

Tables

Heterogeneity in direct replications in psychology and its association with effect size

Table 1

Effect size ρ_{xy} and its heterogeneity as a function of true effect size and measurement reliability.

| | $\rho_{xy} = 0$ | $\rho_{xy} = .3$ | $\rho_{xy} = .5$ |
|---|-----------------|------------------|------------------|
| $\sqrt{R_{xx'}} \times \sqrt{R_{yy'}} = .6$ | 0 | 0.18 | 0.30 |
| $\sqrt{R_{xx'}} \times \sqrt{R_{yy'}} = .7$ | 0 | 0.21 | 0.35 |
| $\sqrt{R_{xx'}} \times \sqrt{R_{yy'}} = .8$ | 0 | 0.24 | 0.40 |

Note:

Values in cells are observed effect sizes arising from the true effect size ρ_{xy} and measurement reliabilites $\sqrt{R_{xx'}} \times \sqrt{R_{yy'}}$. Code to reproduce table: osf.io/kf6pt/

Table 2

Pre-registered multi-lab replication projects

| RP | Paper | Countries | K (US) | Effects | Participants | Sample and Settings | Description of Effects |
|------|---------------------------|-----------|---------|---------|--------------|--|--|
| ML1 | Klein et al. (2014) | 10 | 36 (25) | 16 | 5975 | 26/36 samples were primarily university students, 3 general population and 7 undescribed. 9/36 samples were online, including all the general population ones. | Two correlational effects: 'Gender math attitude' compared implicit attitudes (IAT) towards math between genders and 'IAT correlation math' correlated implicit attitudes with self-reported measures. The remainder were experiments with two independent groups. The groups were primed in some way (Anchoring 1-4; low vs. high category scales; norm of reciprocity; flag priming; currency priming), asked to imagine slightly different situations (Sunk costs; gain vs. loss framing; gambler's fallacy; imagined contact) or asked their agreement with statements presented differently (Allowed vs. forbidden; quote attribution). |
| ML3 | Ebersole et al. (2016) | 2 | 21 (19) | 10 | 2845 | 20/21 samples were university students, 1 general population which was also the only online sample. | Several effects were experiments with two independent groups. The groups were either primed in some way (Power and perspective; warmth perceptions; subjective distance interaction), saw slightly different statements (Elaboration likelihood interaction; credentials interaction) or experienced different situations (weight embodiment). Examined interactions were between treatment conditions and participant characteristics. One priming effect (metaphor) compared two treatment groups with a control. One effect was correlational: 'persistence and conscientiousness' was measured by an unsolvable anagram task and self-report respectively. The Stroop task is a within-person experiment with two conditions and the 'Availability' effect asks participants to judge whether some letters are more common in the first or third position. |
| RRR1 | Alogna et al. (2014) | 10 | 32 (17) | 1 | 4117 | 31/32 samples were undergraduate students between 18-25, 1 general population which was also the only online sample. | Verbal overshadowing 1; Independent two-group experiment. Participants either described a robber after watching a video or listed countries/capitals and after a filler task attempted to identify the robber in a lineup. |
| RRR2 | Alogna et al. (2014) | 8 | 23 (14) | 1 | 2442 | 22/23 samples were undergraduate students between 18-25, 1 general population which was also the only online sample. | Verbal overshadowing 2; Different from 1 only in that the filler task took place before the descriptive task instead of after. |
| RRR3 | Eerland et al. (2016) | 2 | 12 (10) | 3 | 1187 | 11/12 samples were undergraduate students mostly between 18-25, one of which was online. 1 sample was a broader online sample. | Grammar on intentionality/intention attribution/detailed processing; Independent two-group experiment with three outcome variables. Actions either described in imperfect or perfect. |
| RRR4 | Hagger et al. (2016) | 10 | 23 (7) | 1 | 2872 | All samples consisted of in-lab undergraduate students | Ego depletion; Independent two-group experiment. Participants either assigned to a cognitively demanding task or a neutral, and performance was then measured in a subsequent cognitive task. |
| RRR5 | Cheung et al. (2016) | 5 | 16 (9) | 2 | 2071 | All samples consisted of in-lab undergraduate students between 18-25 | Commitment on neglect/exit; Independent two-group experiment with two outcome variables. Participants either primed to think about commitment to or independence from partner. |
| RRR6 | Wagenmakers et al. (2016) | 8 | 17 (8) | 1 | 1894 | All but one sample explicitly consisted of students and all took place in-lab. The last sample was recruited at university grounds. | Facial feedback hypothesis; Independent two-group experiment. Participants either induced to 'smile' or 'pouth' with a pen and then rated funniness of cartoons. |

| | | | | | | | |
|------|---------------------------|----|--------|---|------|---|--|
| RRR7 | Bouwmeester et al. (2017) | 12 | 21 (5) | 1 | 3596 | All samples consisted of in-lab undergraduate students between 18-34. | Intuitive cooperation; Independent two-group experiment. Economic game with money contribution to a common pool either under time pressure or time delay. |
| RRR8 | O'Donnell et al. (2017) | 13 | 23 (9) | 1 | 4493 | All samples consisted of in-lab undergraduate students between 18-25 | Professor priming; Independent two-group experiment. Participants primed with either a 'professor' or 'hooligan' stimuli. Outcome was percentage correct trivia answers. |

Note: “For studies with several effects the number of participants is the average across effects, rounded to the closest whole number. Participant numbers are those used for primary analyses by original authors (i.e., after exclusions). RP = Replication Project, K (US) = no. primary studies (number of US studies), ML = Many Labs, RRR = Registered Replication Report. Code to reproduce table: osf.io/kf6pt/”

Table 3

Heterogeneity across primary effects and statistical power of ten multi-lab replication projects, ordered with respect to estimated between studies variance (τ^2).

| RP | Effect | K | Effect type | Effect size estimate | I^2 (%) | I^2 95% CI | r^* | τ^2 | τ^2 95% CI | Type I Error Rate & Statistical Power | | | |
|------|------------------------------------|----|-------------|----------------------|-----------|----------------|-------|----------|-----------------|---------------------------------------|-------|--------|-------|
| | | | | | | | | | | Level of heterogeneity | | | |
| | | | | | | | | | | Zero | Small | Medium | Large |
| ML1 | Anchoring 3 - Everest | 36 | SMD. | 2.41 | 91.29 | [86.61, 95.23] | 0.95 | 0.00 | [0.01, 0.03] | 0.04 | 0.46 | 0.91 | 1.00 |
| ML1 | Allowed vs. forbidden | 36 | SMD. | 1.93 | 75.56 | [60.32, 85.46] | 0.88 | 0.01 | [0.01, 0.04] | 0.05 | 0.47 | 0.91 | 1.00 |
| ML1 | Anchoring 2 - Chicago | 36 | SMD. | 2.00 | 75.36 | [61.11, 87.15] | 0.89 | 0.01 | [0.00, 0.01] | 0.05 | 0.44 | 0.92 | 1.00 |
| ML1 | Anchoring 4 - Babies | 36 | SMD. | 2.53 | 64.67 | [45.67, 83.33] | 0.97 | 0.00 | [0.00, 0.00] | 0.05 | 0.47 | 0.92 | 1.00 |
| ML1 | Quote Attribution | 36 | SMD. | 0.31 | 52.05 | [24.63, 76.25] | 0.20 | 0.01 | [0.00, 0.03] | 0.04 | 0.43 | 0.91 | 1.00 |
| ML1 | Anchoring 1 - NYC | 36 | SMD. | 1.21 | 40.23 | [10.62, 73.94] | 0.68 | 0.01 | [0.00, 0.02] | 0.05 | 0.45 | 0.92 | 1.00 |
| ML1 | IAT correlation math | 35 | r | 0.39 | 40.05 | [3.93, 64.97] | 0.39 | 0.00 | [0.00, 0.01] | 0.05 | 0.40 | 0.91 | 1.00 |
| RRR3 | Grammar on intentionality | 12 | MD | -0.25 | 38.06 | [0.00, 85.72] | -0.11 | 0.01 | [0.00, 0.08] | 0.06 | 0.22 | 0.68 | 0.97 |
| ML3 | Subjective Distance interaction | 21 | r | 0.02 | 33.51 | [0.00, 76.78] | 0.02 | 0.00 | [0.00, 0.02] | 0.05 | 0.33 | 0.83 | 0.99 |
| ML1 | Gender math attitude | 35 | SMD. | 0.57 | 28.06 | [0.00, 67.34] | 0.35 | 0.01 | [0.00, 0.02] | 0.05 | 0.44 | 0.90 | 1.00 |
| ML3 | Credentials interaction | 21 | r | 0.02 | 24.03 | [0.00, 73.82] | 0.02 | 0.00 | [0.00, 0.02] | 0.05 | 0.30 | 0.81 | 1.00 |
| ML1 | Gambler's Fallacy | 36 | SMD. | 0.61 | 22.85 | [0.00, 69.16] | 0.38 | 0.00 | [0.00, 0.02] | 0.05 | 0.44 | 0.91 | 1.00 |
| ML1 | Imagined Contact | 36 | SMD. | 0.12 | 20.60 | [0.00, 62.50] | 0.07 | 0.00 | [0.00, 0.02] | 0.05 | 0.44 | 0.91 | 1.00 |
| ML1 | Low vs. high category scales | 36 | SMD. | 0.88 | 19.20 | [0.00, 49.95] | 0.58 | 0.04 | [0.03, 0.08] | 0.04 | 0.46 | 0.92 | 1.00 |
| RRR8 | Professor priming | 23 | MD | 0.14 | 17.32 | [0.00, 64.77] | 0.01 | 0.00 | [0.00, 0.02] | 0.05 | 0.34 | 0.83 | 1.00 |
| ML1 | Norm of reciprocity | 36 | SMD. | -0.36 | 17.21 | [0.00, 47.51] | -0.22 | 0.01 | [0.00, 0.02] | 0.05 | 0.43 | 0.91 | 1.00 |
| ML3 | Metaphor | 20 | r | 0.14 | 13.03 | [0.00, 57.02] | 0.14 | 0.00 | [0.00, 0.02] | 0.05 | 0.32 | 0.80 | 0.99 |
| RRR1 | Verbal overshadowing 1 | 32 | RD | -0.03 | 12.23 | [0.00, 46.51] | -0.05 | 0.00 | [0.00, 0.02] | 0.06 | 0.38 | 0.90 | 1.00 |
| ML1 | Sunk Costs | 36 | SMD. | 0.29 | 9.18 | [0.00, 45.93] | 0.19 | 0.00 | [0.00, 0.01] | 0.05 | 0.44 | 0.91 | 1.00 |
| RRR7 | Intuitive-cooperation | 21 | MD | -0.39 | 2.80 | [0.00, 39.28] | -0.01 | 0.00 | [0.00, 0.01] | 0.05 | 0.32 | 0.83 | 1.00 |
| ML3 | Availability | 21 | r | 0.04 | 0.51 | [0.00, 56.09] | 0.04 | 0.00 | [0.00, 0.01] | 0.05 | 0.34 | 0.83 | 1.00 |
| ML1 | Gain vs. loss framing | 36 | SMD. | -0.66 | 0.01 | [0.00, 55.57] | -0.41 | 0.01 | [0.00, 0.02] | 0.05 | 0.43 | 0.91 | 1.00 |
| ML3 | Power and Perspective | 21 | SMD. | 0.03 | 0.01 | [0.00, 57.17] | 0.02 | 0.00 | [0.00, 0.02] | 0.05 | 0.32 | 0.81 | 0.99 |
| RRR3 | Grammar on intention attribution | 12 | MD | 0.00 | 0.00 | [0.00, 70.62] | 0.00 | 0.00 | [0.00, 0.04] | 0.06 | 0.24 | 0.70 | 0.96 |
| ML3 | Conscientiousness and persistence | 21 | r | 0.02 | 0.00 | [0.00, 61.42] | 0.02 | 0.00 | [0.00, 0.01] | 0.05 | 0.29 | 0.79 | 1.00 |
| RRR3 | Grammar on detailed processing | 12 | MD | -0.10 | 0.00 | [0.00, 54.49] | -0.05 | 0.00 | [0.00, 0.02] | 0.06 | 0.24 | 0.70 | 0.97 |
| RRR5 | Commitment on neglect | 16 | MD | -0.05 | 0.00 | [0.00, 53.18] | -0.03 | 0.00 | [0.00, 0.01] | 0.06 | 0.28 | 0.74 | 0.99 |
| ML3 | Warmth Perceptions | 21 | SMD. | 0.01 | 0.00 | [0.00, 47.10] | 0.01 | 0.00 | [0.00, 0.01] | 0.04 | 0.37 | 0.91 | 1.00 |
| RRR4 | Ego depletion | 23 | SMD. | 0.00 | 0.00 | [0.00, 46.91] | 0.00 | 0.00 | [0.00, 0.01] | 0.05 | 0.32 | 0.85 | 1.00 |
| ML1 | Flag Priming | 36 | SMD. | 0.02 | 0.00 | [0.00, 36.23] | 0.01 | 0.00 | [0.00, 0.01] | 0.05 | 0.43 | 0.90 | 1.00 |
| ML1 | Money Priming | 36 | SMD. | -0.02 | 0.00 | [0.00, 33.18] | -0.01 | 0.00 | [0.00, 0.01] | 0.05 | 0.44 | 0.91 | 1.00 |
| RRR2 | Verbal overshadowing 2 | 23 | RD | -0.15 | 0.00 | [0.00, 32.36] | -0.23 | 0.00 | [0.00, 0.01] | 0.06 | 0.31 | 0.83 | 1.00 |
| ML3 | Weight Embodiment | 20 | SMD. | 0.03 | 0.00 | [0.00, 29.97] | 0.02 | 0.00 | [0.00, 0.01] | 0.05 | 0.35 | 0.84 | 1.00 |
| RRR6 | Facial Feedback hypothesis | 17 | MD | 0.03 | 0.00 | [0.00, 25.13] | 0.01 | 0.00 | [0.00, 0.00] | 0.06 | 0.27 | 0.77 | 0.99 |
| ML3 | Elaboration likelihood interaction | 20 | r | 0.00 | 0.00 | [0.00, 18.62] | 0.00 | 0.00 | [0.00, 0.00] | 0.05 | 0.31 | 0.83 | 0.99 |
| RRR5 | Commitment on exit | 16 | MD | -0.06 | 0.00 | [0.00, 17.44] | -0.03 | 0.00 | [0.00, 0.00] | 0.06 | 0.27 | 0.77 | 0.99 |
| ML3 | Stroop effect | 21 | r | 0.41 | 0.00 | [0.00, 13.61] | 0.41 | 0.00 | [0.00, 0.00] | 0.05 | 0.29 | 0.80 | 0.99 |

Note:

Effects were estimated in metafor using REML. The following effects are odds ratios transformed into standardized mean differences: 'Allowed vs. forbidden', 'Gain vs. loss framing', 'Norm of reciprocity', 'Low vs. high category scales'. RP = Replication Project, k = no. primary studies, CI = confidence intervals, r^* = effect sizes as correlations and biserial correlations. Statistical power and type I error rates were simulated, where Zero = simulated type 1 error, and the other headers represent simulated power under small/medium/large heterogeneity ($I^2 = 25/50/75\%$) respectively. SMD = Standardized Mean difference (Hedge's g), MD = Mean Difference, RD = Risk Difference, r = correlation. Code to reproduce table: osf.io/kf6pt/

^a These effects were simulated as standardized mean differences