included studies, and one original study’s sample size was not available. In Soto (2019) effect sizes which were only reported in this dataset as beta coefficients were not converted to Fisher z scores as not enough information was available in Soto’s data set to do so. A total of 100 of 121 studies were included in the current analysis. Following the other large scale replication studies, the signs of the original and replication study effects were inverted for analysis if the original effect was negative.

#### Experimental philosophy replication project (Cova et al., 2018)

##### Sampling frame and deviations from original protocols

Cova et al., (2018) sampled their replication targets from a list of experimental philosophy papers hosted by Yale’s University’s Experimental Philosophy Lab (<http://experimental-philosophy.yale.edu/ExperimentalPhilosophy.html>), and filtered these down to articles published in one of 35 scientific journals that Cova and Strickland judged would likely have published experimental philosophy research. From the remaining articles, they sampled three papers per year for the years 2003 to 2015, selecting the most cited paper (according to Google scholar) and two random studies (barring the year 2003 where there were only two applicable papers). This left them with a list of 38 studies which were assessed for feasibility. They report that four studies were identified as being potentially difficult to replicate (as they required access to special populations, three papers requiring access to non-English speakers and one requiring participants with high-functioning autism). In order to ensure that they achieved at least their goal number of studies, they sampled an additional replacement for each of the original studies they were concerned might be practically infeasible (either picking the second most cited article of the year if the potentially infeasible study was selected as it was the most cited, or an additional random article), but continued to attempt to replicate all studies. In the end, two of the studies that were identified as being potentially infeasible were replicated, and two were not. The “more feasible” replication studies were also included in the final analysis if they were finished by the end of the project’s duration. This left a total of 40 completed replications, 12 replications of most the most cited article of the years in the sampling frame, two replications of the second most cited article in a given year, and 26 studies which were randomly chosen from each year’s set of experimental philosophy studies.

Cova et al. (2018) report that there was a major deviation in protocol for one study, where text was presented on screen for a set duration instead of being removed after participants responded, which they suggest may have reduced participants' incentive to quickly respond.

##### Participants

The research teams were based in Brazil, France, Lithuania, Netherlands, Spain, Switzerland, United Kingdom and the United States. Most studies were replicated using online samples (35 studies) and four used university students, unlike the original studies of which 25 used collage samples, 10 used online samples, and four involved local pedestrians. Cova et al., (2018) report that their average replication sample size was 206.3, sd = 131.8. For the 32 studies reporting a significant result for which they could perform a power analysis, the average level of statistical power to detect the original effect size was 0.88 (SD = 0.14), 26 had a power > 0.80 to detect the original sample size, and 18 had a power > .95.

##### Extraction and inclusion rules

Cova et al. (2018) included three replications of original studies which were non-significant (and which did not claim to provide evidence for the effects under test), these were removed from this analysis. Effect sizes were reported by Cova et al. (2018) and are included in the current study for 33, original and replication studies, out of an original 37 replicated studies with significant original results. The four studies for which no effect sizes were reported performed analyses for which Cova et al. (2018) could not develop reasonable effect size estimates from the reported statistics or data (e.g., a Sobel test, GEE analysis).

#### Many labs 1 (R. Klein et al., 2014)

##### Sampling frame

Many Labs 1 (R. Klein et al., 2014) replicated a non-random sample of 13 effects from 12 original articles. They report that the replicated studies were chosen on the basis of being (a) suitable for online presentation, (b) able to be presented “quickly” and (c) to have a simple, two condition design (except for a single a correlational study which was included). The replication materials were translated into the local language of each sample.

##### Participants

Klein et al., (2014) replicated each effect across 36 samples and settings (27 laboratory, 9 online, and 25 in the USA and 11 in non-US locations) including a total of 6344 participants. The individual studies each included between 4896 and 6336 participants.

##### Extraction and inclusion rules

One of the original studies in Many labs 1 (R. Klein et al., 2014) did not report an effect size or test statistic so is not included in the current sample. No effect size was extractable for one original study, and this effect was excluded for the purposes of the current analysis. Four different operationalisations of anchoring effects were tested, all of which are included in the current analysis, leading to a total of 15 paired data-points being included from this study. The multilevel models reported account for non-independence between these effects.

#### Many labs 2 (R. A. Klein et al., 2018)

##### Sampling frame

Many labs 2 (R. A. Klein et al., 2018) replicated a non-random sample of 28 effects. In order to select the claims that were replicated, they (a) held an open nomination round that fit a set of desired characteristics, elicited ideas within the project team, and asked “independent experts” who worked in the field of psychological science. The criteria for nominations were that the study must “(a) have procedures that can be implemented via a web browser, (b) not have more than two conditions in an experimental design, (c) have one outcome (dependent variable), (d) be correlational or experimental designs, though the latter is preferred, and (e) examine a psychological topic”. There were also a set of desired characteristics, that the study should: “(a) be “important” in at least one of the many variety of ways of demonstrating importance, (b) be brief, (c) be a direct replication of an existing published design from which an original effect size can be determined)" (see <https://osf.io/uazdm/> for the coordinating proposal).

After over 100 claims were nominated, the Klein et al., chose the replicated claims on the basis of the claims being (a) suitable for online presentation, (b) able to be presented quickly, (c) citation count (more being judged more positively), (d) to have a simple, two condition design or, at a lower priority, a correlational design, (e) to have a high level of “general interest”, and (f) to be applicable to a general sample of adults. Secondarily, Klein et al., report that they aimed to ensure a diversity of claims, including (a) claims that had both limited and extensive literatures across samples and settings, (b) effects for which it was expected that there would be high and low effect size heterogeneity, (c) “classic” and “contemporary” effects, (d) effects which cover a broad range of subareas within social and cognitive psychology, (e) effects from a diversity of research teams, and finally (f), effects from a variety of journals.

For all chosen claims, the original study’s corresponding author were contacted (if possible) and asked for the original materials, and asked if there were any limitations or moderators that the replication teams should be aware of before attempting the replication. Klein et al., report that this process eliminated some studies, at which point the effect was eliminated from consideration and replaced with another of the nominated claim or the original study’s authors were asked to help identify alternative studies for replication. In one case, the original authors expressed “strong concerns” about the study being included in this project, and this claim was removed.

After this process, 31 effects remained, which were split into two 30 minutes “slates”, and pilot tested. Three effects were removed for length, and some studies were shortened in order to fit into the two 30 minutes slates after pilot testing. This left 28 effects that were replicated.

##### Participants

Labs were invited to participate in an open call made in 2014. The only criteria for inclusion was that a lab had to agree to include at least 80 participants. A total of 125 samples were recruited in this study. The total sample size was 15305, with ns of at least 3549 in each replication study after exclusions (mean n = 6965.75, sd = 1040.89, median = 7157)

##### Extraction and inclusion rules

A total of 22 of 28 paired original and replication effects sizes are included for this analysis. Four studies from (R. A. Klein et al., 2018) were removed because the original and replication studies examined a difference in effect sizes seen in different conditions, and the effects were not directly tested against each other making it difficult to derive an appropriate effect size. Two additional studies were excluded because their effect sizes were only available as Cohen’s q.

#### Many labs 3 (Ebersole et al., 2016)

##### Sampling frame

A non-random sample of nine original studies were replicated, and one conceptual replication was performed. The conceptual replication was excluded from the current analyses. Ebersole et al., report that they aimed to identify two-condition experiments or correlational designs. Seven out of the nine included studies were included in a single computer delivered experiment, and were followed by extensive individual difference and data quality indicators. Two experiments were exclusively performed in person as, in one case, the study required participants to physically hold a clipboard, and in one as the original authors said that the experiment needed to be administered using a paper and pencil format.

##### Participants

The replication included 20 university participant pools at universities in the United states and Canada (N = 2,696) and an mTurk online sample (N = 737). Individual studies had ns of at least 1335 after exclusions (mean n = 2752, sd = 642.85, median = 3088)

##### Extraction and inclusion rules

Original and replication effect sizes were extracted for all 9 original and replication studies from (Ebersole et al., 2016), excluding a study they term a “conceptual replication”. Most effects (6/9) were converted to correlation coefficients from the Cohen’s d values reported in this replication project. The results of three additional studies reported as partial Eta squared were converted to correlation coefficients from F statistics using the formula:

#### Economics replication project (Camerer et al., 2016)

##### Sampling frame and deviations

Camerer et al., (2016) replicated all 18 between-subject laboratory experimental papers published in the American Economic Review and the Quarterly Journal of Economics between 2011 and 2014. For 17 of 18 studies, the same software was used to conduct the replications as was used in the original experiment. In one study the software used was different as the original program was no longer available.

##### Participants

Two American teams, one Swedish, and one Austrian replication team conducted four or five replications each. Camerer (2016) reports that the replication sample sizes were chosen to ensure that each replication study had at least 90% power to detect the original study’s effect size using the original analysis strategy. Individual studies had ns of at least 40 (mean n = 176.22, sd = 98.93, median = 150).

##### Extraction and inclusion rules

The economics replication project Camerer et al. (2016). Original and replication effect sizes for all 18 studies were reported in correlation coefficients and all are included in this analysis.

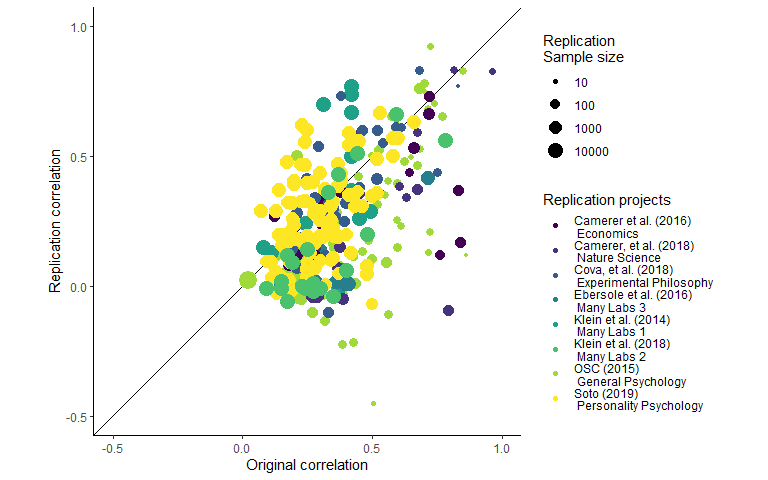
**Supplementary Material 2**

**Plots and multilevel model output of the relationship between original and replication correlation coefficients using varied exclusion criteria**

The following output shows scatter plots and model output for all of the multilevel meta-analyses performed using the varied exclusion criteria explained in the main text.

Table SM 1. Multilevel meta-analysis model estimates and random effects for all data.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Estimate | 95% CI LB | 95% CI UB | SE | p | Random effects |
| -0.14 | -0.21 | -0.07 | 0.04 | < .001 |  |
|  |  |  |  |  | Project variance = 0.008, n = 8 |
|  |  |  |  |  | Article variance = 0.016, n = 229 |
|  |  |  |  |  | Effect variance = 0.012, n = 306 |
|  |  |  |  |  | QE(305) = 3531.9, p < .001 |

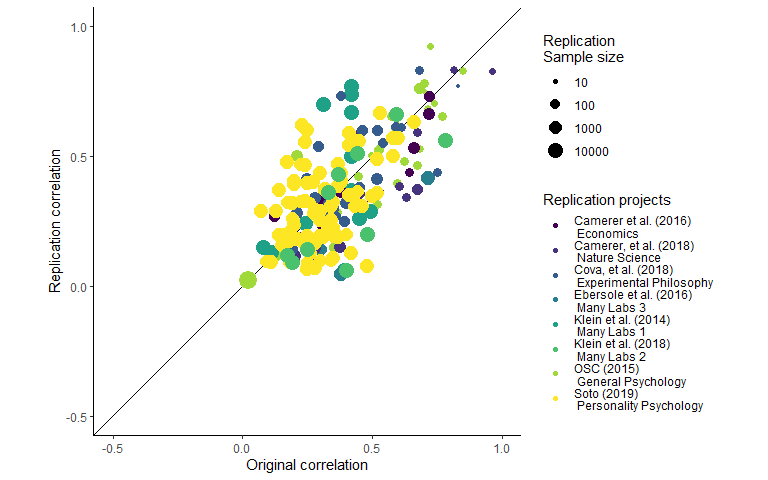


*Figure SM1.* Scatter plot of replication effect sizes (in correlation coefficients) plotted against original effects including all data.

Table SM2.

*Multilevel meta-analysis model estimates and random effects including only statistically significant replications.*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Estimate | 95% CI LB | 95% CI UB | SE | p | Random effects |
| -0.05 | -0.11 | 0.01 | 0.03 | 0.1 |  |
|  |  |  |  |  | Project variance = 0.005, n = 8 |
|  |  |  |  |  | Article variance = 0.014, n = 132 |
|  |  |  |  |  | Effect variance = 0.009, n = 198 |
|  |  |  |  |  | QE(197) = 2715.24, p < .001 |

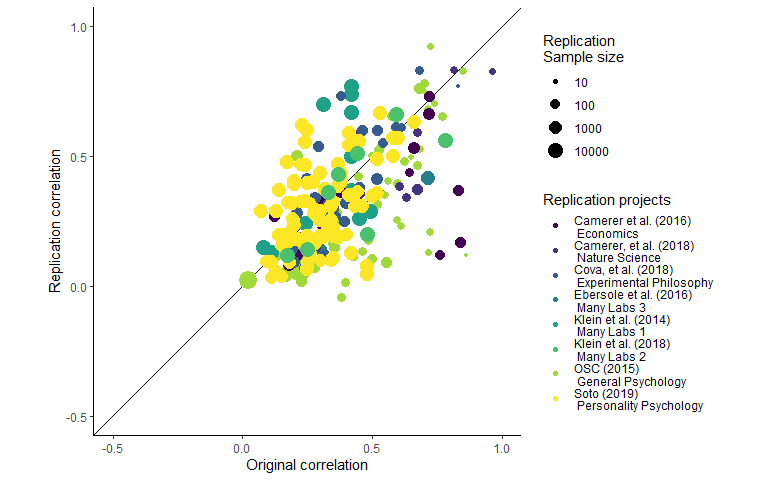


*Figure SM2.* Scatter plot of replication effect sizes (in correlation coefficients) plotted against original effects including only statistically significant replications.

Table SM3.

*Multilevel meta-analysis model estimates and random effects including studies which are not statistically equivalent to the null, using equivalence bounds set as the minimum effect size that would have been statistically significant in the original study.*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Estimate | 95% CI LB | 95% CI UB | SE | p | Random effects |
| -0.08 | -0.15 | -0.01 | 0.04 | 0.03 |  |
|  |  |  |  |  | Project variance = 0.008, n = 8 |
|  |  |  |  |  | Article variance = 0.018, n = 169 |
|  |  |  |  |  | Effect variance = 0.009, n = 237 |
|  |  |  |  |  | QE(236) = 3031.58, p < .001 |



*Figure SM3.* Scatter plot of replication effect sizes (in correlation coefficients) plotted against original effects including studies which are not statistically equivalent to the null, using equivalence bounds set as the minimum effect size that would have been statistically significant in the original study.

**Supplementary materials 3**

**Leave one out cross validation output**

Table [SM4](file:///C:\Users\fsingletonthorn\Documents\PhD\Effect%20size%20adjustment%20testing%20paper\SimplifiedEffectSizeAdjustment%2029%20Jan%202018.docx#loo-cross-validation-output).

*0th, 25th, 50th, 75th and 100th percentiles from leave one out cross validation for each multilevel model, for each exclusion method an, including only the sample indicated in “LOO exclusions”.*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| LOO exclusions | Subsample | Proportion significant | Minimum estimate | 25th percentile | Median | 75th percentile | Maximum estimate |
| Replication project | Only non-equivalent replications | 0.38 | -0.10 | -0.09 | -0.08 | -0.07 | -0.06 |
| Replication project | Only significant replications | 0.12 | -0.07 | -0.06 | -0.06 | -0.04 | -0.03 |
| Replication project | P value as Moderator | 1.00 | -0.10 | -0.09 | -0.07 | -0.07 | -0.06 |
| Replication project | All data | 1.00 | -0.16 | -0.15 | -0.13 | -0.13 | -0.12 |
| Study | All data | 1.00 | -0.14 | -0.14 | -0.14 | -0.14 | -0.13 |
| Study | Only significant replications | 0.01 | -0.06 | -0.05 | -0.05 | -0.05 | -0.04 |
| Study | Only non-equivalent replications | 1.00 | -0.09 | -0.08 | -0.08 | -0.08 | -0.07 |
| Study | P value as Moderator | 1.00 | -0.09 | -0.08 | -0.08 | -0.08 | -0.07 |
| Effect | Only significant replications | 0.00 | -0.06 | -0.05 | -0.05 | -0.05 | -0.04 |
| Effect | Only non-equivalent replications | 1.00 | -0.09 | -0.08 | -0.08 | -0.08 | -0.07 |
| Effect | P value as Moderator | 1.00 | -0.09 | -0.08 | -0.08 | -0.08 | -0.07 |
| Effect | All data | 1.00 | -0.14 | -0.14 | -0.14 | -0.14 | -0.13 |

**Supplementary materials 4**

**Bayesian Mixture Model**

The mixture model results presented in text presents the model developed by Camerer et al,. (2018; see <https://osf.io/xhj4d/> for their detailed description of this model). All priors were chosen to be uninformative or vague. The mixture model assumes that the observed replication effect sizes either come from the null hypothesis, a true effect sampled from a normal distribution with a mean of zero and an estimated precision (tau). This model uses an errors-in-variables approach to account for possible attenuation of effect sizes due to measurement error and estimation uncertainty following Matzke et al. (2017), which means the effect size attenuation factor is the factor change between the estimated true effect of the original and replication study effect size.

Box SM1. The original model reported in Camerer et al. (2018) and reported on in the main text of the current article.

Model{# Mixture Model Priors:alpha ~ dunif(0,1) # flat prior on slope for predicted effect size under H1tau ~ dgamma(0.001,0.001) # vague prior on study precisionphi ~ dbeta(1, 1) # flat prior on the true effect rate# prior on true effect size of original studies:for (i in 1:n){trueOrgEffect[i] ~ dnorm(0, 1)}# Mixture Model Likelihood:for(i in 1:n){clust[i] ~ dbern(phi)# extract errors in variables (FT stands for Fisher-transformed):orgEffect\_FT[i] ~ dnorm(trueOrgEffect[i], orgTau[i])repEffect\_FT[i] ~ dnorm(trueRepEffect[i], repTau[i])trueRepEffect[i] ~ dnorm(mu[i], tau)# if clust[i] = 0 then H0 is true; if clust[i] = 1 then H1 is true and# the replication effect is a function of the original effect:mu[i] <- alpha \* trueOrgEffect[i] \* equals(clust[i], 1)# when clust[i] = 0, then mu[i] = 0;# when clust[i] = 1, then mu[i] = alpha \* trueOrgEffect[i] }  
}

**Supplementary materials 5**

**Conversions**

All statistical tests extracted were transformed into correlation coefficients as follows, using the methods reported in (Open Science Collaboration, 2015).

t statistics:

Where is the observed t statistic and is the degrees of freedom of the t test.

F statistics:

Where is the observed F statistic and is the degrees of freedom of the numerator and is degrees of freedom of the denominator.

Chi square statistics:

Where is the observed statistic and is the associated degrees of freedom.

All values were then transformed into fisher Z transformed correlation coefficients using:

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., . . . Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science, 351*(6280), 1433. doi:10.1126/science.aaf0918

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., . . . Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour, 2*(9), 637-644. doi:10.1038/s41562-018-0399-z

Cova, F., Strickland, B., Abatista, A., Allard, A., Andow, J., Attie, M., . . . Colombo, M. (2018). Estimating the reproducibility of experimental philosophy. *Review of Philosophy and Psychology*, 1-36.

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., . . . Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology, 67*, 68-82. doi:<https://doi.org/10.1016/j.jesp.2015.10.012>

Klein, R., Ratliff, K., Vianello, M., Adams Jr, R., Bahník, S., Bernstein, M., . . . Brumbaugh, C. (2014). Data from investigating variation in replicability: A “many labs” replication project. *Journal of Open Psychology Data, 2*(1).

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., . . . Nosek, B. A. (2018). Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. *Advances In Methods and Practices in Psychological Science, 1*(4), 443-490. doi:10.1177/2515245918810225

Matzke, D., Ly, A., Selker, R., Weeda, W. D., Scheibehenne, B., Lee, M. D., & Wagenmakers, E.-J. (2017). Bayesian inference for correlations in the presence of measurement error and estimation uncertainty. *Collabra: Psychology, 3*(1). doi:<http://doi.org/10.1525/collabra.78>

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251). Retrieved from <http://science.sciencemag.org/content/349/6251/aac4716.abstract>

Soto, C. J. (2019). How replicable are links between personality traits and consequential life outcomes? The Life Outcomes Of Personality Replication Project. *Psychological Science, 30*(5), 711-727.

Viechtbauer, W. (2010). Conducting Meta-Analyses in R with the metafor Package. *Journal Of Statistical Software, 36*(3), 48. doi:10.18637/jss.v036.i03