Supplement

1 Heterogeneity analysis

The prediction interval and p-original values used for models and point estimates were done on the original scale where possible. However, to do heterogeneity analysis, our imputed value for tau is in SMD units, so we have to use SMD units for the original and replication values. This limits the number of applicable original-replication pairs. Prediction intervals are sensitive to whether only studies with SMD are included (different studies report differently and also vary in how close replications were) and whether the scale used was original or SMD (if the variances differed between original and replication).

Of all studies where prediction intervals can be calculated, 62/136 (46%) of the original estimates had prediction intervals that included the replication effect size. For those where we have SMD, 43/112 (38%) of the original estimates had prediction intervals that included the replication effect size, when we use original units where possible. Using SMD units, 33/112 (29%) of the original estimates had prediction intervals that included the replication effect size. Allowing for heterogeneity, with tau=.21, 72/112 (64%) of the original estimates are distributionally consistent.

2 Additional model results

Per our pre-registration, we ran a set of 6 models crossing 3 outcome measures (subjective replication score, whether the replication result was within the prediction interval of the original, and p-original on the hypothesis that both came from the same distribution) with 2 sets of predictors (with or without statistical predictors). These 6 models required 3 tiers of data: the subjective replication score without statistical predictors applies to all the data; the p_original and prediction interval models apply to the subset of data with numeric outcomes that can be compared; and the statistical predictor models need the smaller subset of data with p-values and original standardized effect size in particular.

Due to low sample sizes and large numbers of predictors, even with regularizing priors, the coefficient estimates generally have a lot of uncertainty.

2.1 Sensitivity Analysis

As a check on whether our results were sensitive to the inclusion of pairs that were marginal in some way, we repeated the 6 models including only studies that were not marginal. For sensitivity analyses, we completely excluded some studies where the replication analysis was not the focal analysis of the original (ex. student chose a manipulation check as their primary analysis). We also excluded studies with extremely small or lop-sided sample sizes.

For some studies that were included with some statistical outcomes in the main results, we downgraded them to experimental features only for sensitivity. Here, we felt the replication itself didn't have major problems and we felt confident in the subjective replication score reflecting holistically on the replication. However, there were some mismatches in the statistical reporting that meant we were not confident enough if your numeric calculations for the paper. Common mismatches were where one reported a t-test and the other reported mean and SDs, and when degrees of freedom seemed inconsistent across replication and original (suggesting different groupings or levels of analysis).

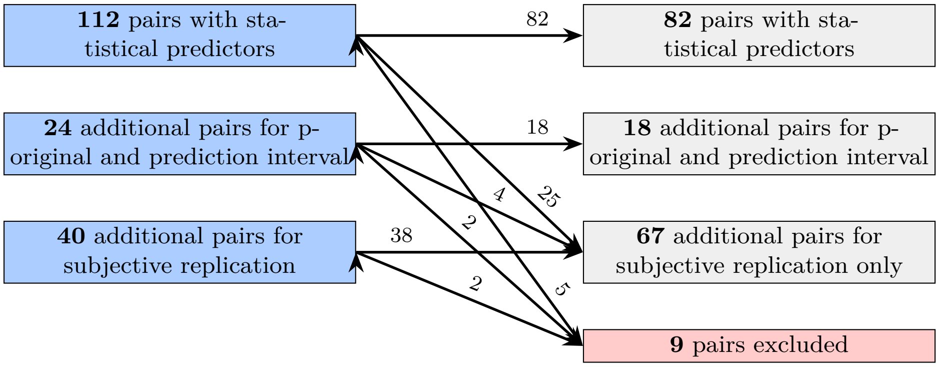
112 pairs with statistical predictors
24 additional pairs for p-original and prediction interval
40 additional pairs for subjective replication
82 pairs with statistical predictors
18 additional pairs for p-original and prediction interval
67 additional pairs for subjective replication only
9 pairs excluded
82
18
38
25
4
2
2
5

Figure 1: Diagram of what studies were downgraded or excluded for the sensitivity analysis.

Table 1: Odds ratios and 95% CrI from ordinal models predicting subjective replication scores. All pairs indicates that all pairs used, regardless of whether they had full statistical information; Full stats indicates that only pairs with full statistical information was used. Sensitivity indicates that it was a sensitivity test.

	All pairs		Full stats		All pairs		Full stats	
					Sensitivity		Sensitivity	
	N=176		N=112		N=167		N=82	
Original effect size	–	–	4.36	[1.31, 16.65]	–	–	13.93	[2.15, 131.6]
Within participants	2.56	[1, 7.36]	2.76	[0.97, 10.71]	2.08	[0.95, 6.03]	4.45	[0.99, 28.08]
Open data	1.33	[0.8, 3.28]	2.11	[0.84, 8.42]	1.36	[0.8, 3.83]	2.05	[0.74, 8.86]
Non psych	1.14	[0.67, 2.63]	1.44	[0.58, 5.64]	1.09	[0.6, 2.33]	1.5	[0.45, 8.23]
Publication year	1.06	[0.78, 1.58]	1.25	[0.66, 3.19]	1.03	[0.76, 1.5]	1.26	[0.54, 3.85]
Other psych	1.04	[0.58, 2.03]	1.11	[0.5, 2.81]	1.03	[0.55, 2]	1.36	[0.45, 5.73]
P-value	–	–	1.05	[0.62, 1.92]	–	–	1.62	[0.71, 5.11]
Stanford	1.04	[0.55, 2.16]	0.94	[0.36, 2.28]	1.05	[0.54, 2.31]	1.05	[0.28, 4.38]
Log original sample size	0.95	[0.63, 1.31]	0.94	[0.44, 1.72]	0.93	[0.6, 1.27]	0.79	[0.24, 1.81]
Log rep/orig sample	0.91	[0.64, 1.2]	0.86	[0.49, 1.32]	0.86	[0.56, 1.14]	0.68	[0.25, 1.28]
Log trials	0.92	[0.55, 1.3]	0.85	[0.38, 1.44]	0.93	[0.53, 1.33]	0.55	[0.16, 1.23]
Social psych	0.74	[0.31, 1.24]	0.83	[0.34, 1.66]	0.74	[0.3, 1.24]	0.54	[0.15, 1.42]
Single vignette	0.75	[0.27, 1.31]	0.76	[0.21, 1.72]	0.69	[0.23, 1.23]	0.82	[0.19, 2.74]
Open materials	0.85	[0.4, 1.35]	0.68	[0.22, 1.37]	0.79	[0.34, 1.28]	0.76	[0.21, 1.98]
Switch to online	0.63	[0.26, 1.1]	0.49	[0.14, 1.17]	0.73	[0.32, 1.16]	0.63	[0.17, 1.59]

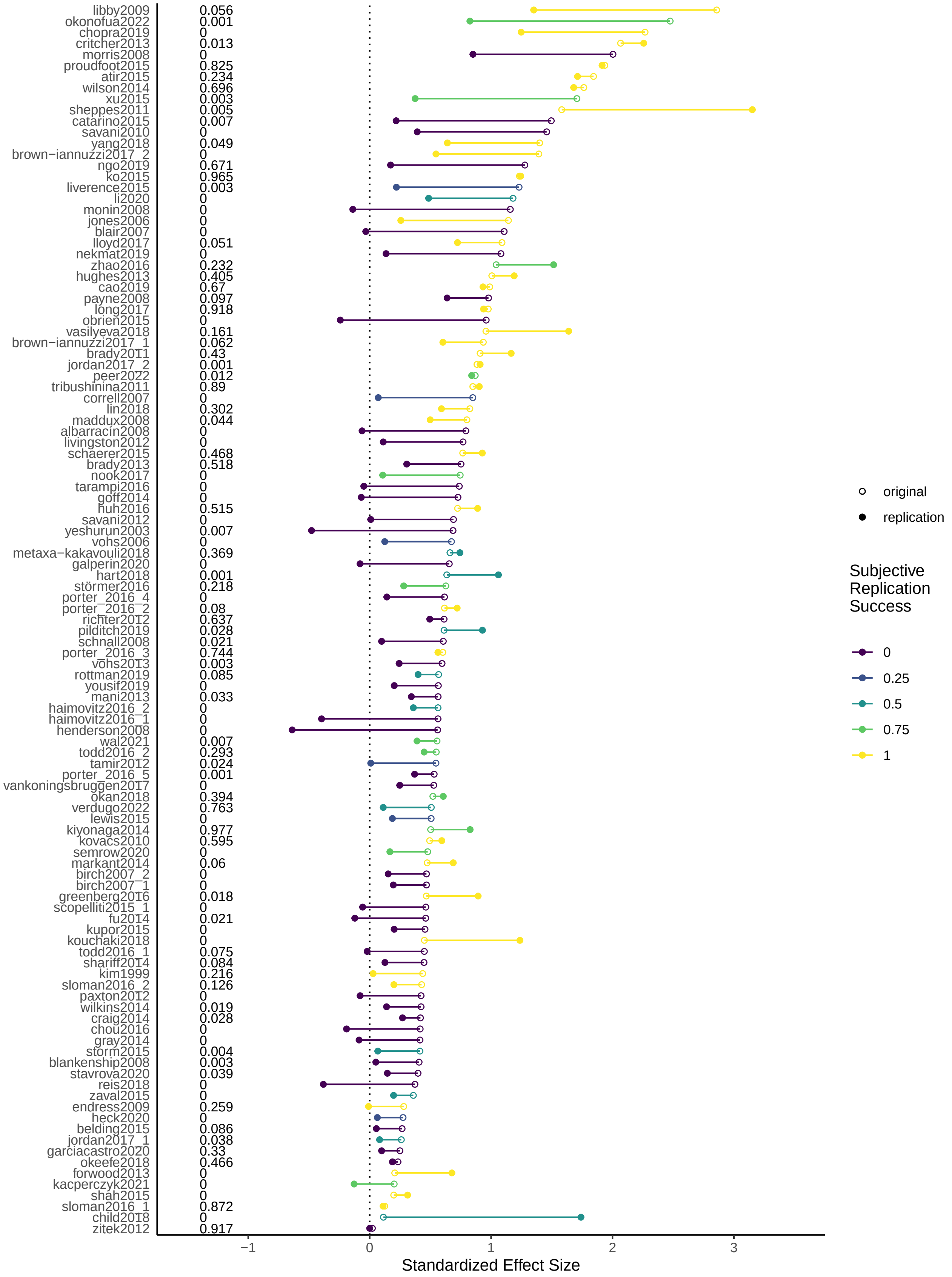

Figure 2: Forest plot of original and replication effect sizes (N=112). Original effect sizes are open dots, replication are closed dots. Coloring indicates subjective replication score, and p-original values are listed on the left side. The median original effect size was .61 and the median replication effect size was .28.

Table 2: Odds ratios and 95% CrI from logistic models predicting whether the prediction interval of the original effect contained the replication effect. All pairs indicates that all pairs with calculable prediction intervals were used, regardless of whether they had full statistical information; Full stats indicates that only data with full statistical information was used. Sensitivity indicates that it was a sensitivity test.

	All pairs		Full stats		All pairs		Full stats	
					Sensitivity		Sensitivity	
	N=136		N=112		N=100		N=82	
Within participants	1.59	[0.86, 5.48]	1.79	[0.83, 7.44]	2.5	[0.93, 13.04]	2.46	[0.89, 15.34]
Publication year	1.11	[0.78, 2.07]	1.37	[0.81, 3.89]	1.01	[0.6, 1.65]	1.29	[0.7, 4.04]
Log trials	1.25	[0.85, 2.49]	1.31	[0.81, 2.95]	1.15	[0.75, 2.37]	1.25	[0.76, 2.92]
P-value	–	–	1.26	[0.77, 3.31]	–	–	1.1	[0.61, 2.75]
Original effect size	–	–	1.17	[0.75, 2.57]	–	–	1.29	[0.74, 4.22]
Social psych	0.98	[0.57, 1.57]	1.11	[0.62, 2.65]	0.93	[0.45, 1.56]	1.04	[0.53, 2.39]
Open data	0.96	[0.52, 1.57]	1.11	[0.59, 2.77]	0.93	[0.41, 1.66]	1.01	[0.47, 2.25]
Open materials	1.16	[0.75, 2.69]	1.11	[0.62, 2.66]	1.16	[0.69, 2.98]	1.17	[0.62, 3.19]
Single vignette	0.87	[0.35, 1.47]	0.99	[0.43, 2.31]	0.91	[0.37, 1.76]	1.04	[0.44, 2.99]
Log rep/orig sample	0.98	[0.66, 1.37]	0.96	[0.57, 1.46]	0.98	[0.63, 1.44]	0.97	[0.52, 1.64]
Stanford	0.89	[0.35, 1.49]	0.95	[0.39, 2.02]	0.85	[0.25, 1.55]	0.78	[0.15, 1.71]
Non psych	1.02	[0.57, 2.01]	0.93	[0.33, 2.13]	1	[0.47, 2.01]	0.83	[0.21, 1.85]
Other psych	0.85	[0.3, 1.41]	0.86	[0.3, 1.74]	0.92	[0.35, 1.74]	0.94	[0.35, 2.02]
Switch to online	0.95	[0.5, 1.47]	0.85	[0.32, 1.51]	0.92	[0.43, 1.61]	0.84	[0.29, 1.56]
Log original sample size	0.93	[0.53, 1.35]	0.67	[0.24, 1.16]	1	[0.62, 1.69]	0.77	[0.28, 1.33]

Table 3: Coefficients and 95% CrI from linear models predictiong the p-original value between the original and replicaiton effects. All pairs indicates that all pairs with calculable prediction intervals were used, regardless of whether they had full statistical information; Full stats indicates that only data with full statistical information was used. Sensitivity indicates that it was a sensitivity test.

	All pairs		Full stats		All pairs		Full stats	
					Sensitivity		Sensitivity	
	N=136		N=112		N=100		N=82	
Within participants	1.59	[0.86, 5.48]	1.79	[0.83, 7.44]	2.5	[0.93, 13.04]	2.46	[0.89, 15.34]
Publication year	1.11	[0.78, 2.07]	1.37	[0.81, 3.89]	1.01	[0.6, 1.65]	1.29	[0.7, 4.04]
Log trials	1.25	[0.85, 2.49]	1.31	[0.81, 2.95]	1.15	[0.75, 2.37]	1.25	[0.76, 2.92]
P-value	–	–	1.26	[0.77, 3.31]	–	–	1.1	[0.61, 2.75]
Original effect size	–	–	1.17	[0.75, 2.57]	–	–	1.29	[0.74, 4.22]
Social psych	0.98	[0.57, 1.57]	1.11	[0.62, 2.65]	0.93	[0.45, 1.56]	1.04	[0.53, 2.39]
Open data	0.96	[0.52, 1.57]	1.11	[0.59, 2.77]	0.93	[0.41, 1.66]	1.01	[0.47, 2.25]
Open materials	1.16	[0.75, 2.69]	1.11	[0.62, 2.66]	1.16	[0.69, 2.98]	1.17	[0.62, 3.19]
Single vignette	0.87	[0.35, 1.47]	0.99	[0.43, 2.31]	0.91	[0.37, 1.76]	1.04	[0.44, 2.99]
Log rep/orig sample	0.98	[0.66, 1.37]	0.96	[0.57, 1.46]	0.98	[0.63, 1.44]	0.97	[0.52, 1.64]
Stanford	0.89	[0.35, 1.49]	0.95	[0.39, 2.02]	0.85	[0.25, 1.55]	0.78	[0.15, 1.71]
Non psych	1.02	[0.57, 2.01]	0.93	[0.33, 2.13]	1	[0.47, 2.01]	0.83	[0.21, 1.85]
Other psych	0.85	[0.3, 1.41]	0.86	[0.3, 1.74]	0.92	[0.35, 1.74]	0.94	[0.35, 2.02]
Switch to online	0.95	[0.5, 1.47]	0.85	[0.32, 1.51]	0.92	[0.43, 1.61]	0.84	[0.29, 1.56]
Log original sample size	0.93	[0.53, 1.35]	0.67	[0.24, 1.16]	1	[0.62, 1.69]	0.77	[0.28, 1.33]