Tables

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

Table 1

Variation in observed effect sizes as a function of true effect size and measurement reliability.

	Observed Effect Sizes			
	Study 1	Study 2	Study 3	
Meta-Analysis	$\sqrt{R_{xx'}} \times \sqrt{R_{yy'}} = .6$	$\sqrt{R_{xx'}} \times \sqrt{R_{yy'}} = .7$	$\sqrt{R_{xx'}} \times \sqrt{R_{yy'}} = .8$	SD(ES)
I: $\mu = 0.0$	0.00	0.00	0.00	0.00
II: $\mu = 0.3$	0.18	0.21	0.24	0.03
III: $\mu = 0.5$	0.30	0.35	0.40	0.05

Note: The values under Study 1, 2 and 3 are observed effect sizes for that study given its measurement reliability $\sqrt{R_{xx'}} \times \sqrt{R_{yy'}}$ and the true effect size μ when within study sample size is infinite. SD(ES) is the standard deviation of the observed effect sizes for meta-analysis I, II and III; equivalent to heterogeneity given infinite within study sample sizes. Code to reproduce table: osf.io/kf6pt/?view_only=e6639d8d4f924739a7310782fbbb4e06

Table 2

Pre-registered multi-lab replication projects

RP	Paper	Countries	K (US)	Effects	N	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).
ML2	Klein et al. (2018)	35	115 (21)	28	6568	79/125 samples were collected in person (typically in labs), remainder online. Mean age in two rounds of data collection were 22.37 and 23.34 years.	Most effects were experiments with two independent groups. Often participants were primed in some way (Structure & Goal Pursuit, Priming Consumerism, Incidental Anchors, Position & Power, Moral Cleansing, Priming Warmth) or asked to imagine slightly different situations (SMS & Well-Being, Less is Better, Moral Typecasting, Intentional Side-Effects, Tempting Fate, Affect & Risk, Trolley Dilemma 1, Framing, Trolley Dilemma 2, Disgust & Homophobia, Choosing or Rejecting). Some groups saw slightly different statements (Correspondence Bias, Intuitive Reasoning), were asked to perform slightly different tasks (Direction & SES, Actions are Choices), or had to read a text with a clear vs. unclear font (Incidental Disfluency). Two correlational effects measured the correlations of Moral Foundations with political leaning, and Social Value Orientation with family size. Two effects examined order effects (Assimilation & Contrasts, Direction & Similarity). Finally, in False Consensus 1 and 2, participants made a binary choice and estimated how many people had made the same choice.
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: 'Conscientiousness and persistence' 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 aged 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.

(continued)

RP	Paper	Countries	K (US)	Effects	N	Sample and Settings	Description of Effects
RRR2	Alogna et al. (2014)	8	23 (14)	1	2442	22/23 samples were undergraduate students aged 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 aged 18-25, one of which was online. 1 sample was a broader online sample.	Grammar's effect on interpretation; Independent two-group vignette experiment with three outcome variables. Participants read about actions either described in imperfect or perfect tense and then rated protagonist's intentions (intentionality/intention attribution/detailed processing).
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 aged 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' by holding a pen in their mouth differently and simultaneously rated funniness of cartoons.
RRR7	Bouwmeester et al. (2017)	12	21 (5)	1	3596	All samples consisted of in-lab undergraduate students aged 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 aged 18-25	Professor priming; Independent two-group experiment. Participants primed with either a 'professor' or 'hooligan' stimuli. Outcome was percentage correct trivia answers.
RRR9	McCarthy et al. (2018)	13	22 (4)	1	5610	All samples consisted of in-lab students aged 18-25	Hostility priming; Independent two-group experiment. Participants descrambled sentences, either 20% or 80% were hostile, then rated ambiguous behavior.
RRR10	Verschuere et al. (2018)	12	19 (4)	1	2294	All samples consisted of in-lab students aged 18-25	Moral reminder; Independent two-group experiment. Participants either recalled the Ten Commandments or books they'd read. Outcome was degree of cheating when reporting results.

Note: For studies with several effects the number of participants is the average across effects, rounded to the closest whole number. N = Participants 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/?view_only=e6639d8d4f924739a7310782fbbb4e06"

Table 3

Heterogeneity across primary effects and statistical power of thirteen multi-lab replication projects, ordered with respect to estimated heterogeneity (I^2).

```
## [1] 12
```

```
## [1] 7
```

```
## [1] 17
```

```
## [1] 31
```

```
##         0%        25%        50%        75%       100% 
## 0.00000000 0.01382603 0.04741243 0.06788318 0.24703601
```

```
##           0%          25%          50%          75%         100% 
## 0.0000000000 0.0007530126 0.0901820876 0.1596068265 0.6925999787
```

```
## [1] 47
```

```
## [1] 47
```

```
## [1] 16
```

```
## [1] 16
```

									Type I Error Rate & Statistical Power			
									Level of heterogeneity			
RP	Effect	K	Effect type	Effect size estimate	I^2(%)	I^2 95% CI	$\hat{\tau}$	$\hat{\tau}$ 95% CI	Zero	Small	Medium	Large
ML2	Intentional Side-Effects	59	r	0.67	93.5	[91.7, 96.5]	0.148	[0.129, 0.205]	0.05	0.48	0.98	1.00
ML1	Anchoring 3 - Everest	36	SMD	2.41	91.3	[86.6, 95.2]	0.693	[0.544, 0.956]	0.05	0.42	0.92	1.00
ML2	Direction & SES	64	r	0.20	88.8	[84.1, 92.2]	0.247	[0.202, 0.301]	0.05	0.53	0.99	1.00
ML1	Allowed vs. forbidden	36	SMD	1.93	75.6	[60.3, 85.5]	0.496	[0.348, 0.685]	0.05^b	0.46^b	0.92^b	1.00^b
ML1	Anchoring 2 - Chicago	36	SMD	2.00	75.4	[61.1, 87.1]	0.358	[0.257, 0.533]	0.04	0.40	0.92	1.00
ML2	Moral Typecasting	60	r	0.45	72.9	[61.7, 82.8]	0.110	[0.085, 0.147]	0.05	0.58	0.98	1.00
ML2	Intuitive Reasoning	57	r	0.40	66.5	[54.4, 80.9]	0.103	[0.080, 0.150]	0.05	0.54	0.98	1.00
ML2	Less is Better	57	r	0.39	64.7	[48.8, 77.0]	0.099	[0.071, 0.133]	0.05	0.57	0.97	1.00
ML2	Moral Foundations	60	r	0.13	64.7	[49.1, 75.7]	0.091	[0.066, 0.118]	0.05	0.55	0.98	1.00
ML2	Correspondence Bias	58	r	0.69	64.7	[46.2, 73.1]	0.064	[0.044, 0.078]	0.05	0.57	0.98	1.00
ML1	Anchoring 4 - Babies	36	SMD	2.53	64.7	[45.7, 83.3]	0.298	[0.202, 0.492]	0.05	0.42	0.91	1.00
ML2	Actions are Choices	57	r	-0.11	63.9	[46.8, 76.0]	0.061	[0.043, 0.081]	0.05	0.52	0.98	1.00
ML2	Trolley Dilemma 1	59	r	0.59	54.1	[31.8, 66.2]	0.080	[0.050, 0.102]	0.05	0.54	0.99	1.00
ML1	Quote Attribution	36	SMD	0.31	52.0	[24.6, 76.3]	0.164	[0.090, 0.282]	0.05	0.45	0.91	1.00
ML2	Social Value Orientation	54	r	0.03	50.2	[28.2, 67.9]	0.069	[0.043, 0.100]	0.05	0.52	0.98	1.00
ML2	False Consensus 2	58	r	0.41	43.2	[18.1, 62.6]	0.063	[0.034, 0.093]	0.05	0.58	0.98	1.00
ML1	Anchoring 1 - NYC	36	SMD	1.21	40.2	[10.6, 73.9]	0.152	[0.064, 0.311]	0.05	0.44	0.91	1.00
ML1	IAT correlation math	35	r	0.39	40.0	[3.9, 65.0]	0.056	[0.014, 0.094]	0.05	0.40	0.92	1.00
RRR3	Grammar on intentionality	12	MD	-0.25	38.1	[0.0, 85.7]	0.227	[0.000, 0.708]	0.06	0.26	0.68	0.96
ML2	Priming Warmth	47	r	-0.01	36.8	[8.2, 62.7]	0.082	[0.032, 0.140]	0.05	0.51	0.97	1.00
ML2	Tempting Fate	59	r	0.11	36.5	[5.9, 53.6]	0.065	[0.021, 0.091]	0.05	0.58	0.98	1.00
ML3	Subjective Distance interaction	21	r	0.02	33.5	[0.0, 76.8]	0.059	[0.000, 0.151]	0.05	0.28	0.83	0.99
ML1	Gender math attitude	35	SMD	0.57	28.1	[0.0, 67.3]	0.112	[0.000, 0.258]	0.05	0.41	0.91	1.00
ML2	Choosing or Rejecting	41	r	-0.06	26.5	[0.0, 52.4]	0.047	[0.000, 0.083]	0.06	0.46	0.94	1.00
ML2	Incidental Anchors	49	r	0.03	24.9	[0.0, 54.7]	0.056	[0.000, 0.107]	0.05	0.49	0.97	1.00
ML3	Credentials interaction	21	r	0.02	24.0	[0.0, 73.8]	0.046	[0.000, 0.137]	0.05	0.30	0.80	1.00
ML1	Gambler's Fallacy	36	SMD	0.61	22.8	[0.0, 69.2]	0.090	[0.000, 0.248]	0.05	0.41	0.90	1.00
ML2	Moral Cleansing	52	r	0.01	22.3	[0.0, 51.5]	0.047	[0.000, 0.090]	0.05	0.53	0.98	1.00
ML1	Imagined Contact	36	SMD	0.12	20.6	[0.0, 62.5]	0.080	[0.000, 0.202]	0.05	0.44	0.91	1.00
ML1	Low vs. high category scales	36	SMD	0.88	19.2	[0.0, 50.0]	0.155	[0.000, 0.318]	0.05^b	0.44^b	0.92^b	1.00^b
RRR9	Hostility priming	22	MD	-0.08	18.0	[0.0, 56.3]	0.096	[0.000, 0.233]	0.05	0.34	0.82	1.00
RRR8	Professor priming	23	MD	0.14	17.3	[0.0, 64.8]	0.857	[0.000, 2.538]	0.06	0.33	0.82	1.00
ML1	Norm of reciprocity	36	SMD	-0.36	17.2	[0.0, 47.5]	0.091	[0.000, 0.190]	0.05^b	0.44^b	0.91^b	1.00^b
ML2	False Consensus 1	59	r	0.48	15.9	[0.0, 40.5]	0.032	[0.000, 0.061]	0.05	0.57	0.98	1.00
ML2	Assimilation & Contrast	59	q	-0.07	15.1	[0.0, 33.3]	0.078	[0.000, 0.131]	0.05	0.52	0.98	1.00
ML3	Metaphor	20	r	0.14	13.0	[0.0, 57.0]	0.047	[0.000, 0.141]	0.06	0.31	0.81	0.99
RRR1	Verbal overshadowing 1	32	RD	-0.03	12.2	[0.0, 46.5]	0.032	[0.000, 0.081]	0.05^b	0.34^b	0.82^b	0.99^b
ML2	Priming Consumerism	54	r	0.07	12.0	[0.0, 49.1]	0.035	[0.000, 0.093]	0.05	0.54	0.97	1.00
ML2	Trolley Dilemma 2	60	r	0.13	11.9	[0.0, 33.2]	0.036	[0.000, 0.069]	0.05	0.57	0.98	1.00
ML1	Sunk Costs	36	SMD	0.29	9.2	[0.0, 45.9]	0.050	[0.000, 0.145]	0.05	0.44	0.93	1.00

(continued)

									Type I Error Rate & Statistical Power			
									Level of heterogeneity			
RP	Effect	K	Effect type	Effect size estimate	I^2(%)	I^2 95% CI	$\hat{\tau}$	$\hat{\tau}$ 95% CI	Zero	Small	Medium	Large
ML2	Framing	55	r	0.22	5.9	[0.0, 36.5]	0.025	[0.000, 0.075]	0.06	0.55	0.98	1.00
ML2	Position & Power	59	r	0.01	3.1	[0.0, 42.2]	0.016	[0.000, 0.074]	0.05	0.58	0.98	1.00
ML2	Disgust & Homophobia	59	q	0.04	3.1	[0.0, 30.3]	0.035	[0.000, 0.131]	0.05	0.54	0.98	1.00
RRR7	Intuitive-cooperation	21	MD	-0.39	2.8	[0.0, 39.3]	0.911	[0.000, 4.321]	0.06	0.32	0.81	1.00
ML2	SMS & Well-Being	59	r	-0.01	1.8	[0.0, 29.8]	0.013	[0.000, 0.063]	0.05	0.55	0.98	1.00
ML3	Availability	21	r	0.04	0.5	[0.0, 56.1]	0.006	[0.000, 0.095]	0.05	0.33	0.82	1.00
ML2	Incidental Disfluency	66	r	-0.02	0.0	[0.0, 27.4]	0.001	[0.000, 0.061]	0.05	0.56	0.99	1.00
ML1	Gain vs. loss framing	36	SMD	-0.66	0.0	[0.0, 55.6]	0.002	[0.000, 0.205]	0.05[b]	0.44[b]	0.91[b]	1.00[b]
ML3	Power and Perspective	21	SMD	0.03	0.0	[0.0, 57.2]	0.002	[0.000, 0.198]	0.05	0.32	0.82	1.00
RRR3	Grammar on intention attribution	12	MD	0.00	0.0[a]	[0.0, 70.6]	0.001	[0.000, 0.185]	0.06	0.24	0.66	0.97
ML3	Conscientiousness and persistence	21	r	0.02	0.0[a]	[0.0, 61.4]	0.000[a]	[0.000, 0.104]	0.05	0.35	0.80	1.00
RRR3	Grammar on detailed processing	12	MD	-0.10	0.0	[0.0, 54.5]	0.000	[0.000, 0.246]	0.06	0.21	0.68	0.97
RRR5	Commitment on neglect	16	MD	-0.05	0.0	[0.0, 53.2]	0.000	[0.000, 0.208]	0.06	0.28	0.75	0.99
ML3	Warmth Perceptions	21	SMD	0.01	0.0	[0.0, 47.1]	0.000	[0.000, 0.158]	0.06	0.39	0.91	1.00
RRR4	Ego depletion	23	SMD	0.00	0.0	[0.0, 46.9]	0.000	[0.000, 0.169]	0.05	0.33	0.84	1.00
RRR10	Moral reminder	19	MD	0.11	0.0	[0.0, 44.1]	0.000	[0.000, 0.392]	0.06	0.31	0.79	0.99
ML1	Flag Priming	36	SMD	0.02	0.0	[0.0, 36.2]	0.000	[0.000, 0.118]	0.05	0.43	0.92	1.00
ML1	Money Priming	36	SMD	-0.02	0.0	[0.0, 33.2]	0.000	[0.000, 0.110]	0.05	0.48	0.92	1.00
RRR2	Verbal overshadowing 2	23	RD	-0.15	0.0	[0.0, 32.4]	0.000	[0.000, 0.065]	0.05[b]	0.31[b]	0.82[b]	0.99[b]
ML3	Weight Embodiment	20	SMD	0.03	0.0	[0.0, 30.0]	0.000	[0.000, 0.122]	0.06	0.34	0.83	1.00
RRR6	Facial Feedback hypothesis	17	MD	0.03	0.0	[0.0, 25.1]	0.000	[0.000, 0.164]	0.06	0.27	0.79	0.99
ML2	Affect & Risk	60	r	-0.04	0.0	[0.0, 21.1]	0.000	[0.000, 0.056]	0.05	0.57	0.99	1.00
ML3	Elaboration likelihood interaction	20	r	0.00	0.0	[0.0, 18.6]	0.000	[0.000, 0.042]	0.05	0.31	0.79	1.00
RRR5	Commitment on exit	16	MD	-0.06	0.0	[0.0, 17.4]	0.000	[0.000, 0.089]	0.06	0.29	0.74	0.99
ML3	Stroop effect	21	r	0.41	0.0	[0.0, 13.6]	0.000	[0.000, 0.027]	0.05	0.30	0.80	1.00
ML2	Structure & Goal Pursuit	52	r	-0.01	0.0	[0.0, 1.9]	0.000	[0.000, 0.013]	0.05	0.53	0.97	1.00
ML2	Direction & Similarity	49	r	0.01	0.0	[0.0, 0.0]	0.000	[0.000, 0.000]	0.05	0.54	0.97	1.00

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'. All ML2 meta-analyses with effect type 'r' except 'Moral foundations' and 'Social Value Orientation' were transformed to correlations from a variety of effect sizes. RP = Replication Project, K = no. primary studies, $\hat{\tau}$ = between studies standard deviation, CI = confidence intervals. Statistical power was 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/?view_only=e6639d8d4f924739a7310782fbbb4e06

[a] Value rounded to zero

[b] These effects were simulated as standardized mean differences