Reproducibility of Experimental Research

Background

Power posing example

Reproducibility Project: Psychology (RPP)

Experimental Economics Replication Project (EERP)

Social Sciences Replication Project (SSRP)

Prediction Markets of Replicability

Background

Reproducibility

Replicability

Types of replications:

Reanalysis replication: verifying that the published results can be reproduced based on the same methods (and code if available) and data.

Direct replication: running the same experiment in the same way as the original experiment, i.e. ideally using exactly the same materials and software as in the original study.

Conceptual replication: testing the same hypothesis as in the original study, but using a different method/design.

Example of Replication: Power Posing

Original Study: Carney, Cuddy and Yap, Psychological Science 2010. Amy Cuddy Ted talk on power posing seen by about 50 million (second highest of all Ted talks).

42 subjects randomly allocated to two treatments (high-power-pose and low-power-pose treatment).

Saliva samples taken before and after power-posing to measure cortisol and testosterone. Measurement of risk taking (a choice between $2 for sure or a 50/50 chance of winning $4) and feelings of power (a scale from 1 to 4) after power-poising.

Reported significantly higher testosterone ($p=0.045$), significantly lower cortisol ($p=0.01$), significantly higher risk taking ($p=0.0495$), and significantly higher feelings of power (0.004) for the high-power-pose treatment compared to the low-power-pose treatment.

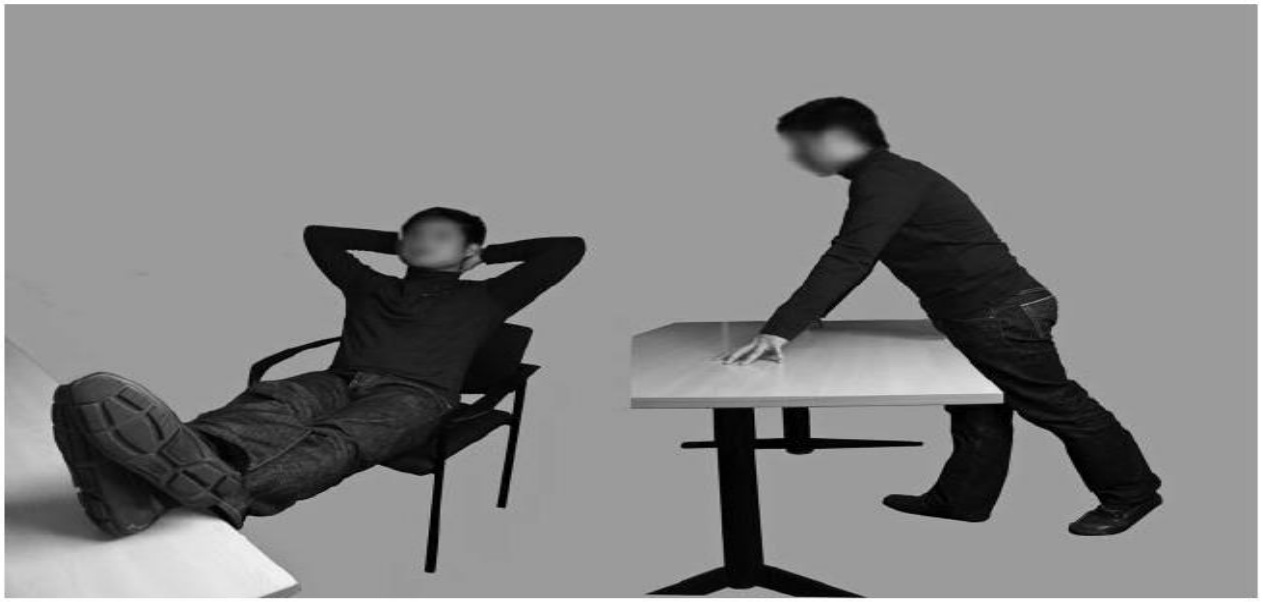


Fig. 1. The two high-power poses used in the study. Participants in the high-power-pose condition were posed in expansive positions with open limbs.

Fig. 2. The two low-power poses used in the study. Participants in the low-power-pose condition were posed in contractive positions with closed limbs.

Replication of Power Posing

Ranehill et al., Psychological Science 2015.

200 subjects (about five times larger than the original study)

No significant difference between high and low power pose treatment on testosterone ($p=0.162$ and point estimate in opposite direction of the original study), cortisol ($p=0.272$), or risk taking ($p=0.215$ and point estimate in the opposite direction of the original study). Found a significant effect of feelings of power ($p=0.017$) in the same direction as the original study (but the size of this effect was only about a third of the original effect size).

How to measure if a study replicates?

Alternative ”replication indicators” suggested in the literature:

A statistically significant effect in the same direction as the original study interpreted as a ”succesful replication” (with p=0.05 typically used as the significance threshold in a double-sided test).

The relative effect size of the replication (the effect size of the replication divided by the effect size of the original study). A continuous measure of the degree of replication success.

The replication 95% confidence interval overlapping the point estimate of the original study interpreted as a ”succesful replication”.

A statistically significant ($p<0.05$) difference between the replication and the original result interpreted as a ”failed replication” (the prediction interval approach suggested by Patil et al. (Psychological Science 2016) based on this)

An effect in the replication that is significantly smaller than a ”small effect” interpreted as a ”failed replication” ($p<0.05$ in a one-sided test). A ”small effect” defined as the effect size the original study would have had 33% power to detect. The ”Small Telescopes Approach” (Simonsohn, Psychological Science 2015). Recommended in this approach that replications should always have 2.5 times the sample size of the original study as that gives 80% power to reject a small effect size if the true effect is zero.

Replication Project: Psychology, RPP (Open Science Collaboration, Science 2015)

Replicated 100 experimental and correlational studies published in 2008 selected from three top journals in psychology (Psychological Science, Journal of Personality and Social Psychology, Journal of Experimental Psychology: Learning, Memory and Cognition).

The last study/experiment typically selected for replication for articles reporting a series of experiments. A key result from this experiment the focus of replication.

100 replications conducted; 97 of the original studies reported a statistically significant finding for the key result being replicated ($p<0.05$ although four studies had p-values between 0.05 and 0.06 and were interpreted as statistically significant in the original articles) and 3 of the original studies reported a null result for the key result being replicated.

Replication teams could join the project and conduct at least one replication. Original authors invited to give feedback on the replication design.

Replication power on average 92% to detect the same effect size as in the original study.

Table 1. Summary of reproducibility rates and effect sizes for original and replication studies overall and by journal/discipline. *df/N* refers to the information on which the test of the effect was based (for example, *df* of *t* test, denominator *df* of *F* test, sample size −3 of correlation, and sample size for *z* and χ^2). Four original results had *P* values slightly higher than 0.05 but were considered positive results in the original article and are treated that way here. Exclusions (explanation provided in supplementary materials, A3) are "replications $P < 0.05$" (3 original nulls excluded; $n = 97$ studies); "mean original and replication effect sizes" (3 excluded; $n = 97$ studies); "meta-analytic mean estimates" (27 excluded; $n = 73$ studies); "percent meta-analytic ($P < 0.05$)" (25 excluded; $n = 75$ studies); and, "percent original effect size within replication 95% CI" (5 excluded, $n = 95$ studies).

			Effect size comparison					Original and replication combined			
	Replications $P < 0.05$ in original direction	Percent	Mean (SD) original effect size	Median original *df/N*	Mean (SD) replication effect size	Median replication *df/N*	Average replication power	Meta-analytic mean (SD) estimate	Percent meta-analytic ($P < 0.05$)	Percent original effect size within replication 95% CI	Percent subjective "yes" to "Did it replicate?"
Overall	35/97	36	0.403 (0.188)	54	0.197 (0.257)	68	0.92	0.309 (0.223)	68	47	39
JPSP, social	7/31	23	0.29 (0.10)	73	0.07 (0.11)	120	0.91	0.138 (0.087)	43	34	25
JEP:LMC, cognitive	13/27	48	0.47 (0.18)	36.5	0.27 (0.24)	43	0.93	0.393 (0.209)	86	62	54
PSCI, social	7/24	29	0.39 (0.20)	76	0.21 (0.30)	122	0.92	0.286 (0.228)	58	40	32
PSCI, cognitive	8/15	53	0.53 (0.2)	23	0.29 (0.35)	21	0.94	0.464 (0.221)	92	60	53

Table 2. Spearman's rank-order correlations of reproducibility indicators with summary original and replication study characteristics. Effect size difference computed after converting *r* to Fisher's *z*. *df/N* refers to the information on which the test of the effect was based (for example, *df* of *t* test, denominator *df* of *F* test, sample size −3 of correlation, and sample size for *z* and χ^2). Four original results had *P* values slightly higher than 0.05 but were considered positive results in the original article and are treated that way here. Exclusions (explanation provided in supplementary materials, A3) are "replications $P < .05$" (3 original nulls excluded; $n = 97$ studies), "effect size difference" (3 excluded; $n = 97$ studies); "meta-analytic mean estimates" (27 excluded; $n = 73$ studies); and, "percent original effect size within replication 95% CI" (5 excluded, $n = 95$ studies).

	Replications $P < 0.05$ in original direction	Effect size difference	Meta-analytic estimate	Original effect size within replication 95% CI	Subjective "yes" to "Did it replicate?"
Original study characteristics					
Original *P* value	−0.327	−0.057	−0.468	0.032	−0.260
Original effect size	0.304	0.279	0.793	0.121	0.277
Original *df/N*	−0.150	−0.194	−0.502	−0.221	−0.185
Importance of original result	−0.105	0.038	−0.205	−0.133	−0.074
Surprising original result	−0.244	0.102	−0.181	−0.113	−0.241
Experience and expertise of original team	−0.072	−0.033	−0.059	−0.103	−0.044
Replication characteristics					
Replication *P* value	−0.828	0.621	−0.614	−0.562	−0.738
Replication effect size	0.731	−0.586	0.850	0.611	0.710
Replication power	0.368	−0.053	0.142	−0.056	0.285
Replication *df/N*	−0.085	−0.224	−0.692	−0.257	−0.164
Challenge of conducting replication	−0.219	0.085	−0.301	−0.109	−0.151
Experience and expertise of replication team	−0.096	0.133	0.017	−0.053	−0.068
Self-assessed quality of replication	−0.069	0.017	0.054	−0.088	−0.055

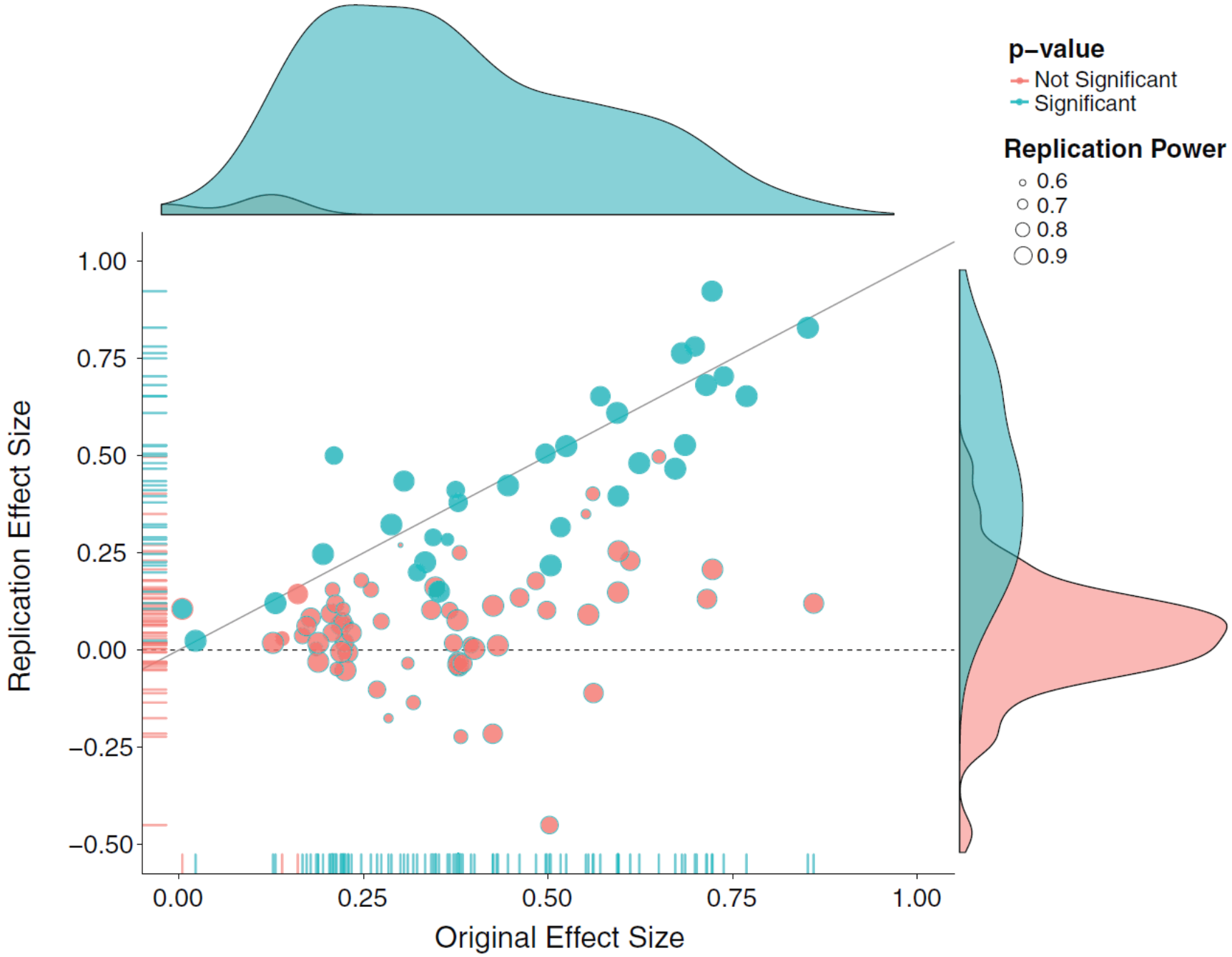


Fig. 3. Original study effect size versus replication effect size (correlation coefficients).

Experimental Economics Replication Project, EERP (Camerer et al. Science 2016)

Replication of 18 between-subject lab experimental papers published in American Economic Review or Quarterly Journal of Economics published in 2011-2014 (all papers meeting the inclusion criteria).

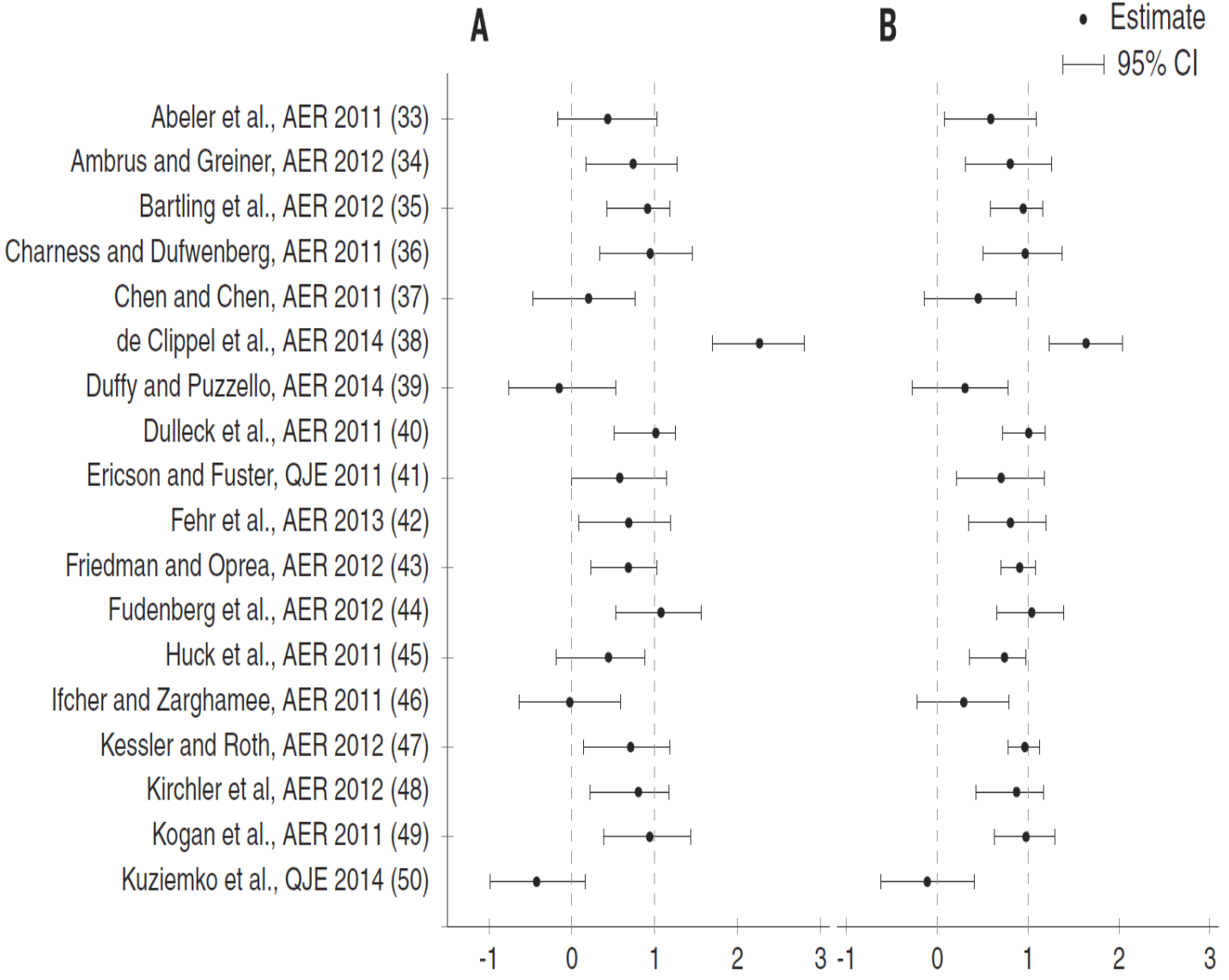
The most central statistically significant between-subject result as emphasized in the original paper selected for replication (if more than one equally central result the following criteria used to select the result: related to efficiency; the last experiment; lottery if ”ties” (n=5). Two original studies had p-values between 0.05 and 0.10, but interpreted as statistically significant by the original authors.

Replication power at least 90% to detect the same effect size as in the original study (average replication power 92%).

Four replication teams (CalTech, National University of Singapore, Stockholm School of Economics, University of Innsbruck).

Replication reports with design sent to the original authors for feedback and approval and then posted at project webpage prior to data collection (preregistered); updated with results after replication and final versions posted after author feedback.

Fig. 1. Replication results.
(**A**) Plotted are 95% CIs of replication effect sizes (standardized to correlation coefficients). The standardized effect sizes are normalized so that 1 equals the original effect size (fig. S1 shows a nonnormalized version). Eleven replications have a significant effect in the same direction as in the original study [61.1%; 95% CI = (36.2%, 86.1%)]. The 95% CI of the replication effect size includes the original effect size for 12 replications [66.7%; 95% CI = (42.5%, 90.8%)]; if we also include the study in which the entire 95% CI exceeds the original effect size, this increases to 13 replications [72.2%; 95% CI = (49.3%, 95.1%)]. AER denotes the *American Economic Review* and QJE denotes the *Quarterly Journal of Economics*.
(**B**) Meta-analytic estimates of effect sizes, combining the original and replication studies. Plotted are 95% CIs ofcombined effect sizes (standardized to correlation coefficients). The standardized effect sizes are normalized as in (A) (fig. S1 shows a nonnormalized version). Fourteen studies have a significant effect in the same direction as the original study in the meta-analysis [77.8%; 95% CI = (56.5%, 99.1%)].



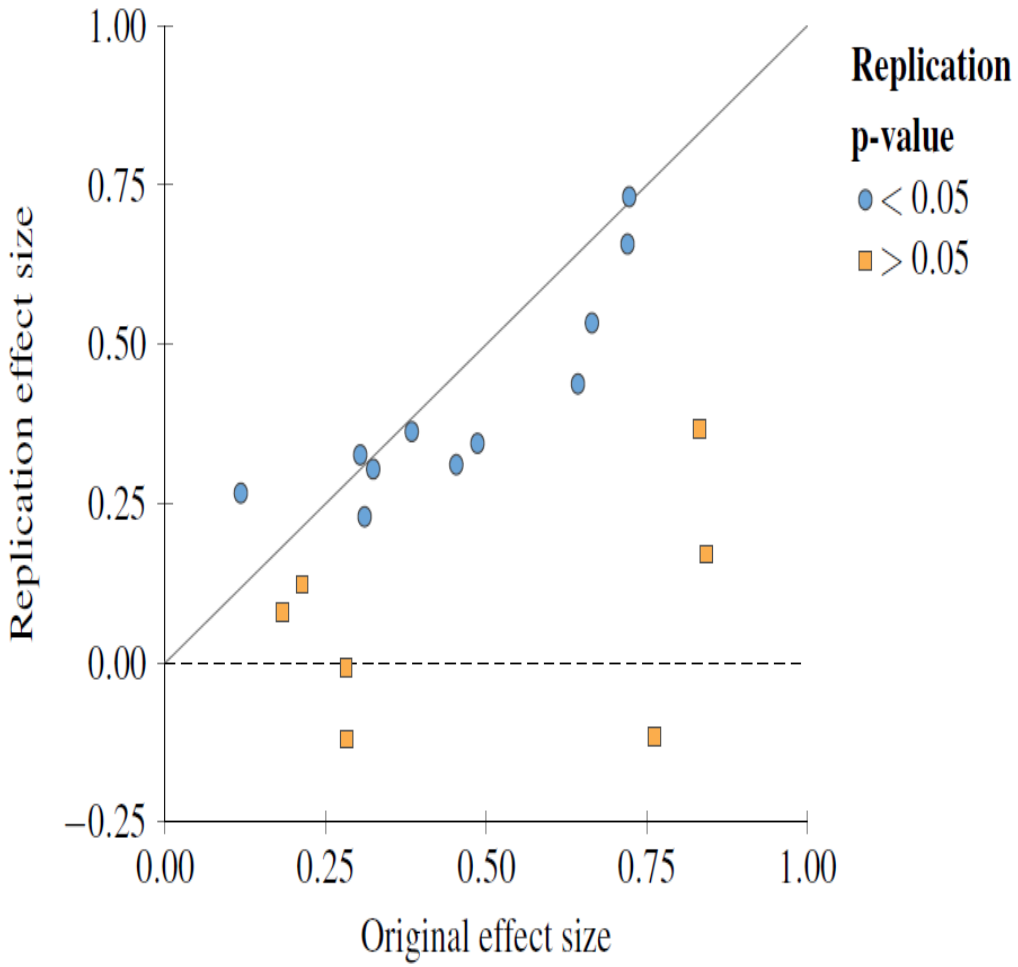


Fig. S3. Original study effect size versus replication effect size (correlation coefficients r).

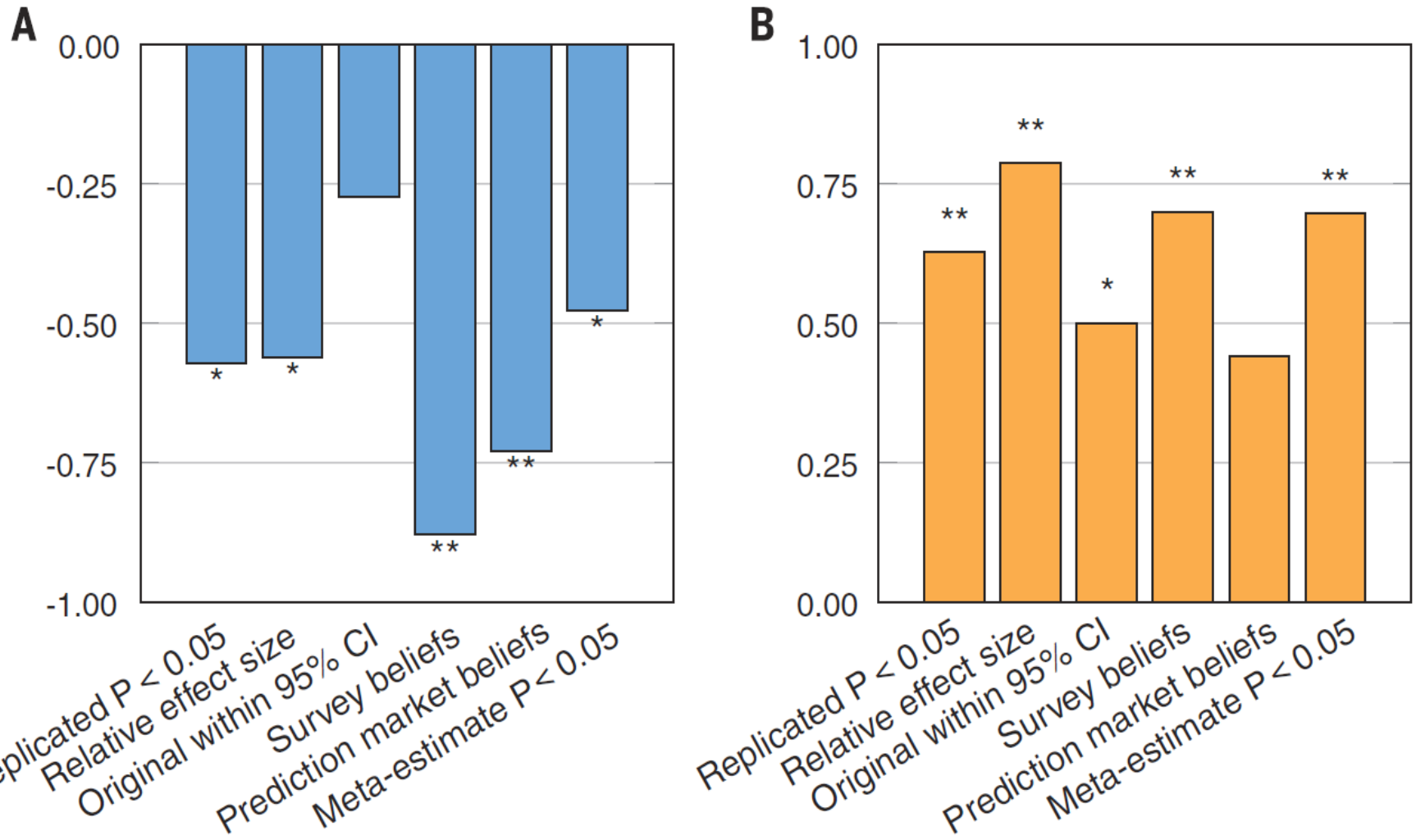


Fig. 3. Correlations between *P* values and sample sizes in original studies and replicability indicators. (**A**) The original *P* value is negatively correlated with all six replicability indicators, and five of these correlations are significant. (**B**) The original sample size is positively correlated with all six replicability indicators, and five of these correlations are significant. Spearman correlation coefficients are shown on the vertical axes. *$P < 0.05$; **$P < 0.01$.

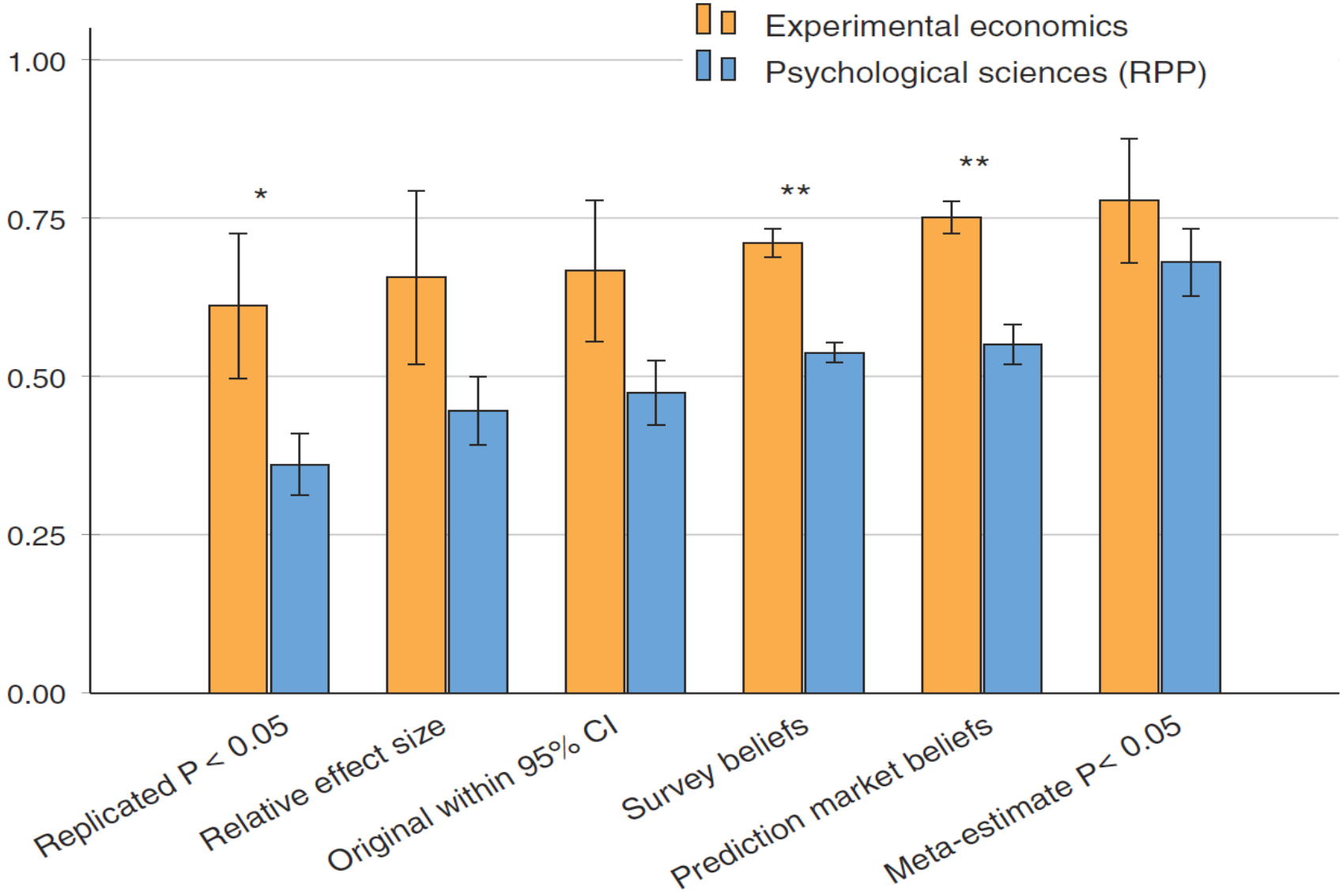


Fig. 4. A comparison of replicability indicators in experimental economics (this study) and psychological sciences (RPP). The graph shows means ± SE for replicability indicators. All six replicability indicators are higher for experimental economics; this difference is significant for three of the replicability indicators. The average difference in replicability across the six indicators is 19 percentage points. Details about the statistical tests are included in the supplementary materials. *$P < 0.05$; **$P < 0.01$.

Social Sciences Replication Project, SSRP (Camerer et al, Nature Human Behaviour 2018)

Replication of 21 between or within subject experimental papers published in Nature or Science in 2010-2015 (all papers meeting the inclusion criteria: experiments performed on students or other accesible subject pools that can be conducted in a standard lab in economics/psychology (including online studies); at least one significant ($p<0.05$) between or within subject treatment effect).

Selection of result to replicate within paper: the first experiment reporting a significant treatment effect included for papers with several experiments; the most prominent significant treatment effect in that experiment as emphasized in the original paper; lottery if ”ties” (n=3).

Two stage replication design: replication power 90% to detect 75% of the original effect size in Stage 1; if not replicated ($p>0.05$ or wrong direction) a second data collection with 90% power to detect 50% of the original effect size pooling Stage 1 and Stage 2. On average sample sizes about five times larger than in the original studies.

Five replication teams (Center for Open Science/University of Virginia, CalTech/Wharton, National University of Singapore, Stockholm School of Economics, University of Innsbruck).

Replication reports with design sent to the original authors for feedback and approval and then posted at OSF prior to data collection (preregistered); updated with results after replication and final versions posted after author feedback.

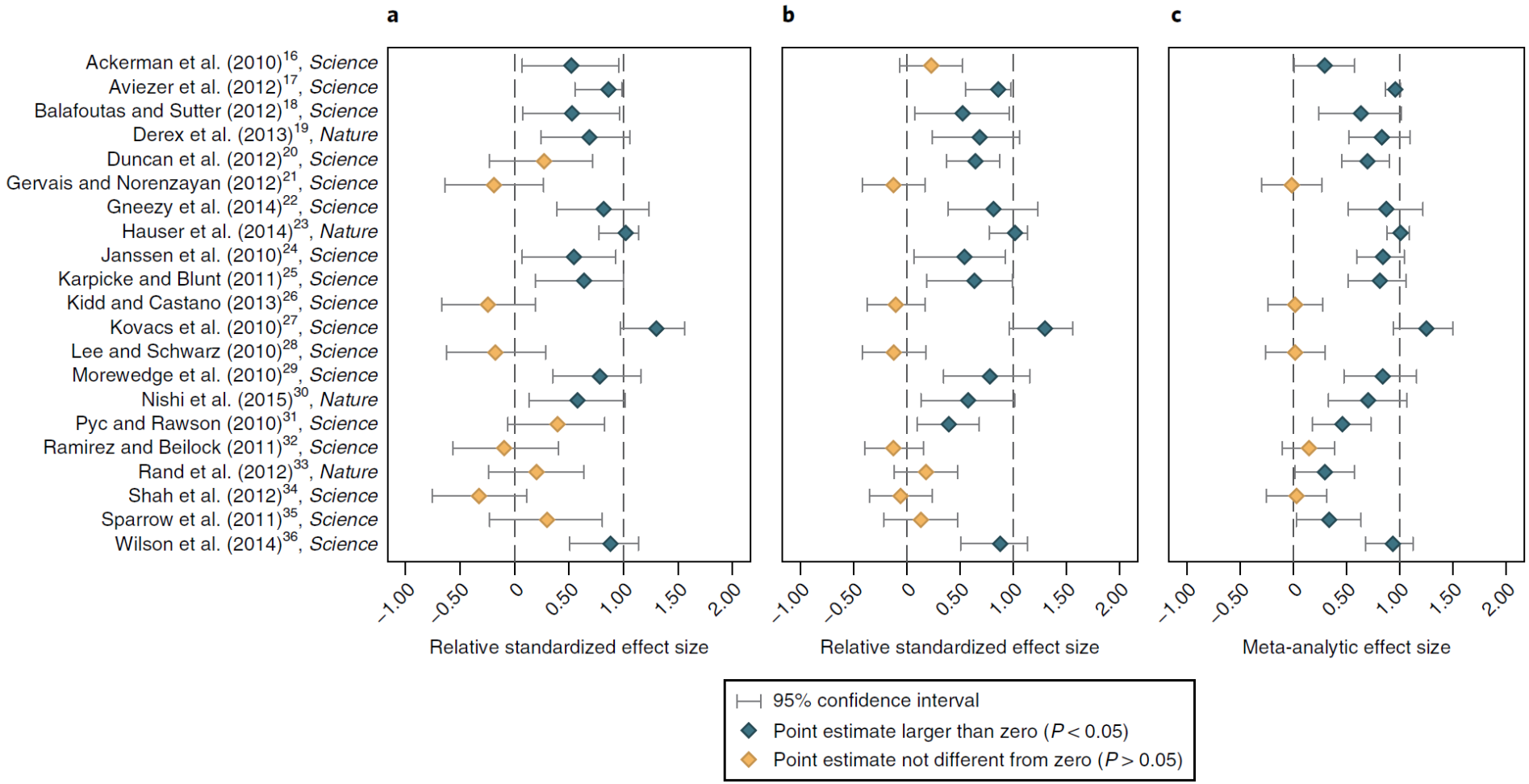
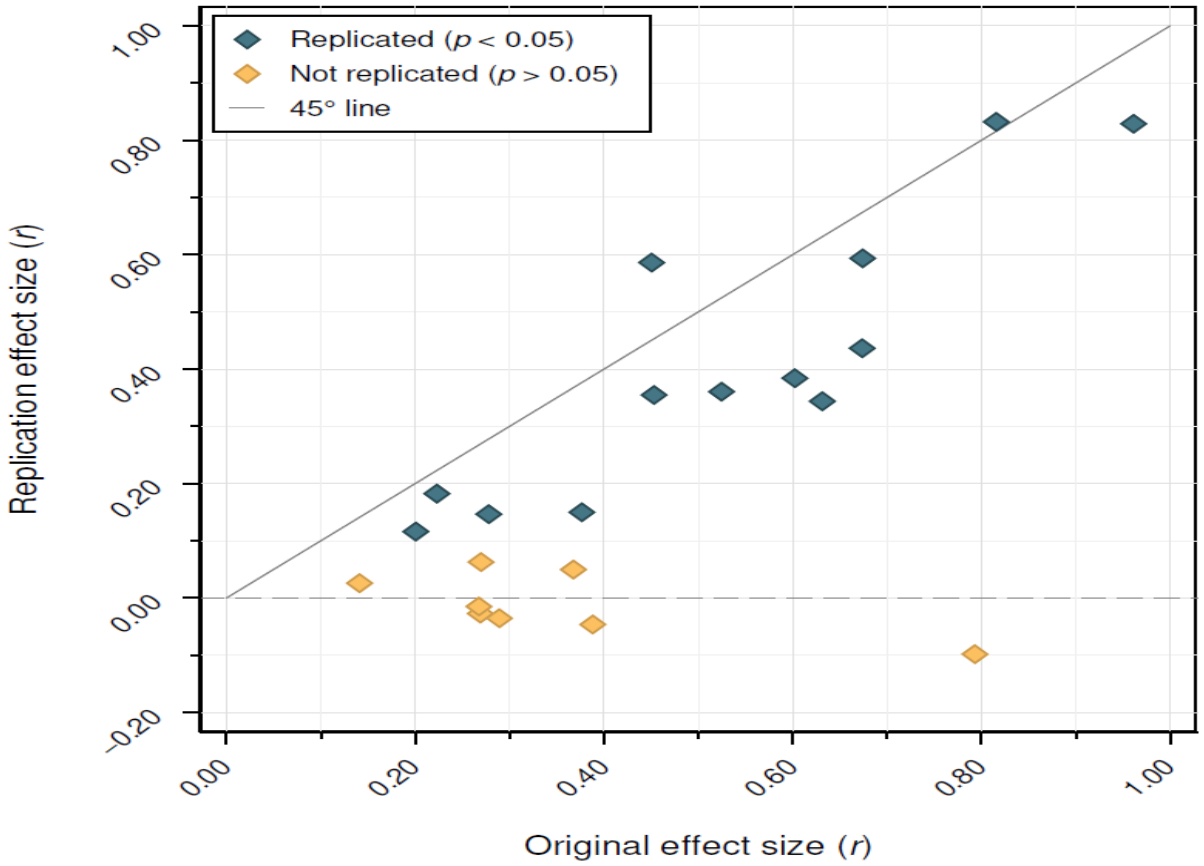


Fig. 1 | Replication results after stage 1 and stage 2. a, Plotted are the 95% CIs of the replication effect sizes (standardized to the correlation coefficients *r*) after stage 1. The standardized effect sizes are normalized so that 1 equals the original effect size. There is a significant effect in the same direction as in the original study for 12 out of 21 replications (57.1%; 95% CI = 34.1–80.2%). **b**, Plotted are 95% CIs of replication effect sizes (standardized to the correlation coefficients *r*) after stage 2 (replications not proceeding to stage 2 are included with their stage 1 results). The standardized effect sizes are normalized so that 1 equals the original effect size. There is a significant effect in the same direction as in the original study for 13 out of 21 replications (61.9%; 95% CI = 39.3–84.6%). **c**, Meta-analytic estimates of effect sizes combining the original and the replication studies. Shown are the 95% CIs of the standardized effect sizes (correlation coefficient *r*). The standardized effect sizes are normalized so that 1 equals the original effect size. Original and zero effect size are indicated by dashed lines. Sixteen out of 21 studies have a significant effect in the same direction as the original study in the meta-analysis (76.2%; 95% CI = 56.3–96.1%). Any deviations from the pre-registered replication protocols are listed towards the end of the Supplementary Methods. There was no deviation from the protocol for 7 replications[17,18,20–22,25,35], minor deviations for 12 replications[19,23,24,26,27,29–34,36], an unintended methodological deviation for one replication[28] and a continuation to the stage 2 data collection by mistake for one replication[16].



Supplementary Figure 4. Original study effect size versus replication effect size (correlation coefficients r). The diagonal line represents a replication effect size equal to the original effect size and the dotted line represents a replication effect size equal to zero. The mean standardized effect size (correlation coefficient, r) of the replications is 0.249 ($SD = 0.283$), compared to 0.459 ($SD = 0.229$) in the original studies. This difference is significant (Wilcoxon signed-ranks test, $n = 21$, $z = 3.667$, $p < 0.001$). The mean *relative* effect size of all the replications is 46.2% [95% CI = (27.0%, 65.5%)]; the mean *relative* effect size of the replications that replicated is 74.5% [95% CI = (60.1%, 88.9%)]; and the mean relative effect size of the replications that did not replicate is 0.3% [95% CI = (−12.4%, 13.1%)]. The Spearman correlation between the original effect size and the replication effect size is 0.574 [$p = 0.007$; 95% CI = (18.9%, 80.6%)].

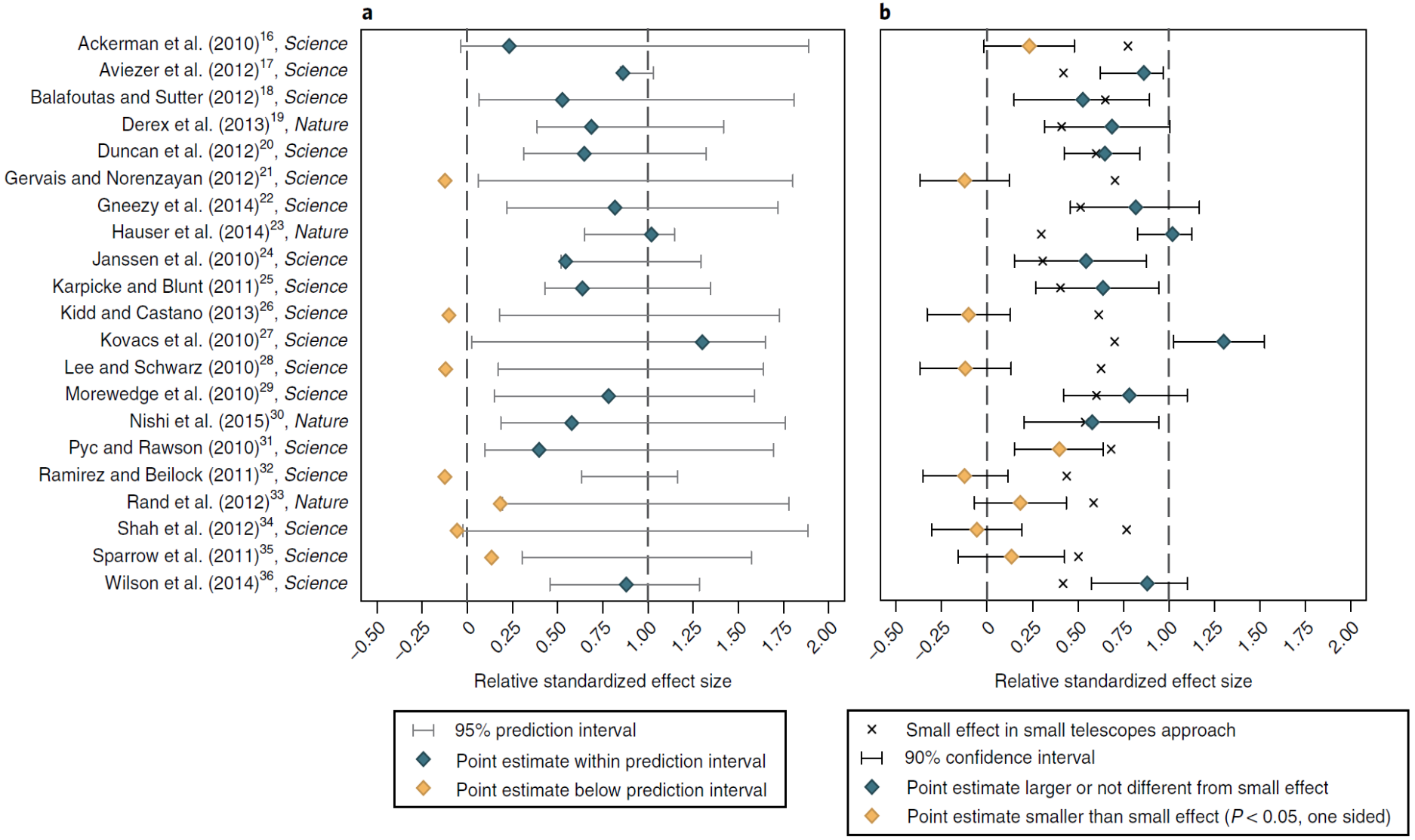


Fig. 2 | Replication results for two complementary replication indicators. a, Plotted are the 95% prediction intervals[47] for the standardized original effect sizes (correlation coefficient *r*). The standardized effect sizes are normalized so that 1 equals the original effect size. Original and zero effect size are indicated by dashed lines. Fourteen replications out of 21 (66.7%; 95% CI = 44.7–88.7%) are within the 95% prediction interval and replicate according to this indicator. **b**, Plotted are the 90% CIs of replication effect sizes in relation to small-effect sizes as defined by the small telescopes approach[46] (the effect size that the original study would have had 33% power to detect). Effect sizes are standardized to correlation coefficients *r* and normalized so that 1 equals the original effect size. A study is defined as failing to replicate if the 90% CI is below the small effect. According to the small telescopes approach, 12 out of 21 (57.1%; 95% CI = 34.1–80.2%) studies replicate.

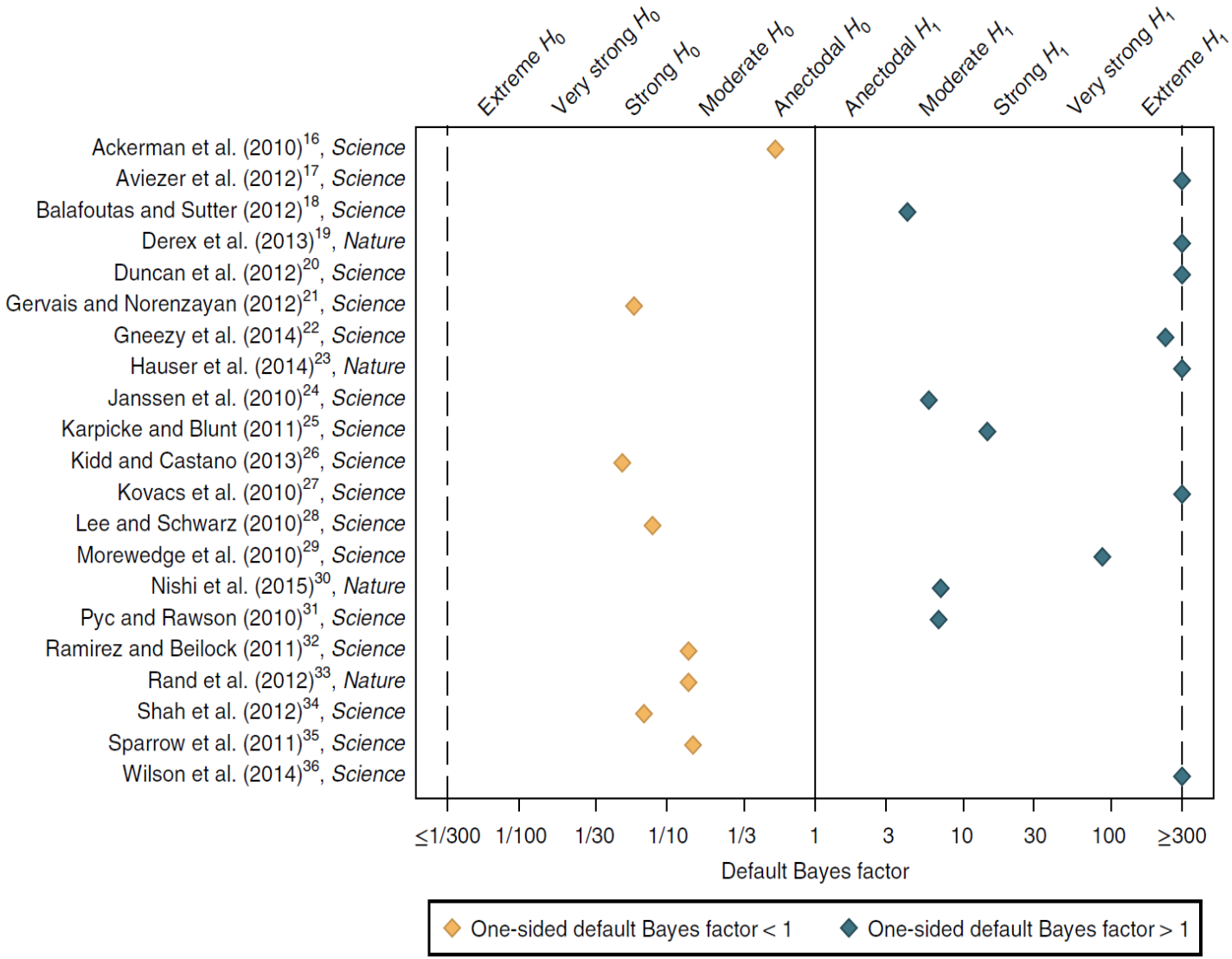
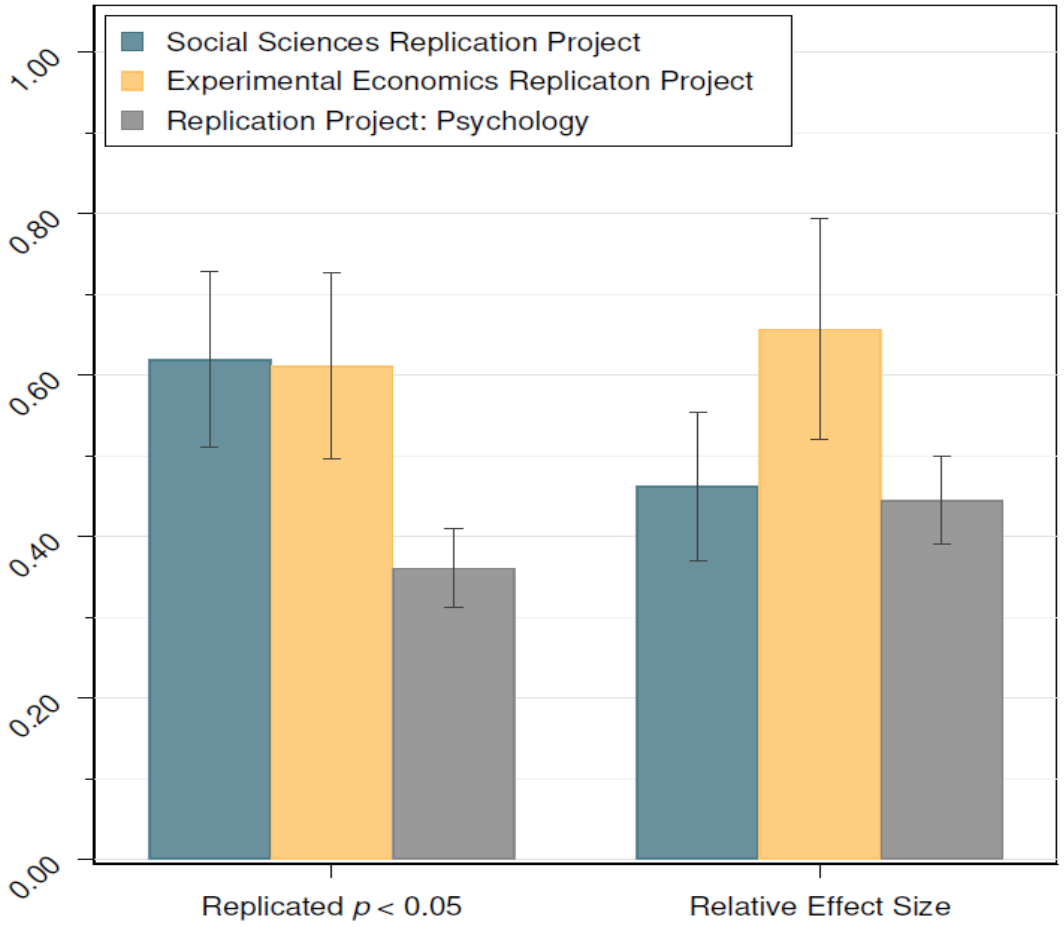


Fig. 3 | Default Bayes factors (one sided) for the 21 replications. A default Bayes factor[48] above 1 favours the hypothesis of an effect in the direction of the original paper and a default Bayes factor below 1 favours the null hypothesis (H_0) of no effect. The evidence categories proposed by Jeffreys[52] are also shown (from extreme support for the null hypothesis to extreme support for the original hypothesis). The default Bayes factor is above 1 and provides evidence in favour of an effect in the direction of the original study for the 13 out of 21 (61.9%) studies that replicated according to the statistical significance criterion. This evidence is strong to extreme for 9 out of 21 (42.9%) studies. The default Bayes factor is below 1 for 8 out of 21 (38.1%) studies, providing evidence in support of the null hypothesis; this evidence is strong to extreme for 4 out of 21 (19.0%) studies. Values more extreme than 1/300 or 300 are represented on the dashed lines. H_1, alternative hypothesis.



Supplementary Figure 9. A comparison of replicability indicators between the Social Sciences Replication Project (SSRP), the Experimental Economics Replication Project (EERP)[8], and the Reproducibility Project: Psychology (RPP)[7]. Error bars denotes ±se.

Additional result (not related to Figure above): Spearman correlation between original p-value and replication outcome -0.40 (p=0.069); in between the correlations of -0.33 and -0.57 in RPP and EERP.

Using Prediction Markets to Estimate Peer Beliefs About Replication (Dreber et al. PNAS 2015)

Prediction markets can be used to predict if a study will replicate or not

Participants trade contracts that pay X (e.g. $1) if a study replicates and 0 otherwise (”short selling” also possible).

The market price can be interpreted as the probability of replication.

Collect and disseminate information among participants.

First study on 44 replications in RPP (41 of these replications were completed)

Survey about beliefs of replication prior to entering the markets.

Researchers in mainly psychology participated (47 and 45 participants in two sets of markets). Endowed with $100 each.

Example of trading page, from EERP

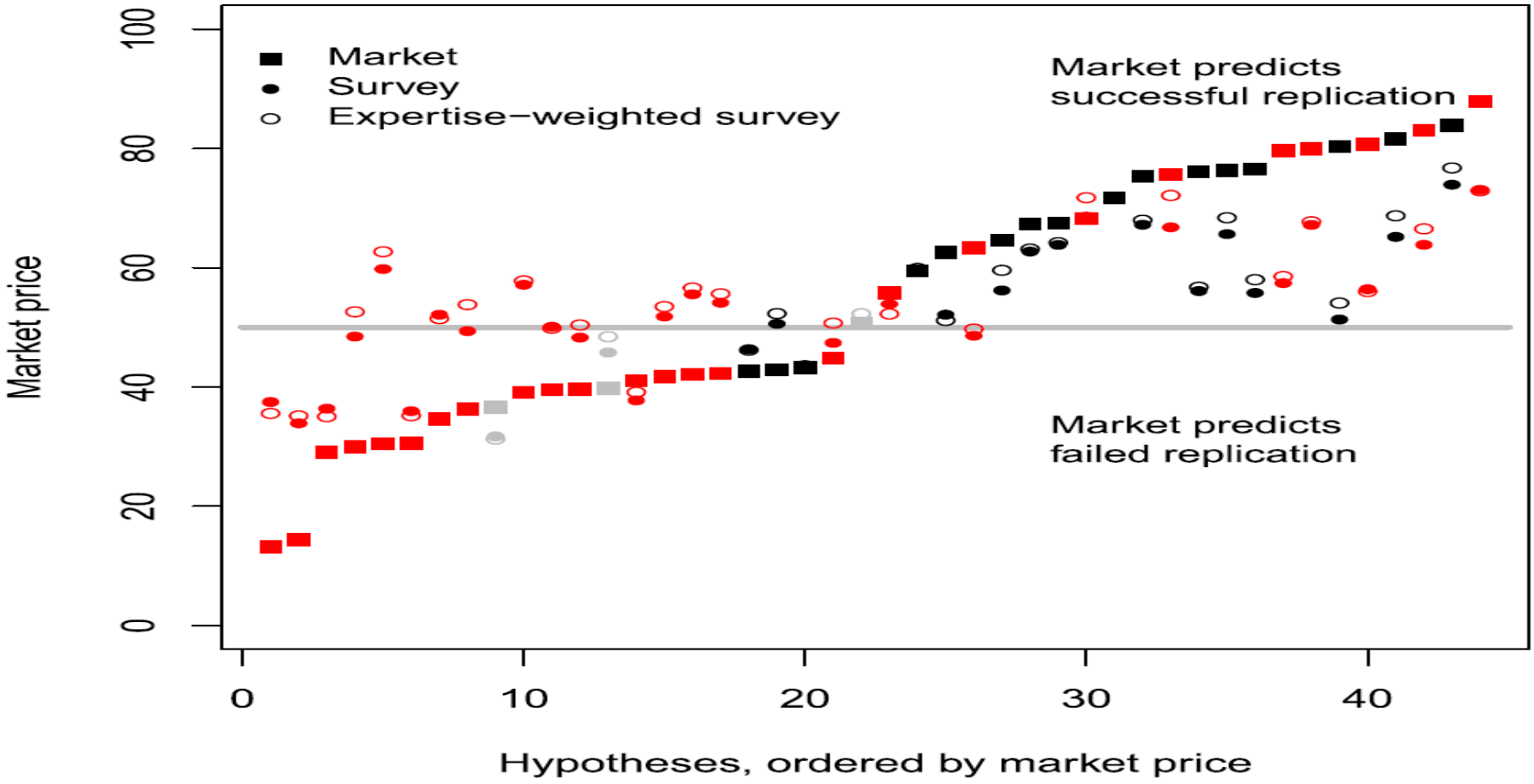


Fig. 1. Prediction market performance. Final market prices and survey predictions are shown for the replication of 44 publications from three top psychology journals. The prediction market predicts 29 out of 41 replications correctly, yielding better predictions than a survey carried out before the trading started. Successful replications (16 of 41 replications) are shown in black, and failed replications (25 of 41) are shown in red. Gray symbols are replications that remained unfinished (3 of 44).

Average market price 55% (compared to 39% replication rate in this subsample of RPP). Market beliefs correctly predicted 71% of the replications and survey beliefs correctly predicted 58% of replication outcomes. Pearson correlation 0.42 between market beliefs and replication outcomes and 0.27 between survey beliefs and replication outcomes.

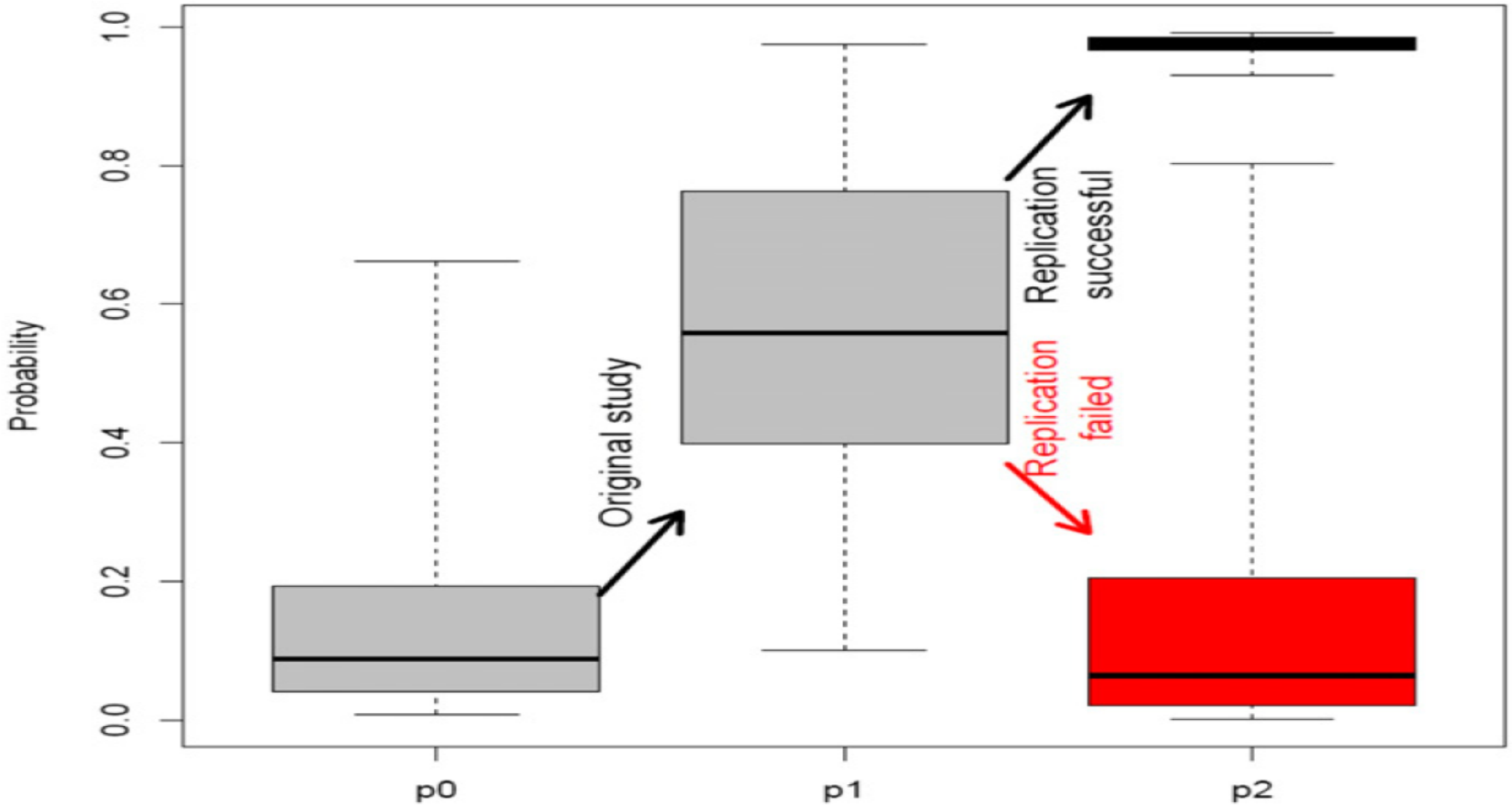


Fig. 3. Probability of a hypothesis being true at three different stages of testing: before the initial study (p_0), after the initial study but before the replication (p_1), and after replication (p_2). "Error bars" (or whiskers) represent range, boxes are first to third quartiles, and thick lines are medians. Initially, priors of the tested hypothesis are relatively low, with a median of 8.8% (range, 0.7–66%). A positive result in an initial publication then moves the prior into a broad range of intermediate levels, with a median of 56% (range, 10–97%). If replicated successfully, the probability moves further up, with a median of 98% (range, 93.0–99.2%). If the replication fails, the probability moves back to a range close to the initial prior, with a median of 6.3% (range, 0.01–80%).

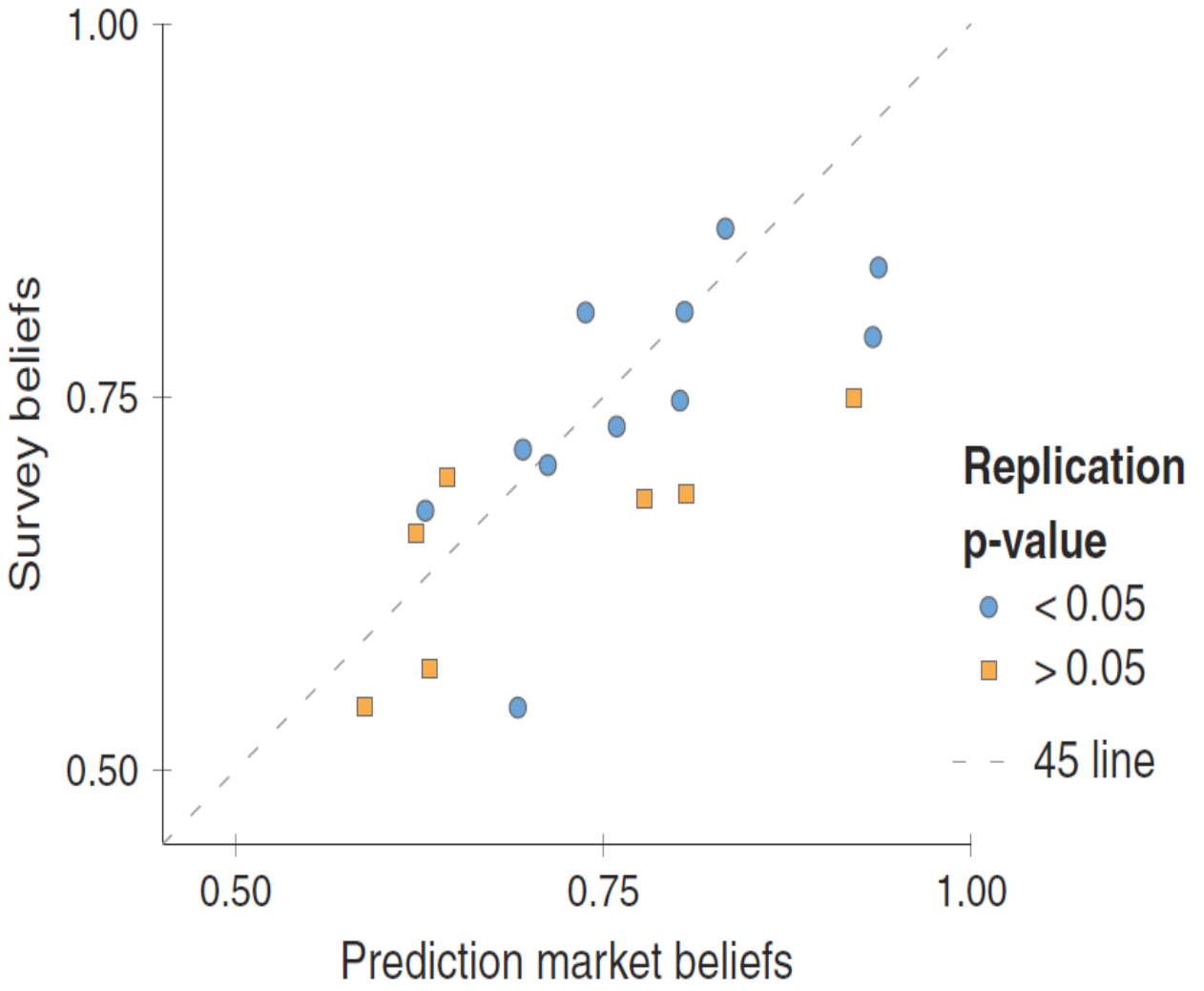
Prediction Markets in EERP (Camerer et al. Science 2016)

Prediction markets of the 18 replications in EERP.

Survey about beliefs of replication prior to entering the markets.

Researchers in mainly economics participated (97 participants). Endowed with $50 each.

Fig. 2. Prediction market and survey beliefs. A plot of prediction market beliefs and survey beliefs, in relation to whether the original result was replicated with $P < 0.05$ in the original direction. The mean prediction market belief in a successful replication is 75.2% [range, 59% to 94%; 95% CI = (69.7%, 80.6%)], and the mean survey belief is 71.1% [range, 54% to 86%; 95% CI = (66.4%, 75.8%)]. The prediction market beliefs and survey beliefs are highly correlated (Spearman correlation coefficient = 0.79, $P < 0.001$, $n = 18$). Both the prediction market beliefs (Spearman correlation coefficient = 0.30, $P = 0.232$, $n = 18$) and the survey beliefs (Spearman correlation coefficient 0.52, $P = 0.028$, $n = 18$) are positively correlated with the ranked degree of replication success.



Prediction Markets in SSRP (Camerer et al. Nature Human Behaviour 2018)

Prediction markets of the 21 replications in SSRP.

Survey about beliefs of replication prior to entering the markets.

Two treatments (one with trading on replication outcomes in Stage 1; one with trading outcomes in both Stage 1 and Stage 2)

Researchers in mainly social sciences participated (114 participants in treatment 1 and 92 participants in treatment 2). Endowed with $50 each.

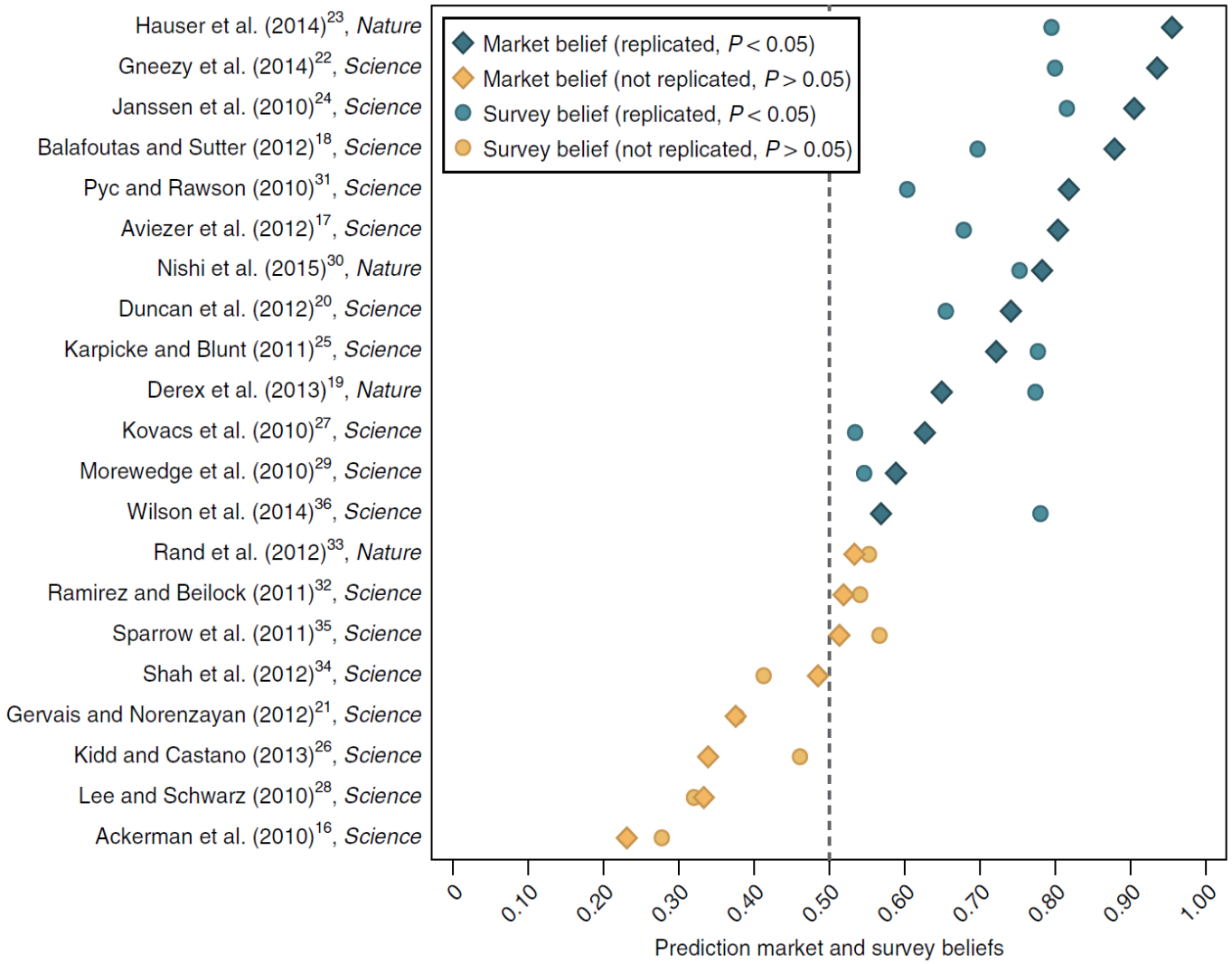


Fig. 4 | Prediction market and survey beliefs. The prediction market beliefs and the survey beliefs of replicating (from treatment 2 for measuring beliefs; see the Supplementary Methods for details and Supplementary Fig. 6 for the results from treatment 1) are shown. The replication studies are ranked in terms of prediction market beliefs on the *y* axis, with replication studies more likely to replicate than not to the right of the dashed line. The mean prediction market belief of replication is 63.4% (range: 23.1–95.5%, 95% CI = 53.7–73.0%) and the mean survey belief is 60.6% (range: 27.8–81.5%, 95% CI = 53.0–68.2%). This is similar to the actual replication rate of 61.9%. The prediction market beliefs and survey beliefs are highly correlated, but imprecisely estimated (Spearman correlation coefficient: 0.845, 95% CI = 0.652–0.936, $P < 0.001$, $n = 21$). Both the prediction market beliefs (Spearman correlation coefficient: 0.842, 95% CI = 0.645–0.934, $P < 0.001$, $n = 21$) and the survey beliefs (Spearman correlation coefficient: 0.761, 95% CI = 0.491–0.898, $P < 0.001$, $n = 21$) are also highly correlated with a successful replication.