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

Anton Olsson-Collentine, Jelte Wicherts, & Marcel A.L.M. van Assen

"hidden moderators"

"minor, seemingly arbitrary and even theoretically irrelevant modifications in procedures..."

"psychological phenomena are not stable across time, situations and persons, therefore being inherently non-reproducible"

Conclusion

Minor and theoretically irrelevant changes in sample population and settings are unlikely to affect research outcomes in psychology

Conclusion

What does this mean?

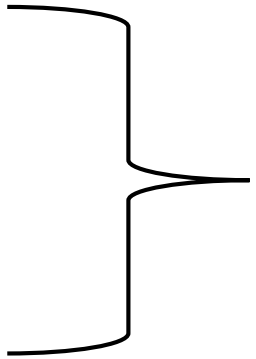
Post hoc hypothesizing about sensitivity to changes in sample population and settings (heterogeneity) is not a credible explanation for 'failed' direct replications

Heterogeneity (contextual sensitivity)

- Heterogeneity = sensitivity to changes in contextual factors

- Difference between two studies examining the same phenomenon? (Generalizability theory)

1) sample population
2) settings
3) treatment variables
4) measurement variables



Contextual factors

Conclusion

What does this mean?

Post hoc hypothesizing about sensitivity to changes in sample population and settings (heterogeneity) is not a credible explanation for 'failed' direct replications

Why not?

Research question

To what extent do study results in psychology depend on sample population and settings?

Data

10 Pre-registered multi-lab replication projects (Many labs 1&3; RRR 1 – 8)

- Median 23 labs per project
- Median 102 participants per lab
- 37 primary effects (= 37 meta-analyses)

Analyses

1) Observed heterogeneity (sensitivity to changes in contextual factors) across effects
2) Estimated power to detect small/medium/large heterogeneity (Higgins, 2003)

Table 3.

Heterogeneity across primary effects and statistical power of ten multi-lab replication projects, ordered with respect to estimated heterogeneity

						Statistical power			
						Level of heterogeneity			
RP	Effect	k	Effect type	Effect size estimate	I^2 (%)	I^2 95% CI	Zero	Small	Medium	Large
ML1	Anchoring 3 – Everest	36	SMD	2.41	91.29	[86.61, 95.23]	0.04	0.46	0.91	1.00
ML1	Allowed vs. forbidden	36	SMD	1.93	75.56	[60.32, 85.46]	0.05[a]	0.47[a]	0.91[a]	1.00[a]
ML1	Anchoring 2 – Chicago	36	SMD	2.00	75.36	[61.11, 87.15]	0.05	0.44	0.92	1.00
ML1	Anchoring 4 – Babies	36	SMD	2.53	64.67	[45.67, 83.33]	0.05	0.47	0.92	1.00
ML1	Quote Attribution	36	SMD	0.31	52.05	[24.63, 76.25]	0.04	0.43	0.91	1.00
ML1	Anchoring 1 – NYC	36	SMD	1.21	40.23	[10.62, 73.94]	0.05	0.45	0.92	1.00
ML1	IAT correlation math	35	R	0.39	40.05	[3.93, 64.97]	0.05	0.40	0.91	1.00
RRR3	Grammar on intentionality	12	MD	-0.25	38.06	[0.00, 85.72]	0.06	0.22	0.68	0.97
ML3	Subjective Distance interaction	21	R	0.02	33.51	[0.00, 76.78]	0.05	0.33	0.83	0.99
ML1	Gender math attitude	35	SMD	0.57	28.06	[0.00, 67.34]	0.05	0.44	0.90	1.00
ML3	Credentials interaction	21	R	0.02	24.03	[0.00, 73.82]	0.05	0.30	0.81	1.00
ML1	Gambler's Fallacy	36	SMD	0.61	22.85	[0.00, 69.16]	0.05	0.44	0.91	1.00
ML1	Imagined Contact	36	SMD	0.12	20.60	[0.00, 62.50]	0.05	0.44	0.91	1.00
ML1	Low vs. high category scales	36	SMD	0.88	19.20	[0.00, 49.95]	0.04	0.46	0.92	1.00
RRR8	Professor priming	23	MD	0.14	17.32	[0.00, 64.77]	0.05	0.34	0.83	1.00
ML1	Norm of reciprocity	36	SMD	-0.36	17.21	[0.00, 47.51]	0.05	0.43	0.91	1.00
ML3	Metaphor	20	R	0.14	13.03	[0.00, 57.02]	0.05	0.32	0.80	0.99
RRR1	Verbal overshadowing 1	32	RD	-0.03	12.23	[0.00, 46.51]	0.06[a]	0.38[a]	0.90[a]	1.00[a]
ML1	Sunk Costs	36	SMD	0.29	9.18	[0.00, 45.93]	0.05	0.44	0.91	1.00
RRR7	Intuitive-cooperation	21	MD	-0.39	2.80	[0.00, 39.28]	0.05	0.32	0.83	1.00
ML3	Availability	21	R	0.04	0.51	[0.00, 56.09]	0.05	0.34	0.83	1.00

Table 3 continued.

RP	Effect	k	Effect type	Effect size estimate	I^2 (%)	I^2 95% CI	Zero	Small	Medium	Large
ML1	Gain vs. loss framing	36	SMD	-0.66	0.01	[0.00, 55.57]	0.05[a]	0.43[a]	0.91[a]	1.00[a]
ML3	Power and Perspective	21	SMD	0.03	0.01	[0.00, 57.17]	0.05	0.32	0.81	0.99
RRR3	Grammar on intention attribution	12	MD	0.00	0.00	[0.00, 70.62]	0.06	0.24	0.70	0.96
ML3	Conscientiousness and persistence	21	R	0.02	0.00	[0.00, 61.42]	0.05	0.29	0.79	1.00
RRR3	Grammar on detailed processing	12	MD	-0.10	0.00	[0.00, 54.49]	0.06	0.24	0.70	0.97
RRR5	Commitment on neglect	16	MD	-0.05	0.00	[0.00, 53.18]	0.06	0.28	0.74	0.99
ML3	Warmth Perceptions	21	SMD	0.01	0.00	[0.00, 47.10]	0.04	0.37	0.91	1.00
RRR4	Ego depletion	23	SMD	0.00	0.00	[0.00, 46.91]	0.05	0.32	0.85	1.00
ML1	Flag Priming	36	SMD	0.02	0.00	[0.00, 36.23]	0.05	0.43	0.90	1.00
ML1	Money Priming	36	SMD	-0.02	0.00	[0.00, 33.18]	0.05	0.44	0.91	1.00
RRR2	Verbal overshadowing 2	23	RD	-0.15	0.00	[0.00, 32.36]	0.06[a]	0.31[a]	0.83[a]	1.00[a]
ML3	Weight Embodiment	20	SMD	0.03	0.00	[0.00, 29.97]	0.05	0.35	0.84	1.00
RRR6	Facial Feedback hypothesis	17	MD	0.03	0.00	[0.00, 25.13]	0.06	0.27	0.77	0.99
ML3	Elaboration likelihood interaction	20	R	0.00	0.00	[0.00, 18.62]	0.05	0.31	0.83	0.99
RRR5	Commitment on exit	16	MD	-0.06	0.00	[0.00, 17.44]	0.06	0.27	0.77	0.99
ML3	Stroop effect	21	R	0.41	0.00	[0.00, 13.61]	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, Estimate = Point estimates of effect sizes, I^2 95% CI = I^2 95% confidence interval. 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/

[a] Odds ratio or risk difference simulated as (standardized) mean difference

Table 3.

Heterogeneity across primary effects and statistical power of ten multi-lab replication projects, ordered with respect to estimated heterogeneity

						Statistical power			
						Level of heterogeneity			
RP	Effect	k	Effect type	Effect size estimate	I^2 (%)	I^2 95% CI	Zero	Small	Medium	Large
ML1	Anchoring 3 – Everest	36	SMD	2.41	91.29	[86.61, 95.23]	0.04	0.46	0.91	1.00
ML1	Allowed vs. forbidden	36	SMD	1.93	75.56	[60.32, 85.46]	0.05[a]	0.47[a]	0.91[a]	1.00[a]
ML1	Anchoring 2 – Chicago	36	SMD	2.00	75.36	[61.11, 87.15]	0.05	0.44	0.92	1.00
ML1	Anchoring 4 – Babies	36	SMD	2.53	64.67	[45.67, 83.33]	0.05	0.47	0.92	1.00
ML1	Quote Attribution	36	SMD	0.31	52.05	[24.63, 76.25]	0.04	0.43	0.91	1.00
ML1	Anchoring 1 – NYC	36	SMD	1.21	40.23	[10.62, 73.94]	0.05	0.45	0.92	1.00
ML1	IAT correlation math	35	R	0.39	40.05	[3.93, 64.97]	0.05	0.40	0.91	1.00
RRR3	Grammar on intentionality	12	MD	-0.25	38.06	[0.00, 85.72]	0.06	0.22	0.68	0.97
ML3	Subjective Distance interaction	21	R	0.02	33.51	[0.00, 76.78]	0.05	0.33	0.83	0.99
ML1	Gender math attitude	35	SMD	0.57	28.06	[0.00, 67.34]	0.05	0.44	0.90	1.00
ML3	Credentials interaction	21	R	0.02	24.03	[0.00, 73.82]	0.05	0.30	0.81	1.00
ML1	Gambler's Fallacy	36	SMD	0.61	22.85	[0.00, 69.16]	0.05	0.44	0.91	1.00
ML1	Imagined Contact	36	SMD	0.12	20.60	[0.00, 62.50]	0.05	0.44	0.91	1.00
ML1	Low vs. high category scales	36	SMD	0.88	19.20	[0.00, 49.95]	0.04	0.46	0.92	1.00
RRR8	Professor priming	23	MD	0.14	17.32	[0.00, 64.77]	0.05	0.34	0.83	1.00
ML1	Norm of reciprocity	36	SMD	-0.36	17.21	[0.00, 47.51]	0.05	0.43	0.91	1.00
ML3	Metaphor	20	R	0.14	13.03	[0.00, 57.02]	0.05	0.32	0.80	0.99
RRR1	Verbal overshadowing 1	32	RD	-0.03	12.23	[0.00, 46.51]	0.06[a]	0.38[a]	0.90[a]	1.00[a]
ML1	Sunk Costs	36	SMD	0.29	9.18	[0.00, 45.93]	0.05	0.44	0.91	1.00
RRR7	Intuitive-cooperation	21	MD	-0.39	2.80	[0.00, 39.28]	0.05	0.32	0.83	1.00
ML3	Availability	21	R	0.04	0.51	[0.00, 56.09]	0.05	0.34	0.83	1.00

Table 3 continued.

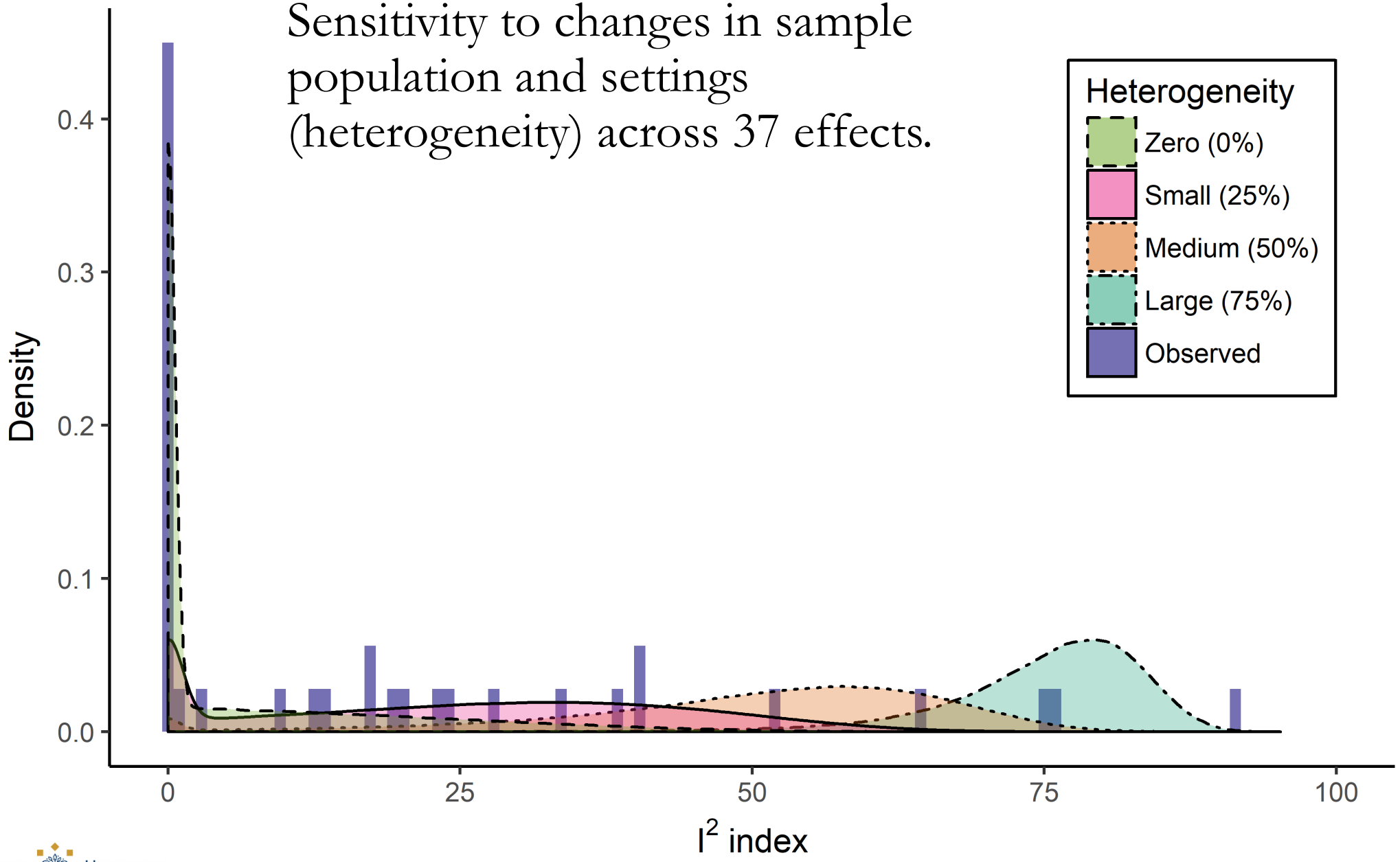
RP	Effect	k	Effect type	Effect size estimate	I^2 (%)	I^2 95% CI	Zero	Small	Medium	Large
ML1	Gain vs. loss framing	36	SMD	-0.66	0.01	[0.00, 55.57]	0.05[a]	0.43[a]	0.91[a]	1.00[a]
ML3	Power and Perspective	21	SMD	0.03	0.01	[0.00, 57.17]	0.05	0.32	0.81	0.99
RRR3	Grammar on intention attribution	12	MD	0.00	0.00	[0.00, 70.62]	0.06	0.24	0.70	0.96
ML3	Conscientiousness and persistence	21	R	0.02	0.00	[0.00, 61.42]	0.05	0.29	0.79	1.00
RRR3	Grammar on detailed processing	12	MD	-0.10	0.00	[0.00, 54.49]	0.06	0.24	0.70	0.97
RRR5	Commitment on neglect	16	MD	-0.05	0.00	[0.00, 53.18]	0.06	0.28	0.74	0.99
ML3	Warmth Perceptions	21	SMD	0.01	0.00	[0.00, 47.10]	0.04	0.37	0.91	1.00
RRR4	Ego depletion	23	SMD	0.00	0.00	[0.00, 46.91]	0.05	0.32	0.85	1.00
ML1	Flag Priming	36	SMD	0.02	0.00	[0.00, 36.23]	0.05	0.43	0.90	1.00
ML1	Money Priming	36	SMD	-0.02	0.00	[0.00, 33.18]	0.05	0.44	0.91	1.00
RRR2	Verbal overshadowing 2	23	RD	-0.15	0.00	[0.00, 32.36]	0.06[a]	0.31[a]	0.83[a]	1.00[a]
ML3	Weight Embodiment	20	SMD	0.03	0.00	[0.00, 29.97]	0.05	0.35	0.84	1.00
RRR6	Facial Feedback hypothesis	17	MD	0.03	0.00	[0.00, 25.13]	0.06	0.27	0.77	0.99
ML3	Elaboration likelihood interaction	20	R	0.00	0.00	[0.00, 18.62]	0.05	0.31	0.83	0.99
RRR5	Commitment on exit	16	MD	-0.06	0.00	[0.00, 17.44]	0.06	0.27	0.77	0.99
ML3	Stroop effect	21	R	0.41	0.00	[0.00, 13.61]	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, Estimate = Point estimates of effect sizes, I^2 95% CI = I^2 95% confidence interval. 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/

[a] Odds ratio or risk difference simulated as (standardized) mean difference

30/37 (81%) of CI include zero

Sensitivity to changes in sample population and settings (heterogeneity) across 37 effects.



Summary

Reasons to believe zero to small heterogeneity is the norm

- Only 7/37 (19%) show significant heterogeneity
- Under zero heterogeneity expect 17/37 (46%) non-zero estimates, actual 25/37 (68%)

Note also

- Low power to distinguish between zero/small heterogeneity
- Good power to detect medium/large heterogeneity (avg. 85/99+%)

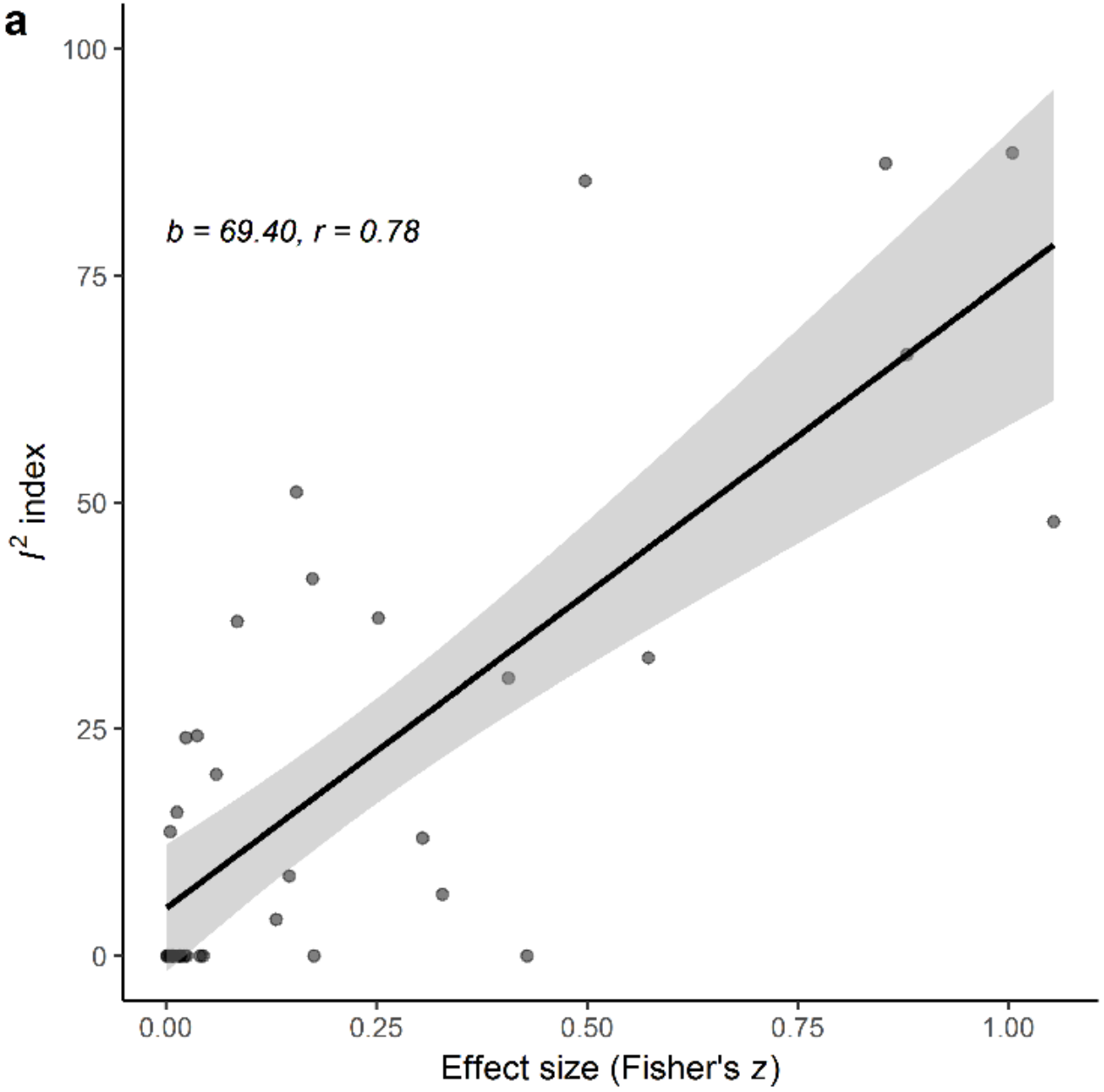
Conclusion

What does this mean?

Post hoc hypothesizing about sensitivity to changes in sample population and settings (heterogeneity) is not a credible explanation for 'failed' direct replications

Is *a priori* unlikely

Heterogeneity strongly correlated with effect size ($r = .70$ - $.78$)



Caveats

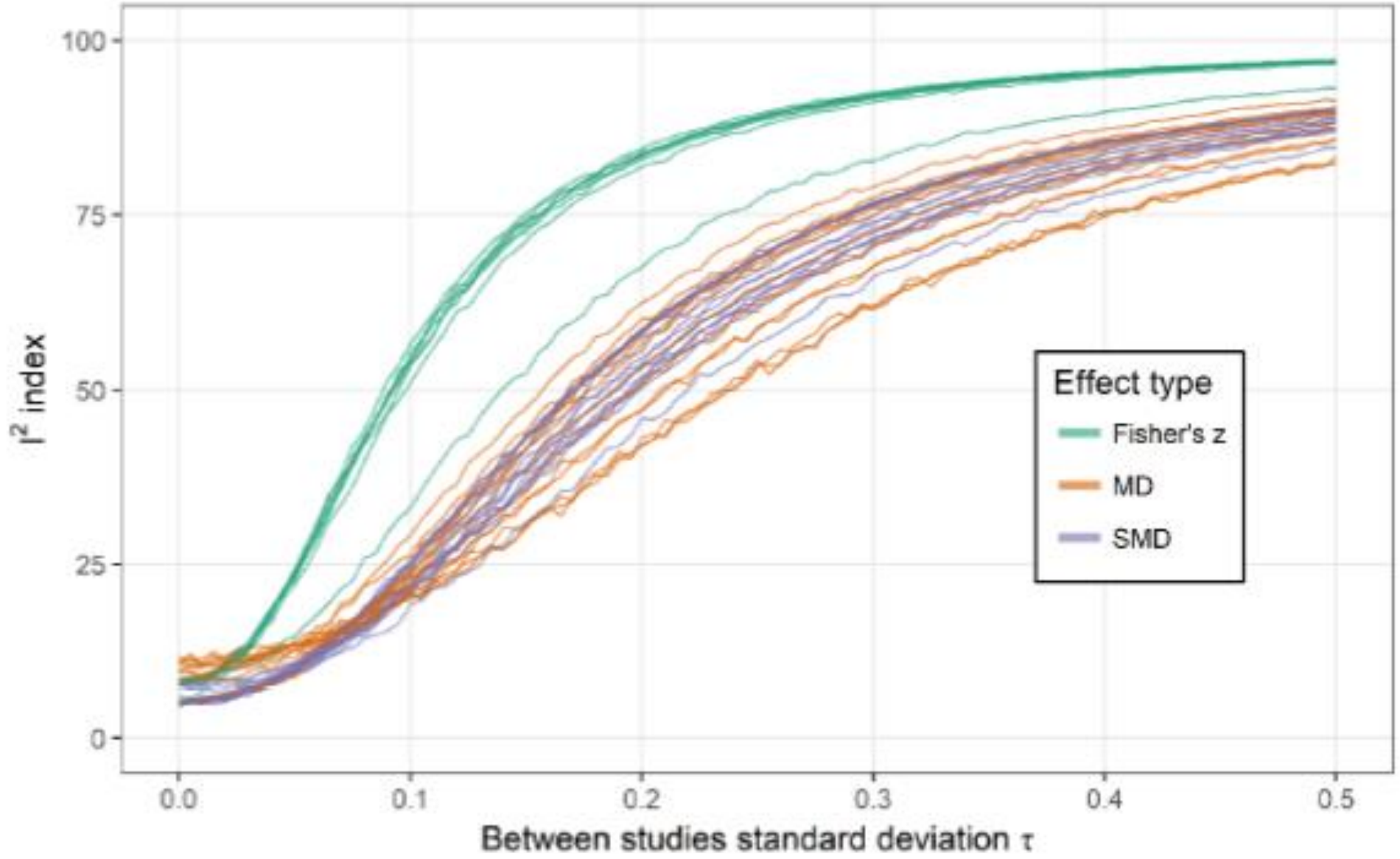
- Small sample (37 psychological effects)
  - See also ML2 (+28 effects)

- Only varied sample population and settings, not treatment/measurement variables
  - Effects may be sensitive to more extensive changes to contextual factors (van Erp et al. 2017; Stanley et al. 2017)

Conclusion

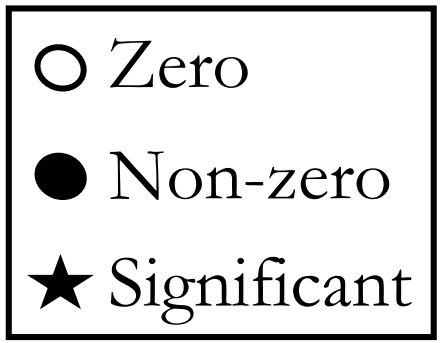
Minor and theoretically irrelevant changes in sample population and settings are unlikely to affect research outcomes in psychology

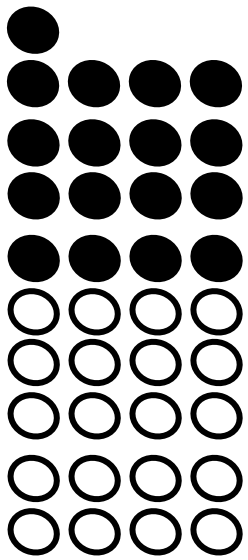
Thanks!

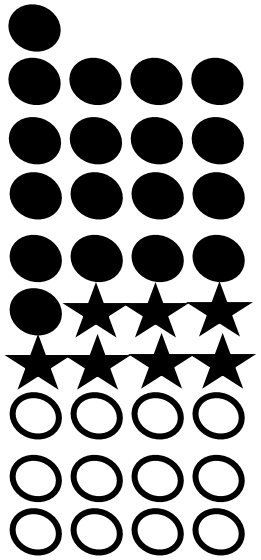

100
75
50
25
0
I^2 index
0.0
0.1
0.2
0.3
0.4
0.5
Between studies standard deviation τ
Effect type
Fisher's z
MD
SMD

Expected
17/37 (46%)

Observed
25/37 (68%)



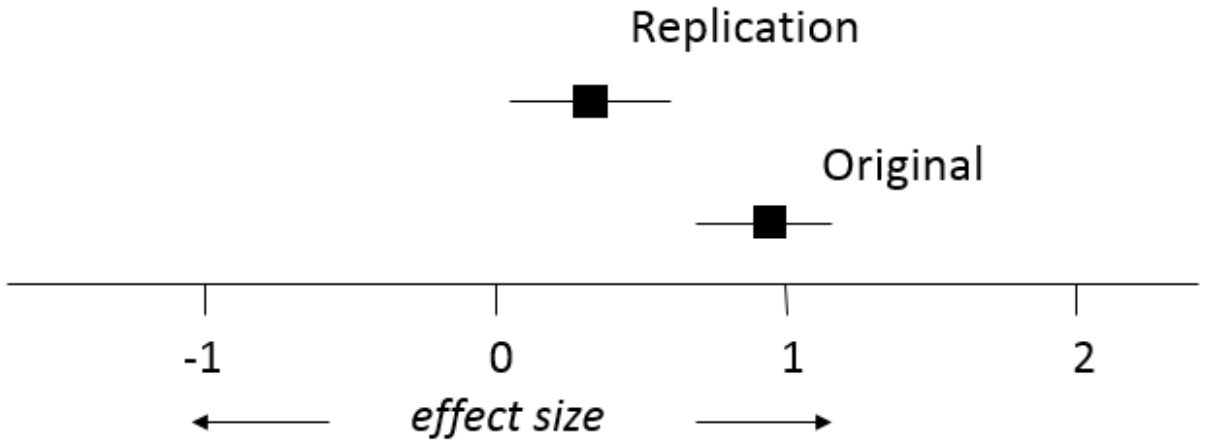




Heterogeneity (contextual sensitivity)

- Heterogeneity = sensitivity to changes in contextual factors

- Original study vs. replication



Why different results?

- Sampling variance
- Garden of forking paths (*p*-hacking)
- Heterogeneity