Table 3. Summary of Effect Sizes Across Studies

	Original study		RP:P replication		ML5: RP:P protocol		ML5: revised protocol	
Original study	*N*	*r*	*N*	*r*	*N*	*r*	*N*	*r*
Albarracín et al. (2008), Experiment 5	36	.38 [.05, .64]	88	−.03 [−.24, .18]	580	.04 [−.04, .12]	884	.09 [.03, .14]
Albarracín et al. (2008), Experiment 7	98	.21 [.01, .39]	105	.16 [−.03, .34]	878	.01 [−.19, .21]	808	−.07 [−.17, .03]
Crosby, Monin, & Richardson (2008)	25	.25 [.02, .46]	30	.18 [−.03, .40]	140	.15 [−.01, .30]	136	.14 [−.08, .34]
Förster, Liberman, & Kuschel (2008)	82	.43 [.23, .59]	71	.11 [−13, .34]	736	.03 [−.02, .09]	720	.05 [−.07, .16]
LoBue & DeLoache (2008)	48	.48 [.22, .70]	48	.18 [−.10, .46]	286	.01 [−.19, .21]	259	.04 [−.02, .10]
Payne, Burkley, & Stokes (2008)	70	.35 [.12, .54]	180	.15 [.00, .29]	545	.05 [−.13, .22]	558	−.16 [−.44, .15]
Risen & Gilovich (2008)	122	.19 [.01, .36]	226	.00 [−.13, .13]	2,811	−.04 [−.08, −.01]	701	−.01 [−.13, .11]
Shnabel & Nadler (2008)	94	.27 [.07, .45]	141	−.10 [−.27, .07]	1,361	.02 [−.03, .08]	1,376	.09 [.04, .14]
van Dijk, van Kleef, Steinel, & van Beest (2008)	103	.38 [.20, .54]	83	−.04 [−.26, .18]	436	.06 [−.06, .18]	119	.23 [−.01, .44]
Vohs & Schooler (2008)	30	.50 [.15, .74]	58	.10 [−.17, .35]	279	.04 [−14, .22]	342	.05 [−.16, .25]

Note: Values in brackets are 95% confidence intervals. RP:P = Reproducibility Project: Psychology; ML5 = Many Labs 5.

For the second test, we conducted a random-effects meta-analysis on the estimates of the effect of protocol within each replication study. We calculated the strength of the effect of protocol on the Pearson's *r* scale for each of the 10 studies. A meta-analysis of these 10 estimates suggested that these effect sizes were not reliably different from zero, $b = 0.014$, 95% CI = [−.02, .05], $SE = 0.01$, $t = 0.968$, $p = .335$. Across studies, the point estimates for revised protocols were thus, on average, 0.014 units larger than the point estimate for RP:P protocols on the Pearson's *r* scale. Overall, the effect of protocol within each study had a fairly small amount of heterogeneity, $\tau = .034$ (95% CI = [0, .06]) on the Fisher's *z* scale. However, the *Q* statistic suggested significant heterogeneity, $Q = 21.81$, $p = .010$, $I^2 = 60.89\%$. Collapsing the data across protocols, we found that only one of the individual studies (Ebersole et al.'s, 2020, replication of Payne, Burkley, & Stokes, 2008) showed at least a small amount of heterogeneity, as indicated by a τ value greater than .10 ($\tau = .16$ for this study)

Exploratory analyses: other evaluations of replicability

Both of our primary tests of the effect of formal peer review on increasing effect sizes of replications failed to reject the null hypothesis and yielded very weak effect sizes with narrow confidence intervals. Nevertheless, two of the revised protocols showed effects below the $p < .05$ threshold (p values of .009 and .005), whereas none of the RP:P protocols did so. Although this pattern might appear to support the hypothesis that expert peer review could improve replicability, counting the number of "significant" replications is not a formal test (Mathur & VanderWeele, 2020). This pattern could have occurred by chance, and indeed, the formal statistical tests do not suggest that the difference is systematic. Perhaps formal peer review does not improve the replicability of findings more than trivially, but perhaps it did for these two studies? Of the two statistically significant effects obtained with the revised protocol, the observed effect sizes were 76% and 67% smaller than the those reported for the original studies. Comparing the RP:P and revised protocols for each of these findings indicated that for only one of the two tests was the revised protocol's effect size significantly larger ($p = .601$ for Chartier et al.'s, 2020, replication of Albarracín et al.'s, 2008, Experiment 5; $p = .012$ for Baranski et al.'s, 2020 replication of Shnabel & Nadler, 2008). It is possible that the expert feedback did reliably improve the effect size for the replication of Shnabel and Nadler (2008), but given the number of tests, it is also plausible that this difference occurred by chance. Therefore, even the most promising examples of formal peer review increasing replicability fail to provide reliable support.

We also examined the cumulative evidence for each of the 10 findings. Figure 2 shows the evidence from