Study	Hypothesis
Ackerman et al. (2010), Science	Participants that evaluate a resume while using a heavier clipboard will rate the resume as better overall compared to the participants that evaluate the resume while using a lighter clipboard. The original study used F-test for a two condition comparison, $p < 0.05$. Original test statistics: Heavy Condition: $N = 26$, $M = 5.80$, $SD = 0.76$; Light Condition: $N = 28$, $M = 5.38$, $SD = 0.79$. $F(1, 52) = 4.08$, $p = 0.049$. If there were no covariates in the model, we will convert the F to t for comparison with the replication tests.
Aviezer et al. (2012), Science	The body context is diagnostic for the affective valence of the situation during peak intensity moments (tests the hypothesis of a higher mean valence rating of winning bodies versus losing bodies in the 'body treatment' in Experiment 1; within subjects variation, paired t-test, $t(14) = 13.07$, $p < 0.0001$, p. 1226 and Fig. 1c).
Balafoutas and Sutter (2012), Science	With preferential treatment of women – i.e., each woman's performance is automatically increased by one unit in the competition – more women will choose to compete (a comparison of the fraction of women who chose the tournament scheme rather than the piece rate scheme in the 'preferential treatment one (PT1)' versus the 'control treatment (CTR)'; $\chi^2(1) = 5.62$, $p = 0.018$, p. 580). (This hypothesis was picked by lottery instead of comparing PT2 to CTR; $\chi^2(1) = 10.89$, $p = 0.001$, p. 580).
Derex et al. (2013), Nature	The probability of maintaining cultural diversity (that is, observing both tasks in the group) increases with group size; $\chi^2(1) = 16.3$, the p-value < 0.0001 (exact 0.000054) (p. 389; measured at the group level with group sizes, 2, 4, 8, and 16).
Duncan et al. (2012), Science	Similar objects are more accurately identified as being similar if they are preceded by new objects than if they are preceded by old objects (a comparison of the fraction of objects rated as similar in trials where they are preceded by new objects compared to trials where they are preceded by old objects in Study 1b (within-subject variation), $t(14) = 3.41$, $p = 0.0042$, p. 486).
Gervais and Norenzayan (2012), Science	Priming analytic thinking via images of 'The Thinker' increases religious disbelief compared to viewing control images of a visually similar artwork; a t-test, $p < 0.05$ using a two-tailed test. Original test statistics: $N = 57$ (31 in Control condition, 26 in Disbelief condition); Control belief in god (100-pt scale): $M = 61.55$, $SD = 35.68$; Disbelief: $M = 41.42$, $SD = 31.47$; $t(55) = 2.24$; $p = 0.029$ (reported as $p = 0.03$).
Gneezy et al. (2014), Science	The likelihood of choosing a charity is higher when potential donors know that the overhead is already paid for, than when the donors pay for overhead themselves (a comparison of the fraction choosing to donate to '*charity: water*' between the '50% overhead, covered treatment' and the '50% overhead treatment', $z = 3.00$, $p < 0.01$ (exact $p = 0.0027$), p. 633). (This hypothesis was picked by lottery instead of comparing the 'no overhead treatment' and the '50% overhead treatment', $z = 3.27$, $p < 0.01$, p. 633.)
Hauser et al. (2014), Nature	Choosing an extraction level for all group members using median voting leads to a higher degree of sustainability of a common pool than allowing each individual to choose their own extraction amount. That is, a comparison of the average probability that the common pool was sustained by the first generation between the voting treatment and the unregulated treatment (in both treatments there is an 80% probability that a new generation occurs and an extraction threshold of 50%). To evaluate this hypothesis, a linear probability model with a treatment dummy variable is used; see the 1st generation regression equation in Table S1; $p = 1.427e^{-10}$ (reported as $p < 0.001$) in a t-test ($t(38) = 8.696$) of the treatment dummy variable coefficient.
Janssen et al. (2010), Science	Communication increases average earnings in a common-pool resource game with spatial and temporal resource dynamics. A comparison of net earnings between the *NCP* condition and the *C* condition in periods 1 to 3 showed p-value < 0.001 with the Mann-Whitney test ($z = 5.761$ and $p = 8.362e^{-9}$).
Karpicke and Blunt (2011), Science	In a memory test one week after learning, Retrieval Practice leads to participants recalling more correct information than Concept-Mapping. A t-test, $p < 0.05$ using a two-tailed test, comparing the Retrieval Practice and Concept Mapping conditions. Original test statistics: $N = 40$ (20 in each condition); Mean performance= 0.67 in the Retrieval Practice condition and 0.45 in the Concept Mapping condition. The comparison between Retrieval Practice and Concept Mapping was reported as $F(1, 38) = 21.63$; $p = 0.000039$.
Kidd and Castano (2013), Science	Reading literary fiction improves affective Theory of Mind (a comparison of the mean Reading the Mind in the Eyes Test (RMET) score between the literary fiction treatment and the nonfiction treatment in experiment 1; ANOVA test, $F(1, 82) = 6.40$ and $p = 0.0133$ (reported as $p = 0.01$, p. 378).
Kovacs et al. (2010), Science	Participants automatically project agents' beliefs and store them in a way similar to that of their own representation about the environment. A comparison of the mean reaction time between the 'P-A- treatment' and the 'P-A+ treatment' in Study 1 (within subject variation), shows that reaction time is shorter in the P-A+ treatment; results show that $t(23) = 2.42$, p-value $= 0.02$ (exact $p = 0.0238$).
Lee and Schwarz (2010), Science	Hand washing will significantly reduce the need to justify one's choice by increasing the perceived difference between alternatives. Specifically, the mean difference between the rankings of the chosen and rejected albums before and after making the choice will be greater for the soap examining condition compared to the soap hand washing condition. F-test assessing the interaction between before-after and hand-washing condition, $p < 0.05$. Original test statistics: (i) *Soap examining condition:* Mean difference between chosen and rejected, before making choice: $M = 0.14$, $SD = 1.01$. Mean difference between chosen and rejected, after making choice: $M = 2.05$, $SD = 1.96$. (ii) *Soap hand washing condition:* Mean difference between chosen and rejected, before making choice: $M = 0.68$, $SD = 0.75$. Mean difference between chosen and rejected, after making choice: $M = 1.00$, $SD = 1.41$. Interaction of before-after and hand-washing: $F(1, 38) = 6.74$, $p = 0.0133$ (reported as $p = 0.01$).
Morewedge et al. (2010), Science	Repeatedly imagining eating a food subsequently reduces the actual consumption of that food (a comparison of the 30-repetition treatment and the control treatment in experiment 1; independent samples t-test, $t(30) = 2.78$, $p = 0.0092$, provided by the original authors. The analysis in the original study pools the variance across the 30-repetition, the 3-repetition, and the control condition and reports an ANOVA result of $F(1, 46) = 4.50$, $p = 0.0393$, p. 1531.) (This hypothesis was picked by lottery instead of comparing the mean consumption of M&M's between the 30-repetition treatment and the 3-repetition treatment; $F(1, 46) = 5.81$, $p < 0.05$, p. 1531).
Nishi et al. (2015), Nature	In initially unequal situations, wealth visibility leads to greater inequality than when wealth is invisible (a comparison of the mean Gini coefficient between the visible and high initial inequality treatment and the invisible and high initial inequality treatment; OLS regression of the session/round Gini coefficient as the dependent variable and multiway clustering of standard errors at the session and round level; regression equation (5) in Table S2, $p = 0.0044$ of a t-test of the treatment dummy variable coefficient, $t(198) = 2.881$).
Pyc and Rawson (2010), Science	Retrieval of mediators is greater with test-restudy practice than with restudy practice; a comparison of mean mediator retrieval between the test-restudy and the restudy treatments within the CMR treatment, p. 335, $t(34) = 2.37$ and p-value $= 0.02$, t-value and p-value from authors). Note that a successful retrieval in each of the final test questions is defined as correctly recalling any of the keyword mediators that had been generated during session 1.
Ramirez and Beilock (2011), Science	In a high-pressure in-lab math test, those writing for 10 minutes about their deepest thoughts and feelings regarding the upcoming test improve more on that test compared to simply sitting quietly; an F-test, $p < 0.05$ using a two-tailed test. Original test statistics: $N = 20$ (10 in each condition); Expressive writing $M_{pre} = 0.86$ ($SD = 0.09$), $M_{post} = 0.91$ ($SD = 0.05$), Control $M_{pre} = 0.82$ ($SD = 0.09$), $M_{post} = 0.70$ ($SD = 0.11$); $F(1, 18) = 30.53$; $p = 0.00003$ (reported as $p < 0.01$, p. S11).
Rand et al. (2012), Nature	Priming intuition increases cooperation in a public goods game compared to priming reflection (a comparison of the mean contribution in a public goods game between the 'intuition-good'/'reflection-bad' treatments and the 'intuition-bad'/'reflection-good' treatments; a Tobit regression (with robust standard errors) with a treatment dummy variable, regression equation (1) in Table S11; $z = 2.617$, $p = 0.0089$ in a z-test of the treatment dummy variable coefficient).
Shah et al. (2012), Science	Low-wealth subjects, that are given fewer chances to win in repeated 'Wheel of Fortune' type word puzzle games, perform worse in a subsequent attention task (Dots-Mixed task) than do high-wealth individuals (a comparison of the mean performance on the Dots-Mixed task between the 'poor treatment' and the 'rich treatment'; ANOVA test, $F(1, 54) = 4.16$ and $p = 0.046$, p. 683).
Sparrow et al. (2011), Science	Computer terms are more accessible than general words after answering a block of hard trivia questions; measured as longer color-naming reaction times in a Modified Stroop Task after priming with computer terms compared to priming with non-computer terms (paired t-test, within subject variation; $t(45) = 3.26$, $p = 0.0021$, study 1, p. 776, and Fig. 1).
Wilson et al. (2014), Science	An external activity from a list (e.g. watching television or reading a book) for 12 minutes is rated as being more enjoyable than a 12 minute 'thinking period' entertaining themselves with their thoughts (a higher average self-rated enjoyment (the mean of three nine-point scale items) in the 'external activities' treatment than in the 'standard thought instructions' treatment in Study 8, $t(28) = 4.83$, $p = 0.000044$, p. 76).