nature human behaviour

Article

https://doi.org/10.1038/s41562-023-01749-9

High replicability of newly discovered social-behavioural findings is achievable

In the format provided by the authors and unedited

Table of Contents

Section 1. The Four Participating Laboratories

The prospective replication project originated from a 2012 meeting about the observation of declining effect sizes between original findings and subsequent investigations organized by Jonathan Schooler at UC Santa Barbara and funded by the Fetzer Franklin Fund. The four participating laboratories were UC Santa Barbara (PI: Jonathan Schooler, Department of Psychology), Stanford University (PI: Jon Krosnick, Department of Communication), UC Berkeley (PI: Leif Nelson, Marketing, Haas School of Business), and University of Virginia (PI: Brian Nosek, Department of Psychology). Each laboratory had prior and subsequent experience independent of this project conducting replications of others' work and having their own studies replicated by others–sometimes successfully and sometimes unsuccessfully[1-5]. This includes some members of the present team failing to replicate prior findings from other members of the present team[4]. Thus, the team was composed of replicators and original authors from prior replication efforts that had experienced a diversity of replication failures and successes concordant with the variability in observing replication success across the social-behavioral sciences.

Section 2. Interactions between laboratories

By default, labs were discouraged from interacting with each other about replication designs and implementation to maximize independence of tests. Originators of each discovery wrote a complete methods section and provided specialized materials (e.g., videos) if any. The project manager received final methods and circulated them on the planned schedule to the replication labs. If replication labs needed to seek clarification from the originating lab to conduct the replication study, the interaction was limited to clarification of the key question and replication labs returned to designing their replication study independently. Overall, for 14 of the 48 (29%) independent replications there was no interaction between the originating lab and the replication lab beyond sharing the methods section and key materials. Supplementary Table 1 summarizes the interactions that occurred between laboratories.

Supplementary Table 1: Instances of replicating labs requesting additional information from the originating lab to run a direct replication of the study. Columns indicate which school requested the additional information.

Study Name	UCSB Replication	UVA Replication	Berkeley Replication	Stanford Replication
Tumor				Additional Info Given: Clarified 'reverse coded'
Minimal Groups				
Cookies				
Label			Helped Stanford design materials and ran a pilot for them.	
Self-control				
Orientation			Additional Info Requested	
Referrals			Additional Info Requested	Additional Info Given: Clarified randomization
Ads				
Fast Social Desirability (FSD)			Additional Info Requested	
Prediction			Additional Info requested	
Fairness				
Ostracism	Additional Info Given: Javascript help on Safari			
Misattribution			Additional info requested	Additional Info Given
Redemption	.qsf shared		.qsf shared	.qsf shared
Worse				
Misreporting			Additional Info Given	

Section 3. Studies

The complete list of studies can be found in Supplementary Table 2.

Supplementary Table 2. Links to materials documenting all 16 confirmatory tests and replications, including data collection materials, data, analysis syntax, output, and report of findings, and discovery-oriented research, if any. Preregistration documents describe design and analysis plans. Order of Replication 1-4 matches the order the replications were run in (see Supplementary Information, Section 6, Table 3).

Study Name	Confirmatory Test	Replication 1	Replication 2	Replication 3	Replication 4
Tumor	Project: https://osf.io/4n8pf/	Project: https://osf.io/5ypsq/?view_only=2870c6f18eda4a4c8d6c769366ebf33d	Project: https://osf.io/8de3f/	Project: https://osf.io/u8hq9/?view_only=680c20470a904cd8831182f7a1595f6f	Project: https://osf.io/utck4/
	Prereg: https://osf.io/zm87e	Prereg: https://osf.io/z9kuz	Prereg: https://osf.io/k62nh	Prereg: https://osf.io/58n4b	Prereg: https://osf.io/yagxp
Minimal Groups	Project: https://osf.io/y8adg/	Project: https://osf.io/y7u5v/?view_only=31be66b525544598b7c029e273bb452a	Project: https://osf.io/8kc59/?view_only=b8a9beddea0c4d899739503eb6c516f5	Project: https://osf.io/jr9pc/	Project: https://osf.io/kzwa6/
	Prereg: https://osf.io/txj9e	*Prereg*: https://osf.io/k5p4f	Prereg: https://osf.io/hwd8m	Prereg: https://osf.io/sdzh5	Prereg: https://osf.io/uytk9
Cookies	Project: https://osf.io/8xdwc/	Project: https://osf.io/5h2gw/	Project: https://osf.io/2ntf3/	Project: https://osf.io/dc4xm/	Project: https://osf.io/3vz4k/?view_only=da10896b68fe4420bf6c65a3a7bd64f6

	Prereg: https://osf.io/74vu2	Prereg: https://osf.io/72xgd	Prereg: https://osf.io/v9658	Prereg: https://osf.io/tnkw9	Prereg: https://osf.io/vmcy2
Label	Project: https://osf.io/f5zdr/	Project: https://osf.io/a5w8d/	Project: https://osf.io/zmkn6/?view_only=c77fbee4f1b240a2837849912dadbe60	Project: https://osf.io/wn9af/	Project: https://osf.io/jq64y/?view_only=3421e6ac51c642b98c5dcf9507cd27ea
	Prereg: https://osf.io/dw3fm	Prereg: https://osf.io/zyst7	Prereg: https://osf.io/7n9yg	Prereg: https://osf.io/n42fr	Prereg: https://osf.io/rtd9b
Self-control	Project: https://osf.io/h2pwm/	Project: https://osf.io/xjas9/?view_only=9c94526a2de24ba8b8a796e2e9e9a00f	Project: https://osf.io/pkuv9/?view_only=83d22ae30533439ab7dd33b7c7595522	Project: https://osf.io/fbnkg/	Project: https://osf.io/5xqya/?view_only=63fedd7e94964d5a814d94cd073b2935
	Prereg: https://osf.io/5x9ha	Prereg: https://osf.io/qmj98	Prereg: https://osf.io/v2qjs	Prereg: https://osf.io/j3sud	Prereg: https://osf.io/kpucv
Orientation	Project: https://osf.io/s6qdv/	Project: https://osf.io/5ygj8/	Project: https://osf.io/pd4s9/?view_only=97405c457b8349ef844196ee267e795b	Project: https://osf.io/t62bd/	Project: https://osf.io/rtb34/?view_only=d00211f1229c4635b69a84444cd72f08
	Prereg: https://osf.io/54tz3	Prereg: https://osf.io/xq89s	Prereg: https://osf.io/8nzwm	Prereg: https://osf.io/x3pjn	Prereg: https://osf.io/f9yng

Referrals	Project: https://osf.io/v3thd/	Project: https://osf.io/bsg7f/?view_only=9129739897b04a4ba673297a9e3e6ce4	Project: https://osf.io/wtz9g/	Project: https://osf.io/e5u42/?view_only=1c32808a01ee4c8c816480825ad5bebf	Project: https://osf.io/jepfa/
	Prereg: https://osf.io/9yw62	Prereg: https://osf.io/q6rnz	Prereg: https://osf.io/qxgwm	Prereg: https://osf.io/pjr9c	Prereg: https://osf.io/qye3h
Ads	Project: https://osf.io/yhux4/	Project: https://osf.io/zc8p9/	Project: https://osf.io/zq8gy/	Project: https://osf.io/txme4/?view_only=8cf18a2babc1499e98ef57dbb9926a80	Project: https://osf.io/rpxa8/?view_only=4a111fdbe8ef404aa80dcf83c9ac1fa7
	Prereg: https://osf.io/24ndc	Prereg: https://osf.io/myx9c	Prereg: https://osf.io/tw9rh	Prereg: https://osf.io/vxkep	Prereg: https://osf.io/9u2nx
Fast Social Desirability (FSD)	Project: https://osf.io/yxwc3/	Project: https://osf.io/jse68/	Project: https://osf.io/m6avf/?view_only=fda7fcac8d754d4bbe033ec34a134414	Project: https://osf.io/duzkv/	Project: https://osf.io/2gteq/?view_only=8f38e1adaba149cb8bb1fb1477a759d9
	Prereg: https://osf.io/zacb3	Prereg: https://osf.io/63z4g	Prereg: https://osf.io/yu8hw	Prereg: https://osf.io/b2489	Prereg: https://osf.io/v3exz
Prediction	Project: https://osf.io/edt62/	Project: https://osf.io/vp9rj/	Project: https://osf.io/vgy4q/	Project: https://osf.io/cnem6/?view_only=e458fc4deb394549ab532dc9ef647f1e	Project: https://osf.io/u8hq9/

	Prereg: https://osf.io/yrx2t	Prereg: https://osf.io/ys62a	Prereg: https://osf.io/3gvqh	Prereg: https://osf.io/nc2t9	Prereg: https://osf.io/58n4b
Fairness	Project: https://osf.io/ctb4m/	Project: https://osf.io/49n8s/?view_only=81819e97d13e4099b3c2c94052efefd4	Project: https://osf.io/2zq9u/?view_only=c8c0e29dbd2248db8dbb2554358be61c	Project: https://osf.io/x82uz/	Project: https://osf.io/vfsq4/
	Prereg: https://osf.io/63x4k	Prereg: https://osf.io/xmyj7	Prereg: https://osf.io/d9c23	Prereg: https://osf.io/52z34	Prereg: https://osf.io/zcusf
Ostracism	Project: https://osf.io/gdf75/	Project: https://osf.io/szmbf/?view_only=312b6136155849a79f3416933a05789b	Project: https://osf.io/ac48m/	Project: https://osf.io/bq8de/?view_only=8d3b69aaf35f4f6c8284337f5e38a3ba	Project: https://osf.io/cjqn7/
	Prereg: https://osf.io/tdypq	Prereg: https://osf.io/w874x	Prereg: https://osf.io/sk43u	Prereg: http://aspredicted.org/blind.php?x=wf558n	Prereg: https://osf.io/qtw5m
Misattribution	Project: https://osf.io/7rqfw/	Project: https://osf.io/f9xp2/	Project: https://osf.io/hfybe/?view_only=37e7d76855184e11a82bc10adf592f90	Project: https://osf.io/yrkdh/?view_only=661a7304e4b84d2b96a46d4c7a3be190	Project: https://osf.io/dmc8p/?view_only=55f9c59766aa4f41b4eb94f020ccebe8
	Prereg: https://osf.io/7y5af	Prereg: https://osf.io/b8tjv	Prereg: https://osf.io/kdx3u	Prereg: https://osf.io/dz2ef	Prereg: https://osf.io/bhe24/

Redemption	Project: https://osf.io/3tckf/	Project: https://osf.io/k3g89/	Project: https://osf.io/6dz87/?view_only=9fe7c043d0004e67865fab9790f3bb2b	Project: https://osf.io/v5tp3/?view_only=30349ce9220f4787966cd4ee4f03b69d	Project: https://osf.io/qnrxb/
	Prereg: https://osf.io/bwpzg	Prereg: https://osf.io/5pqaz	Prereg: https://osf.io/ut5qz	Prereg: https://aspredicted.org/blind.php?x=hg4ix3	Prereg: https://osf.io/39at4
Worse	Project: https://osf.io/zm9nc/	Project: https://osf.io/nexgz/	Project: https://osf.io/gkz7f/?view_only=1ad5d5efbd0742daa81aec677c484f60	Project: https://osf.io/ap8bk/	Project: https://osf.io/qjkry/
	Prereg: http://aspredicted.org/blind.php?x=vz232n	Prereg: https://osf.io/zh4vx	Prereg: http://aspredicted.org/blind.php?x=zg2w6a	Prereg: https://osf.io/69cqm	Prereg: https://osf.io/nzxw3
Misreporting	Project: https://osf.io/3ud4s/	Project: https://osf.io/8n2gh/?view_only=8442e9ad4bb549ae912e7d3efdfd626d	Project: https://osf.io/s52d4	Project: https://osf.io/3rvhc/	Project: https://osf.io/hmfp5/
	Prereg: https://osf.io/2szy7	Prereg: http://aspredicted.org/blind.php?x=3sj645	Prereg: https://osf.io/2szy7	Prereg: https://osf.io/z23qj	Prereg: https://osf.io/7dp4e

Section 4. Additional Results for the Study Described in the Main Text

Power

For all 16 confirmatory tests, the average power was 0.802, with a median of 0.993 and a range of 0.000 to 1.000. The observed replication rate of 90% in the 64 replication attempts was slightly larger than the expected positive result rate based on power estimates. This could occur if some of the confirmatory tests underestimated their true effect sizes, as appeared to be the case for two of the three confirmatory tests that showed null results but had statistically significant treatment effects in the expected direction for most or all of the replication studies. Three of the confirmatory tests produced evidence consistent with the null hypothesis. These were the 'Prediction' study (BF = 5.166), 'Redemption' (BF = 5.981), and 'Misreporting' (BF = 14.673; Bayes Factors produced by using a naive prior and .707 scaling factor and are meant only for exploratory purposes[12]). Aggregating the self-confirmatory test with their own replications for these studies suggests a small but reliable treatment effect in the expected direction for 2 of those 3 null confirmatory tests.

Alternate Metric of Replicability

A common metric to assess replicability is whether the 95%CI between an original study and its replication overlap. This was used in the 100-study Reproducibility Project: Psychology[1], for example. Estimating the 95%CI of the self-confirmatory test ESs, 49/64 (77%) replications ESs fell within those intervals, 15/16 (94%) self-replications, 34/48 (71%) independent replications. Thus, using 95%CI overlap, treating the replication ESs as point estimates with no sampling error, produced smaller replication success rates.

Testing whether effects decline

Confirmation Versus Self-Replication

A first test of unusual possible explanations for declining effect sizes is to compare the effect size of a labs' confirmation studies versus the corresponding self-replications. To test this within-effect decline, we analyzed differences between the self-confirmation study and the self-replication of the same effect by the same lab. A negative average change would be evidence of within-lab decline effects. We predicted that the more studies run between the confirmation study and the self-replication, the greater the decline. Finally, if observation effects contribute to declining effect sizes [S8, S13] we would expect to see larger declines in non-blinded studies than in blinded studies.

First, we calculated differences in standardized effect sizes, pooling effect size estimates across the two half-samples from each replication, taking

$$d_{Sjk} = \frac{1}{2}\left(d_{1R_{jk}jk} + d_{2R_{jk}jk}\right) - \frac{1}{2}(d_{10jk} + d_{20jk})$$

with standard error given by

$$\sigma_{Sjk} = \frac{1}{2}\sqrt{\sigma^2_{1R_{jk}jk} + \sigma^2_{2R_{jk}jk} + \sigma^2_{10jk} + \sigma^2_{20jk}}$$

Let $\tilde{R}_{jk} = R_{jk} - \frac{1}{16}\sum_{j=1}^{4} \quad \sum_{k=1}^{4} \quad R_{jk}$ be the grand-mean centered number of the self-replication. We estimated the parameters of the following meta-regression equation:

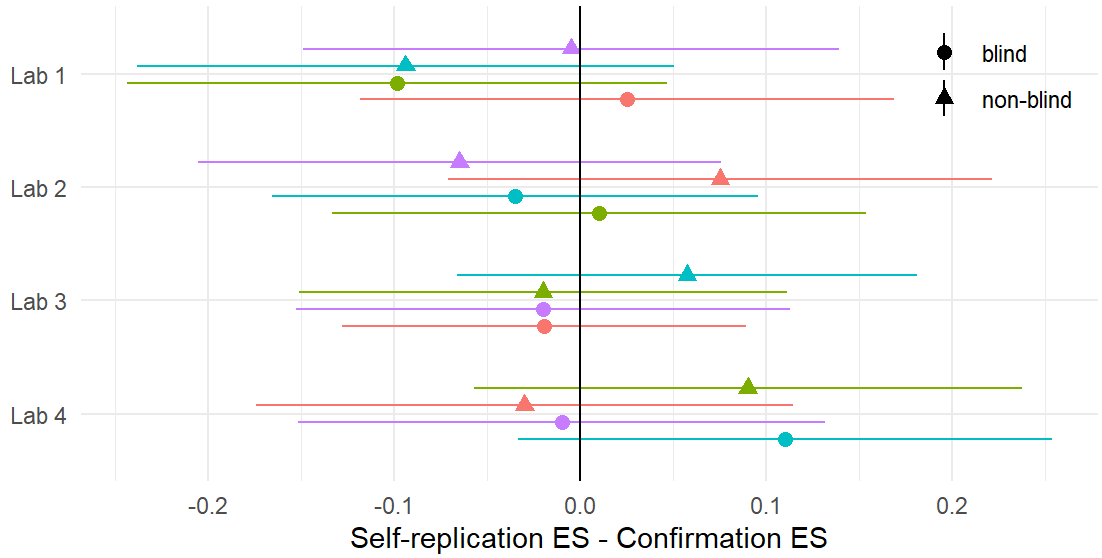
$$d_{Sjk} = \beta_0 + \beta_1\tilde{R}_{jk} + \beta_2 B_{jk} + u_{jk} + e_{jk}$$

where the sampling error term e_{jk} is assumed to have known variance σ^2_{Sjk}. Note that the random effect u_{jk} captures between-study variation in within-lab decline effects. In this model, β_0 represents the average within-study, within-lab difference between self-replication experiment

and confirmation experiment; β_1 represents the average exposure effect, which is the difference in within-study decline effects for self-replication studies conducted after one further intervening replication; and β_2 represents the difference in within-study decline effects between blinded and non-blinded studies. The hypothesis test for $\beta_0 = 0$ used $\alpha = .05$. The tests of β_1 and β_2 were treated as exploratory.

This analysis tested whether effect sizes declined in between a lab's self-confirmation and their own self-replication. This analysis holds all aspects of the lab constant. The only difference is the self-replication was run at a later time in a different group of participants drawn from the same population, with between 0 and 3 replications run in between the two.

In no case was there a statistically significant difference between the magnitude of the confirmation study effect size and the magnitude of the self-replication effect size. On average, self-replication effect sizes were the same size as the confirmation study sizes (d = -0.003, p = .864, 95%CI = -0.035 to 0.030).



Supplementary Figure 1: Difference between a confirmation study (four in total) and the self-replication (four in total) in SD units (shape is mean difference); error bars represent 95%Cis (two-tailed).

There was also no effect of whether the studies were 'blind' (*b* = -0.009, *p* = .760, 95%CI = -0.074 to 0.055) nor the number of replications run in between the two (*b* = -0.008, *p* = .435, 95%CI = -0.030 to .015). As the primary DV is not the replication effect size but the difference between self-confirmation and replications, all predictors are main effects, although the coefficients can be interpreted as though they were interactions had a full model (with replication ES as the DV and confirmation ES as a predictor) been run. Thus, when a lab replicated its own pre-registered study using the same pre-registered procedures, the two tests produced nearly identical effect sizes.

Slope Across Replications

The third test of unusual possible reasons for declining effects looks at the change in effect size over time as replications accumulate, and the interaction of such a decline with blinding. If observer effects cause declines in effect sizes, then we would expect the slope of the

temporal decline to be greater in unblinded studies than in blinded studies, which might show little or no decline.

To examine temporal decline, we first aggregated effect size estimates across the half-samples from each confirmation study and each of the replication studies. Let $d_{\bullet ijk} = \frac{1}{2}\left(d_{1ijk} + d_{2ijk}\right)$, with standard error $\sigma_{\bullet ijk} = \frac{1}{2}\sqrt{\sigma_{1ijk}^2 + \sigma_{2ijk}^2}$. To formally test for time trends across waves, we estimated the following meta-regression model based on the aggregated effect size estimates:

$$d_{\bullet ijk} = \alpha_k + \beta_1(i) + \beta_2 B_{jk} + \beta_3(i)B_{jk} + u_{ijk} + e_{ijk}$$

where α_k is a fixed effect for each lab, representing the average effect size in confirmation studies originating from that lab, β_1 is the average change in effect size for each successive replication study, β_2 is the average difference in effect sizes between blinded studies and unblinded studies, and β_3 represents the difference in slopes between blinded and unblinded studies (i.e., the interaction between the temporal decline and blinding). The model also includes random effects for each confirmation or replication attempt of each study (u_{ijk} for i=0,...,4; j=1,...,4; k=1,...,4). The random effects are allowed to covary within study according to an auto-regressive structure, such that $\text{Var}(u_{ijk}) = \tau^2$, $\text{Cov}(u_{ijk}, u_{i'jk}) = \tau^2 \rho^{|i' - i|}$, and $\text{Cov}(u_{ijk}, u_{i'j'k'}) = 0$ when $j \neq j'$ or $k \neq k'$. The sampling error term e_{ijk} is assumed to have known variance $\sigma_{\bullet ijk}^2$.

We tested the hypothesis $\beta_1 = 0$ to examine temporal decline and $\beta_3 = 0$ to examine whether temporal declines are moderated by blinding. Hypothesis tests for β_1 and β_3 used test-wise alpha levels of $\alpha = .025$ to control the family-wise error rate.

Across replications, there was no consistent change in effect size (b = -0.002, p = .701, 95%CI = -0.015 to 0.01). These results did not change when removing the fixed effect for the lab and were not significantly different for 'blind' and 'not blind' studies (b = 0.02, p = 0.104).

Thus, preregistered and independent replications of preregistered novel findings produced treatment estimates that were stable over replications. Studies that were 'observed' before being replicated did not exhibit a statistically-significantly different slope of change at p = .05 (two-tailed) than those that were kept blind; although given the size of the interaction (p = .104), the power with only eight studies per blind and not-blind conditions was too low to confidently reject the null hypothesis.

Lab-specific variation.

In addition to testing decline, we also examined whether there is lab-specific variation in effect sizes, including variation across originating labs as well as which lab conducted a given replication. Questions about these sources of variation are ancillary to the tests of decline effects, and so we examined them in a separate family of hypothesis tests. We estimated the parameters of the following meta-regression equation:

$$d_{\bullet ijk} = \alpha_k + \gamma_{L_{ijk}} + \beta_1 \quad (i) + \beta_2 B_{jk} + \beta_3 (i) B_{jk} + u_{0jk} + (i) u_{1jk} + e_{ijk}$$

This model elaborates upon the previous model by including fixed effects $\gamma_{L_{ijk}}$ for the lab conducting the replication experiment. In the event of non-convergence, we planned to re-

estimate the model after constraining random effects variance components to zero as necessary to achieve convergence, but this was not necessary.

We then tested two hypotheses pertaining to the originating lab effects and the replication lab effects. First, we tested the hypothesis $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$ to examine whether average effect sizes of the confirmation studies differed across labs. Second, we examined whether average effect sizes varied depending on the lab conducting the replication study by testing the hypothesis $\gamma_1 = \gamma_2 = \gamma_3 = \gamma_4$. We tested these hypotheses using likelihood ratio tests (i.e., using model-based methods, rather than robust variance estimation) because of their greater power. We also conducted corresponding tests based on robust variance estimation methods (i.e., robust Approximate Hotelling's T^2 tests) as sensitivity analyses.

	Likelihood Ratio		Approx. Hotelling's T-squared		
Effects	**Chi-square**	**p-value**	**F statistic**	**Denominator d.f.**	**p-value**
Originating lab effects	2.94	0.40061696	0.65	5.85	0.611535
Replication lab effects	20.18	0.00015557	5.18	12.26	0.015406

As a further sensitivity analysis, we re-estimated the model after removing the occasion predictor, the blinding indicator, and their interaction, as well as simplifying the random effects structure to a study-specific intercept, leaving:

$$d_{\bullet ijk} = \alpha_k + \gamma_{L_{ijk}} + u_{0jk} + e_{ijk}$$

We then repeated the above hypothesis tests under the reduced model.

	Likelihood Ratio		Approx. Hotelling's T-squared		
Effects	**Chi-square**	**p-value**	**F statistic**	**Denominator d.f.**	**p-value**
Originating lab effects	2.61	0.45556	0.70	6.01	0.586202
Replication lab effects	34.82	< 0.0001	3.87	12.31	0.036972

750/750 Split Sample Halves

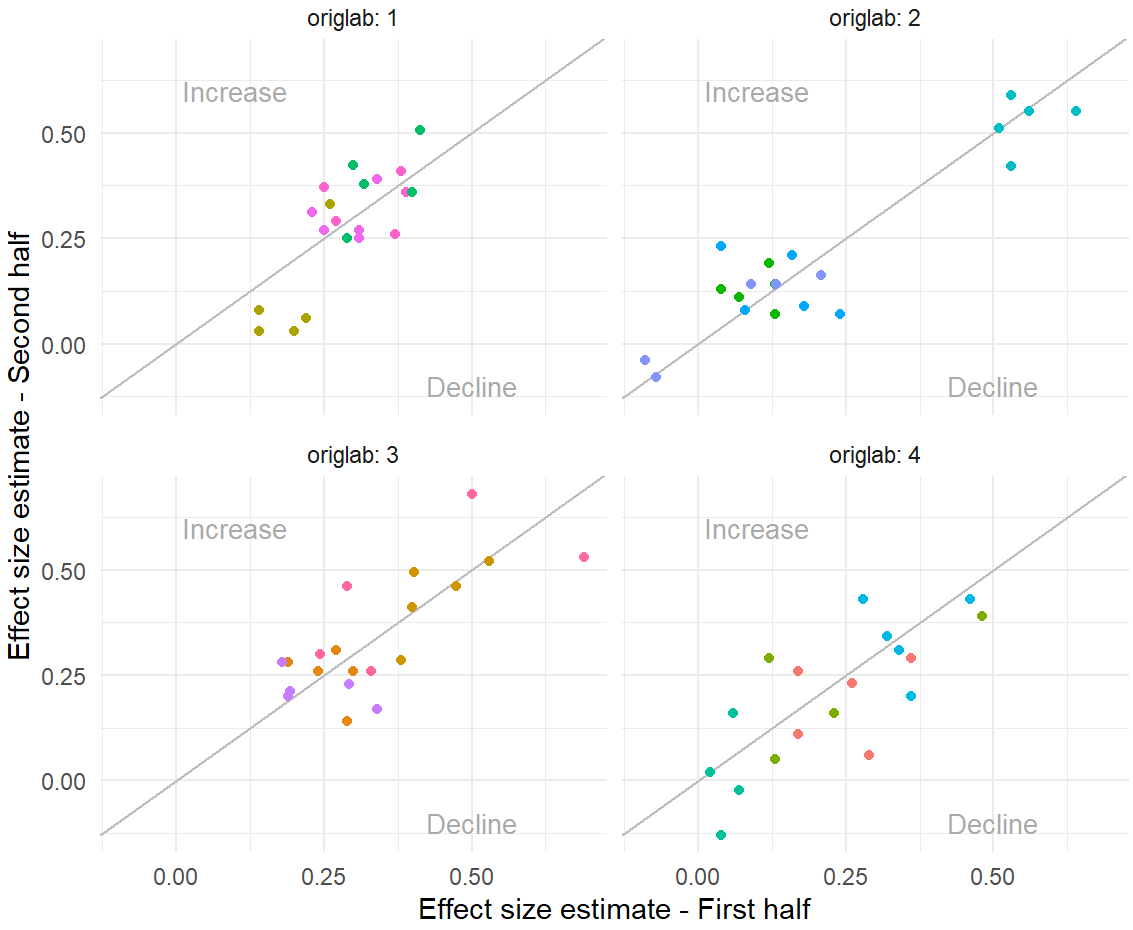
The final test of usual possible reasons for declining effects assessed time-based decline within each experiment, observer effects, and their interaction. We made three predictions:

Time-based decline. Randomly assigning participants to two different half samples allows for a test of the hypothesis that effect sizes of experiments decline over time, with the main difference between the two samples being time of collection. That is, as participants were randomly assigned to the first or second period of data collection, we can test for a causal, time-based decline. We predicted effect sizes to be smaller in the second 750 participants than in the first 750 participants.

Observer effects: analysis order. To test for observer effects, labs were assigned to analyze the first 750 sample or the second 750 sample in a random order. The hypothesis that observation would impact effect sizes leads to the prediction that the 750 that was analyzed first would have a larger effect size than the 750 analyzed second. Including this analysis provides the opportunity to fail to observe evidence for exotic interpretations of declining effect sizes [S8, S13].

Interaction. We included an interaction term in the model of both observer effect order and data collection order.

Supplementary Figure 2 below depicts the effect size estimates from the first and second half-samples of each experiment (including the initial confirmation study and subsequent replications), with separate plots for each of the originating labs. Each study is represented in a different color. If decline occurred, the effect size estimates would tend to fall in the lower triangle of the plot.



Supplementary Figure 2: Effect size of the 1st 750 participants run vs. the effect size observed in the 2nd 750 participants within each lab. Participants were randomly assigned to take the study as part of the 1st or 2nd 750, so this is exogenous temporal variation. Point estimates along the diagonal line correspond to observing the same effect size in both groups of participants.

To formally test these predictions, we estimated the parameters of a meta-regression equation that included terms for the sample half, the order of analysis, and their interaction. Let H_{hijk} = ½ when h=2 and H_{hijk} = -½ when h=1. We estimated the parameters of the following meta-regression equation using the data from both halves of the confirmation and replication experiments from all 16 confirmatory tests:

$$d_{hijk} = \alpha_k + \beta_1 H_{hijk} + \beta_2 A_{hijk} + \beta_3 H_{hijk} A_{hijk} + u_{jk} + v_{ijk} + e_{hijk}$$

where α_k is a fixed effect for each lab, representing the average effect size in studies originating from that lab, β_1 is the average change in effect size from first half to second half of the sample

across experiments (the order of data collection effect), β_2 is the average difference in effect sizes between samples observed first and samples observed second (the order of observation effect), and β_3 represents the difference between the change in effect sizes between experiments where the first half was analyzed first and experiments where the first half was analyzed second (i.e., the interaction between the time effect and the observer effect). The equation also includes random effects for each study (u_{jk}, for j=1,...,4; k=1,...,4) and experiment nested within study (v_{ijk}, for i=0,...,4; j=1,...,4; k=1,...,4). The sampling error term e_{hijk} is assumed to have known variance σ^2_{hijk}. We tested the hypothesis $\beta_1 = 0$ to examine time-based decline, $\beta_2 = 0$ to examine observer effects, and $\beta_3 = 0$ to examine the interaction. Hypothesis tests for β_1 and β_2 used test-wise alpha levels of $\alpha = .025$ to control the family-wise error rate. The test of β_3 was treated as exploratory.

As a specification check for the tests of time-based decline and observer effects, we also estimated these effects using differences in effect sizes between sample halves. The main advantage of modeling the differences in effect sizes is that it requires weaker assumptions than fitting a model for the joint distribution of the effect size estimates.

For the test of time-based decline, let $d_{-ijk} = d_{2ijk} - d_{1ijk}$ denote the decline in effect sizes from the first half sample to the second half sample, with standard error calculated as $\sigma_{-ijk} = \sqrt{\sigma^2_{1ijk} + \sigma^2_{2ijk}}$. Let $A_{-ijk} = (A_{2ijk} - A_{1ijk})$, so that $A_{-ijk} = 1$ if the first half sample was analyzed first and $A_{-ijk} = -1$ if the first half-sample was analyzed second. We estimated the following meta-analytic model:

$$d_{-ijk} = \beta_1 + \beta_2 A_{-ijk} + u_{jk} + v_{ijk} + e_{ijk}$$

where the sampling error term e_{ijk} is assumed to have known variance σ^2_{-ijk}. The meta-regression coefficients have the same interpretation as in the previous model: β_1 is the average change in effect size from the first half to the second half of the sample and β_2 is the average difference in effect sizes between samples observed first and samples observed second. The random effects terms u_{jk} and v_{ijk} now capture study-level and experiment-level variation in the time-based decline, rather than variation in the original effect size estimates.

For the test of observer effects, we used the same approach as above, but based on the difference between effects observed first and those observed second. Let

$$d_{Aijk} = 2(A_{1ijk}d_{1ijk} + A_{2ijk}d_{2ijk})$$

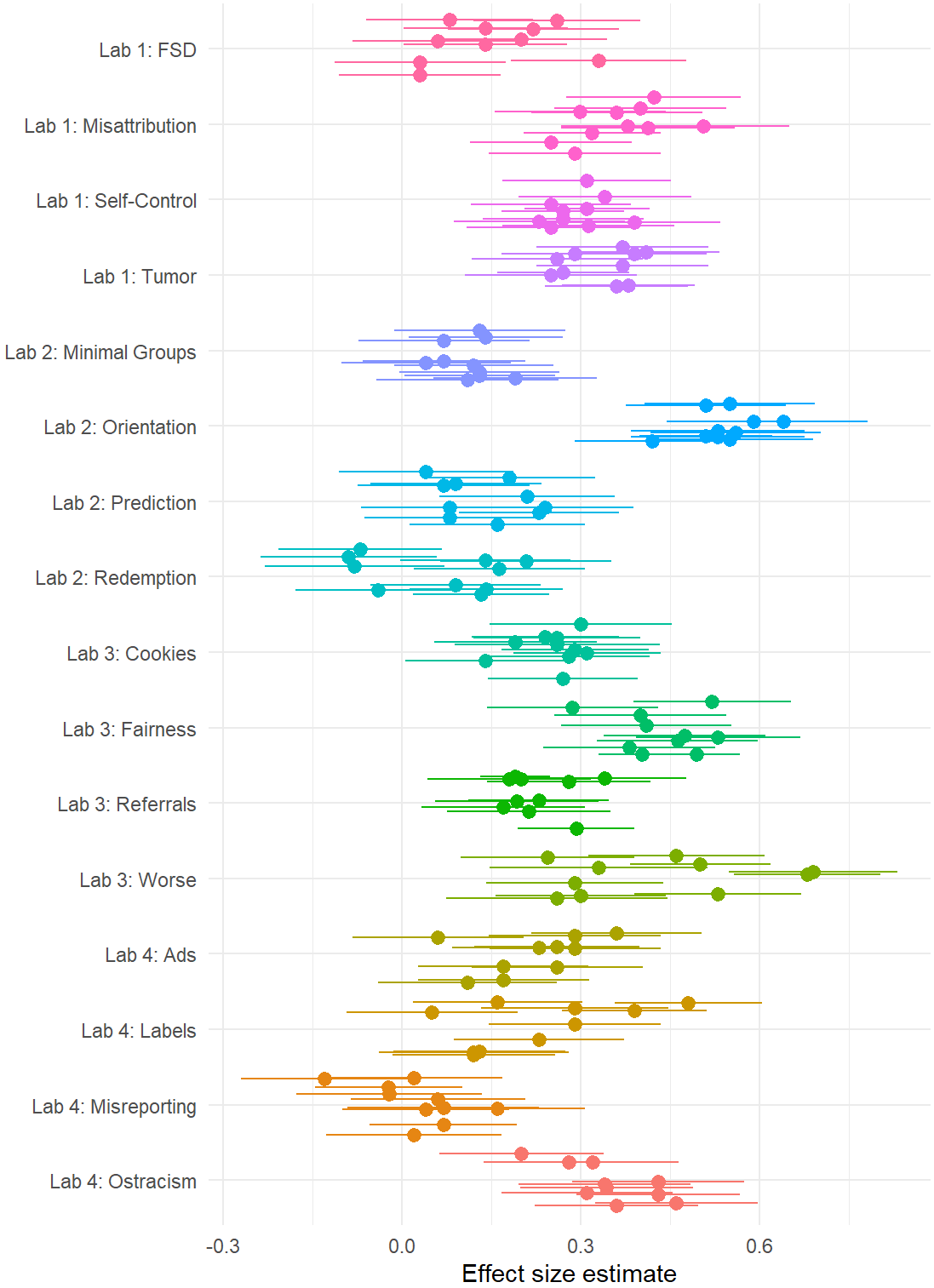
with standard error given by $\sigma_{Aijk} = \sqrt{\sigma^2_{1ijk} + \sigma^2_{2ijk}}$. We estimated the following meta-analytic model:

$$d_{Aijk} = \beta_2 + \beta_1 A_{-ijk} + u_{jk} + v_{ijk} + e_{ijk}$$

where the sampling error term e_{ijk} is now assumed to have known variance σ^2_{Aijk}. The meta-regression coefficients have the same interpretation as in the original model: β_2 is the average difference in effect size from the half-sample analyzed first to the half-sample analyzed second and β_1 is the average change in effect size from the first half to the second half of the sample. The random effects terms u_{jk} and v_{ijk} now capture study-level and experiment-level variation in the observer order effects, rather than variation in the original effect size estimates.

The effect size did not differ depending on whether participants were in the 1st 750 participants or the 2nd 750 participants (b = -0.004, p = .658, 95%CI = -0.023 to 0.015), the order in which the data collected from those two groups were analyzed (b = 0.002, p = .789, 95%CI = -0.016 to -0.021), or the interaction between the two (b = 0.006, p = .867, 95%CI = -0.065 to

0.077). These results were the same when using the difference score between the two 750s of data collection and dropping the non-significant interaction term. Estimating the parameters of the equation with the difference between the 1st and 2nd 750 as the DV showed the same null result (b = 0.002, p = .791, 95%CI = -0.016 to 0.021). The magnitude of effect was the same based on whether the 1st or 2nd 750 was analyzed first (b = -0.003, p = .663, 95%CI = -0.023 to 0.015).



Supplementary Figure 3.: Effect size (circles) and 95%CI from each 750/750 split replication of each of the 16 effects.

Section 5. Prediction Studies of the Novel Discoveries

Study 1: Predictability of Study Outcomes by Practicing Scientists

To gauge how predictable the results of the confirmation studies were for practicing scientists in a range of fields, we conducted a survey of attendees to the Metascience 2019 conference held September 5-8, 2019 at Stanford University. Email invitations were sent out on September 4, 2019 to all who registered, said they would attend the conference, and provided their email address (N = 494). Attendees had a wide range of research experience, from undergraduates to tenured professors, in fields such as Psychology, Medicine, Philosophy, Sociology, Communication, Political Science, Biology, Physics, Statistics, Economics. 68% of the people who answered had a Ph.D. (27% in Psychology, 73% in a different field). Of those who did not have a Ph.D., 41% were enrolled in a graduate program pursuing a Ph.D. in psychology. The survey was approved by the Stanford University Institutional Review Board.

Participants were told: “Here are the descriptions of 15 experiments that were recently run. Each study involved two between-subjects conditions with 750 participants in each condition. After you read each description, you will predict the results that might be observed.” Participants then read a brief description of 12 of the 16 experiments (descriptions of four of the studies were not obtained before the survey invitations were sent). Each account described two conditions (condition A and condition B) to which participants had been randomly assigned. One or two sentences described what participants were asked to do in both experimental conditions. The description of each study was submitted by the lab creating the study and modified by experts in question wording and survey methodology for clarity and completeness. An example, chosen because prediction rates for this study most closely matched the average prediction success across all participants, is:

Condition A: Participants read about an individual being attacked by someone who had low self-control because of a brain injury.

Condition B: Participants read about an individual being attacked by someone who had low self-control because of his/her genes.

DV: Participants decided whether the attacker should be found guilty of assault and battery.

After reading such a description of each study, participants were asked: "Please indicate the percent chance that each of these three possible outcomes will be the result of this study." They were given three options: that the dependent variable was higher in Condition A than Condition B, that the dependent variable was higher in Condition B than in Condition A, and that the dependent variable was not significantly different in the two conditions. These response options were customized to each study. For the example problem described above, the response options were:

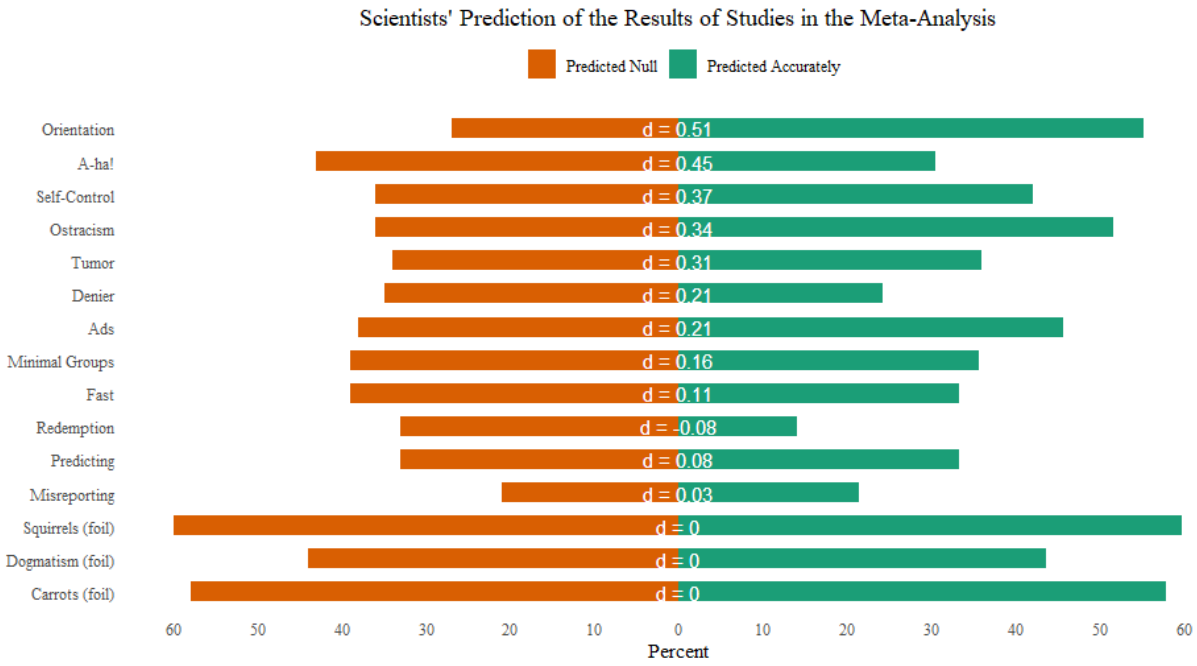
- Significantly more participants in Condition A indicated "guilty" than in Condition B
- Significantly more participants in Condition B indicated "guilty" than in Condition A
- The number of participants in the two conditions who indicated "guilty" was not significantly different from one another

Participants were asked: "Please indicate the percent chance that each of these three possible outcomes will be the result of this study. Please type numbers between zero percent and 100%, and your three answers must add up to 100%."

In addition to the 12 confirmatory test descriptions, three other descriptions were provided, from pilot studies that had been run but yielded non-significant treatment effects. These studies also involved two between-subjects conditions. Data collection ended two days after the link was sent out and the results of the survey were incorporated into a presentation at the conference. This survey was preregistered prior to data collection. Information about the survey is at https://osf.io/3yhbe/, where all study descriptions can be found. Data collection was

stopped on September 8, 2019, before a presentation of the study on which this survey was based. 72 people completed the questionnaire, yielding a response rate of 14.6%.

Results of the survey indicate the 'surprisingness' of the results of the studies in this meta-analysis. Expert scientists were able to predict the correct response only 42% of the time. Interestingly, participants incorrectly predicted individual studies producing a statistically significant effect would not 38% of the time. And, 20% of the time, participants predicted significant treatment effects in the opposite direction of what was actually observed.



Supplementary Figure 4.: Percent of scientists surveyed and asked to predict the results of 12 of the studies in this meta-analysis plus three foils. Bars on the right show participants who predicted the outcome accurately. Bars on the left show participants who predicted there would be no significant difference between the groups. Labels inside the bars are the absolute value of treatment effect sizes from the Confirmatory tests. The Spearman correlation between prediction accuracy and observed effect size was .104, p = .712, two-tailed).

Study 2: Predictability of Study Outcomes Compared with Similar Findings that Showed Low Replicability

To gauge whether the results of the confirmation studies were unusually obvious, we recruited respondents via Prolific and asked them to predict the findings of our confirmation studies that showed high replicability and the findings from a previous large-scale multi-lab replication project in the social-behavioral sciences (referred to as the "comparison effects")[1,3] that showed much lower replicability. If our highly replicable findings are likewise much easier to predict than less replicable findings, then our findings may be unusually "trivial" in comparison to other social-behavioral research. That could mean that our high replicability is more a consequence of choosing obviously true findings rather than introduction of rigor-enhancing practices. Descriptions of the studies were written by independent researchers, blind to our hypothesis, with experience in writing such descriptions of studies to assess lay people's conceptions of research.

Participants answered six prompts, three of which were sampled at random from among the 16 effects identified in the confirmation studies and three of which were sampled at random from among 20 comparison effects. For each prompt, we recorded:

1. the respondent's prediction regarding the sign of the effect;
2. the respondent's rating of their level of understanding from 1 ("Not at all well") to 5 ("Very well");
3. the respondent's confidence rating for their prediction, from 1 ("Not at all sure") to 5 ("Extremely sure");
4. response time (in seconds) to the sign prediction item; and
5. The respondent's prediction regarding whether the effect would replicate (yes or no).

Our primary aim was to test the equivalence between the rate of correct sign predictions for effects in the present project and the rate of correct sign predictions for the comparison

effects. Specifically, we designed the study to the focal null hypothesis that the average predictability of the confirmation study effects exceeds the average predictability of the comparison effects by a threshold of 5 percentage points or greater, using the conventional alpha level of .05. We tested our focal hypothesis by directly estimating predictability rates of each effect, using the results to calculate the average predictability of both sets of effects, and then testing whether the difference between the confirmation study effect and comparison effects is no more than 5 percentage points. For consistency with the (one-sided) equivalence test, we reported 90% confidence intervals.

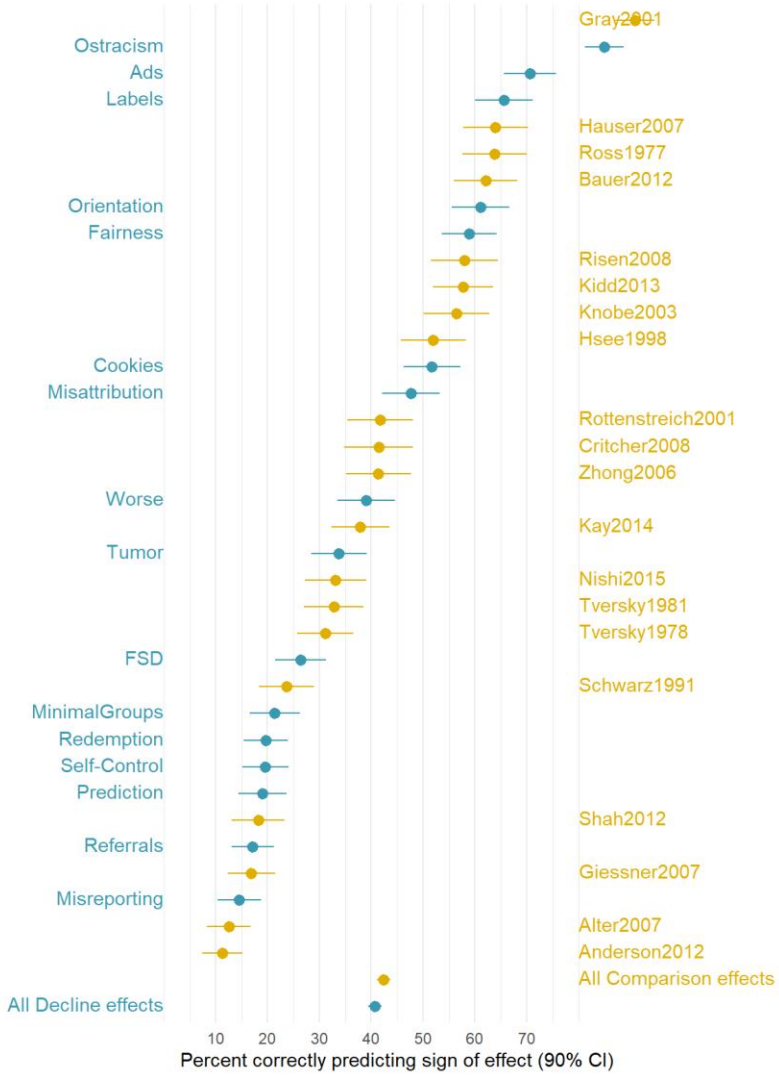
To determine an adequate sample size for the test, we used Monte Carlo simulation to generate data under a normal ogive item response model and conservative assumptions about the variation in participants' prediction abilities and variation in the predictability of each set of prompts. Based on these assumptions and a goal of 95% power if the two sets of studies had identical predictability, we set a target sample size of $N = 1200$ respondents (after excluding any ineligible respondents.

We planned to exclude respondents if they met any of the following criteria: a) reported holding a Ph.D. in Psychology ($N = 14$), b) failed the Captcha item at the end of the survey ($N = 0$), c) indicated at the end of the survey that they did not provide serious responses ($N = 13$ agreeing with "I have just clicked through, please throw my data away."), or d) had an average level of understanding of 2.0 ("slightly") or less ($N = 16$). Additionally, we excluded responses to specific prompts if the respondent indicated the lowest level of understanding ("Not at all well"). This resulted in exclusion of an additional 67 responses from 62 unique respondents.

The survey design and analysis plan were pre-registered prior to data collection beginning. The preregistered analytic plan, survey materials, data, and complete results are

available at https://osf.io/43fn7/. Data collection on Prolific occurred on December 12, 2022. The protocol (# 156-19-0689) was deemed exempt by the Office of Research on Human Subjects (IRB) at the University of California, Santa Barbara.

We obtained a total of 1223 partial or complete responses to the survey. After applying all exclusion criteria, the analytic sample included $N = 1180$ participants. About 55% of respondents held a bachelor's degree or higher degree; 1% were currently enrolled in a doctoral program in Psychology. About 56% reported having taken a college or university course in Psychology.



Supplementary Figure 5. Percent of lay people surveyed and asked to predict the results of 16 of the studies in this meta-analysis plus 20 similar studies with a known replication rate.

Our primary outcome analysis examined whether respondents could correctly predict the sign of the effect from a written description. The figure above depicts the percentage of correct predictions by effect, ordered from most to least predictable. We estimated that the average predictability of confirmation study effects differed from the predictability of the comparison effects by -1.6 percentage points, 90% CI: [-3.4, 0.1]. Under the null of a 5-percentage point difference, this result corresponds to $p<.0001$.

As a secondary outcome, we repeated the analysis using directional confidence scores, calculated by multiplying the self-reported confidence level (shifted by 1) by -1 if the respondent's prediction for sign of the effect was incorrect and +1 if the respondent's prediction was correct. Following the same analytic approach as for the primary outcome, we estimate that the average directional confidence score of confirmation study effects differed from that of the comparison effects by -0.04 points, 90% CI: [-0.12, 0.04]. Participants reported having a very similar level of understanding for the confirmation study effects and the comparison effects. Average response times to the confirmation study prompts were approximately 1 second longer than response times to comparison prompts. In pre-planned sensitivity analyses, we found that all results were robust to the exclusion of 410 item responses from $N = 281$ participants who rated understanding the item "slightly" and to the exclusion of $N = 14$ respondents with a median response time of 4 seconds or less.

One potential concern with this analysis is that the two sets of effects might have similar levels of predictability because participants were inattentive or put little cognitive effort into making predictions. If inattentiveness or low-effort responding were the explanation, we would expect to see little variation across effects in the percentage of respondents who correctly predict the true sign (i.e., all effects would be predicted at near-chance levels). We used an item response theory (IRT) model to investigate this potential alternative explanation by fitting a 1-parameter, normal ogive IRT model to the binary indicator for whether each participant correctly predicted the sign of each effect. This model allows participants to vary in their predictive skill and effects to vary in their predictability. Using REML estimation, we estimated between-participant variance of 0.02 and between-effect variance of 0.37. The between-effect variance was significantly different from null ($\chi^2(1)=1200$, $p<.0001$).

If the effects vary in their degree of predictability, then one might expect that prompts describing larger effect sizes would be more predictable. Using the pooled effect size estimates from multi-lab replications of each finding, we examined whether the absolute magnitude of effect sizes was associated with predictability within the confirmation study effects and the comparison effects. Based on the 1-parameter normal ogive IRT model, clear associations were evident for the confirmation study prompts (β=2.80, p=.003) and comparison study prompts (β=0.66, p=.002). Absolute effect size magnitude explained approximately 35% of the between-effect variance in predictability.

This evidence is consistent with our interpretation that the newly discovered findings in this project that demonstrated high replicability are no more predictable or obvious in advance than prior discoveries that demonstrated low replicability. We believe that the reason for higher replicability is the adoption of rigor-enhancing practices to improve the likelihood of discovering replicable findings, and conducting high-quality replications of those findings. Documentation for this study is available at https://osf.io/43fn7/.

Section 6. Order of Replications

Labs were assigned the order to conduct replications in a Latin square design to equate lab-specific effects across order of replications (see Supplementary Table 3).

Supplementary Table 3: Order of data collection, analysis, and blinding status.

Wave 1	Confirmation	Replication 1	Replication 2	Replication 3	Replication 4
Lab 1	Lab 1 1st 750	Lab 4 1st 750	Lab 2 2nd 750	Lab 3 2nd 750	**Lab 1** 1st 750
Lab 2	Lab 2 2nd 750	Lab 3 1st 750	Lab 4 2nd 750	Lab 1 2nd 750	**Lab 2** 1st 750
Lab 3	Lab 3 1st 750	Lab 2 2nd 750	Lab 1 1st 750	Lab 4 1st 750	**Lab 3** 2nd 750
Lab 4	Lab 4 2nd 750	Lab 1 2nd 750	Lab 3 1st 750	Lab 2 1st 750	**Lab 4** 2nd 750

Wave 2	Confirmation	Replication 1	Replication 2	Replication 3	Replication 4
Lab 1	Lab 1 2nd 750	Lab 3 2nd 750	Lab 4 2nd 750	**Lab 1** 1st 750	Lab 2 1st 750
Lab 2	Lab 2 1st 750	Lab 1 1st 750	Lab 3 1st 750	**Lab 2** 2nd 750	Lab 4 2nd 750
Lab 3	Lab 3 2nd 750	Lab 4 1st 750	Lab 2 1st 750	**Lab 3** 2nd 750	Lab 1 2nd 750
Lab 4	Lab 4 1st 750	Lab 2 2nd 750	Lab 1 2nd 750	**Lab 4** 1st 750	Lab 3 1st 750

Wave 3	Confirmation	Replication 1	Replication 2	Replication 3	Replication 4
Lab 1	Lab 1	**Lab 1**	Lab 3	Lab 2	Lab 4

	1st 750	2nd 750	1st 750	1st 750	2nd 750
Lab 2	Lab 2 2nd 750	**Lab 2** 1st 750	Lab 1 2nd 750	Lab 4 2nd 750	Lab 3 1st 750
Lab 3	Lab 3 1st 750	**Lab 3** 2nd 750	Lab 4 1st 750	Lab 1 1st 750	Lab 2 2nd 750
Lab 4	Lab 4 2nd 750	**Lab 4** 1st 750	Lab 2 2nd 750	Lab 3 2nd 750	Lab 1 1st 750

Wave 4	Confirmation	Replication 1	Replication 2	Replication 3	Replication 4
Lab 1	Lab 1 2nd 750	Lab 2 1st 750	**Lab 1** 1st 750	Lab 4 2nd 750	Lab 3 2nd 750
Lab 2	Lab 2 1st 750	Lab 4 2nd 750	**Lab 2** 2nd 750	Lab 3 1st 750	Lab 1 1st 750
Lab 3	Lab 3 2nd 750	Lab 1 1st 750	**Lab 3** 1st 750	Lab 2 2nd 750	Lab 4 2nd 750
Lab 4	Lab 4 1st 750	Lab 3 2nd 750	**Lab 4** 2nd 750	Lab 1 1st 750	Lab 2 1st 750

Note. Yellow Highlighting corresponds to replications whose results were blinded until the completion of all studies from that cycle. The results of non-highlighted replications were analyzed and reported to the rest of the team as soon as they were completed. The 1st or 2nd 750 refers to which 750 was analyzed first. Bolded text corresponds to self-replication.

Section 7. Additional Procedures

Video Instructions

For confirmation and replication studies: video instructions were put before each study. These videos were short, produced by each individual lab, and merely welcomed participants to the study. This was done to add greater individuality to the experiments and more closely emulate a laboratory procedure in which an experimenter would greet participants.

Blinding

In the event that substantial decline in effect sizes was observed, we planned to compare findings where the outcomes of prior investigations were known versus unknown. Before data collection began in the confirmation stage, each study was designated as either ‘not blind’ or ‘blind’. Studies that were not blind proceeded as any study would. Studies that were ‘blind’ were blinded to prevent the lab from interacting with, analyzing, or observing the dependent measure in their study until all other laboratories had collected their replications. Thus, a ‘blind’ study would be run, the data collected, but the primary DV untouched, not looked at, nor analyzed until all other labs had replicated the work. For any blind study, all replications were also kept ‘blind’ until data collection was completed for the last replication.

Participants and the 750/750 Split

To facilitate examination of unusual reasons for declining effects, had they been observed, we planned to make comparisons within each data collection based on the first and second half of the data collected. Because minimal decline was observed, this plan and analyses are not emphasized in the main text. We retain a full description of this feature of the design here for completeness.

All participants from each confirmation and replication were collected using the labs' individual online sample provider, using the demographic criteria. Online sample providers were given the following instructions:

> To draw the sample to be sufficiently large to include more than enough participants to achieve 1500 completed interviews (and passing attention/quality control checks) within two weeks of inviting all of those people to complete the survey. River sampling and routers were not allowed to be used to obtain participants, as they do not allow selection of the entire potential sample and randomly splitting it in half before data collection begins.
>
> Furthermore, data collections involving a confirmation or replication phase study were collected in two immediate waves of 750 respondents instead of collecting all 1500 participants simultaneously. Hence, after the full sample had been drawn, but before participants were invited to take the study, the drawn sample was divided into two truly random halves. This random assignment was done using random numbers obtained from the Random.org random integer generator (https://www.random.org/integers/?mode=advanced/).
>
> Specifically, the online sample providers were instructed to download random numbers from the generator in batches of 10000, with each integer having a random value between 1 and 10000, using 1 column, decimal numeral system, and having "Generate your own personal randomization right now" checked. One lab drew their own numbers and supplied them to the sample provider; the other three labs instructed their online sample provider to draw the numbers themselves. Each number drawn was appended to one participant in the full sample, until all participants had been assigned

one number each. Participants who were assigned even random numbers were treated as belonging to the sample that was first invited to complete the questionnaire and people who were assigned odd random numbers were treated as belonging to the second sample.

Participants in the first sample were sorted in an ascending order according to the random.org number assigned to each person. Participants in the second sample were sorted in an ascending order according to the random.org number assigned to each person. Beginning with the first person in the sorted list of first sample participants, enough participants were invited so that 750 completed studies, with participants passing the attention/quality control check(s), was finished collecting within two weeks of the first invitation sent.

After 750 participants from the first sample completed the questionnaire and passed the attention/quality control check(s), the second sample was invited using the same procedure to yield 750 completed interviews passing the attention/quality control check(s) by the end of the 14th day after the data collection began. None of the participants in the second sample were allowed to be invited before the first sample had finished collecting and been closed for further collection.

Meta-Study Design

The overall design involved four laboratories, each of which conducted four original confirmatory studies on research topics of interest that came out of discovery-oriented research. Each study involved a two-group, between-subjects manipulation. Multiple outcomes could be assessed, but labs had to designate a single focal outcome. After a lab identified a finding during the discovery phase that was interesting and eligible for inclusion, they conducted a confirmation

study using a new sample of participants. After completing the confirmation studies, each original study was then replicated four times, once by each lab, with order of replications assigned using a Latin square design (see Supplementary Table 3).

Sample Splits

Each confirmation study and replication study was conducted using a target sample of at least 1500 participants, split into two halves. When inviting participants to take part in each survey, participants were randomly assigned to be invited to be a part of the first 750 half sample or the second 750 half sample. Sampling firms frequently collected more data than necessary to meet the 1500 participant target. Some of this was due to how they managed participant invitations and tracking, and some of this was a function of oversampling so that there would still be 1500 participants after applying exclusion rules for failing attention checks.

Data Analysis

Confirmation studies that were not ‘blind’ were analyzed, and a results section was distributed to the other labs within one week of data collection finishing. To match this amount of time among ‘blind’ studies, one week was artificially imposed as a delay after one lab ran their study and the next lab launched their replication.

Blinding

As part of the a priori design to assess unusual reasons for declining effects, had they been observed, each initial confirmation study and each replication study was assigned to either a) analyze the first half-sample and then the second half-sample or b) analyze the second half-sample and then the first half-sample. Confirmation studies were randomly assigned to order of observation, blocking by lab. Replication studies were randomly assigned to order of

observation, blocking by study within labs. If observer effects cause the decline effect, then whichever 750 was analyzed first should yield larger effect sizes than the 750 that was analyzed second.

Confirmation studies and all of their replications were assigned to be either blind or not blind (two per wave, all labs having two blind studies throughout the project). Blinded studies had all data and replications collected before the dependent variable was observed. Thus, there may be an effect of knowing the results of a study or its confirmation that would influence other labs in their data collection, programming, or analysis efforts. If blinded studies showed systematically different effect sizes or different changes in effect size over replications, we would have evidence for causal effects of interacting with data on subsequent replication efforts. For confirmation vs. self-replication tests, we predicted larger declines in non-blinded studies than in blinded studies. This was tested by including a binary term for whether the study was blinded [blind = 1, normal = 0] into the meta-regression. For the change in effect size across 750/750 splits, we predicted blind studies would show a flatter slope than the decline seen in non-blind studies. This was tested by including a binary term for whether the study was blinded [blind = 1, normal = 0] into the meta-regression.

Confirmatory Analyses

Effect Sizes

All study results were transformed into a Cohen's d effect size metric, as all studies involved two between-subjects conditions. In the majority of cases this involved computing the effect size from the means and standard deviations. In cases where regression was used conditioning on other variables, or outcomes were dichotomous, effect sizes were calculated

using alternate formulae to produce an equivalent Cohen's d. This included transforming an unstandardized regression coefficient using *n*s and the SD of the dependent variable for regression outcomes. Binary outcomes were analyzed with a probit regression and the marginal predicted probabilities and standard errors were used to construct the d[6-7].

Coding of Replicability Rate

Studies were coded as 1 if they produced a statistically significant result in the direction expected by theory and 0 if they did not reach statistical significance [no study showed statistically significant effects in the opposite direction].

Lab-specific Variation

We then tested two hypotheses, pertaining to the originating lab effects and the replication lab effects. First, we tested the hypothesis $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$ to examine whether average effect sizes of the confirmation studies differ across labs where α_i is a fixed effect for each lab. Second, we examined whether average effect sizes vary depending on the lab conducting the replication study by testing the hypothesis $\gamma_1 = \gamma_2 = \gamma_3 = \gamma_4$ (where γ_i is a fixed effect for each lab). This was done using likelihood ratio tests (i.e., using model-based methods, rather than robust variance estimation) because of their greater power. We also conducted corresponding tests based on robust variance estimation methods (i.e., robust Approximate Hotelling's T^2 tests) as sensitivity analyses. As a further sensitivity analysis, we re-estimated the model after removing the time variable (0 = confirmatory test/intercept, 1-4 for 1st, 2nd, 3rd, and 4th replication), the blinding indicator, and their interaction, as well as simplifying the random effects structure to a study-specific intercept; we then repeated the above hypothesis tests under the reduced model.

Confirmation versus Self-Replication

We calculated differences in standardized effect sizes, pooling effect size estimates across the two half-samples from each replication, subtracting the confirmation effect size from the self-replication effect size. This way, a negative value corresponds to a decline in effect size. This analysis was done using a random effects meta-analysis with robust standard errors on the difference in the self-replication effect size from the confirmation effect size. We included meta-regression fixed effects for the number of replications run by other teams in between the confirmation and self-replication.

750/750 splits

Confirmatory tests and replication studies had 1500 or more participants each who were themselves assigned to serve as part of either the first or second set of 750 participants. The 750/750 split was included to investigate the stability of effects over time when all other factors are held constant. Companies were instructed to draw from their panels enough participants so that 1500 completed surveys passing attention and/or quality checks could be obtained. These *participants* were then randomly assigned by the companies to be invited to take each confirmation and replication as part of the 1st or 2nd 750 participants. However, based on some characteristics of observed recruiting during data collection, we are not confident that all of the panels effectively implemented this design with random assignment.

This analysis looked at the difference in effect size between the set of participants randomly assigned to be in the 1st 750 group versus the effect size of if they had been randomly assigned to be in the 2nd 750 group. We dummy coded the data to which half group participants were in [1st half = 1, 2nd half = 0]. To examine the effect of analyzing one half of the results

before the other, labs were randomly assigned to analyze the 1st 750 first or second before aggregating the data into a complete N ≳ 1500. Thus, we include a dummy variable in the meta-regression for whether the 1st 750 was analyzed first or second [1st = 1, 2nd = 0]. Although we did not predict an interaction between these two dummy variables, we included it anyway as an exploratory analysis. This model also included a random effect for each study and experiment nested within study.

As a specification check, we also estimated these effects using differences in effect sizes between sample halves. The main advantage of modeling the differences in effect sizes is that it requires weaker assumptions than fitting a model for the joint distribution of the effect size estimates. Thus, we subtracted the effect size from the 1st 750 group from the 2nd 750 group, so negative values of the variable would represent a decline in the magnitude of effect size from 1st to 2nd groups. In this model, the random effects terms now capture study-level and experiment-level variation in the time-based decline effects and variation in the analysis order effects, rather than variation in the original effect size estimates.

Slope Across Replications

The third test looked at the change in effect size over time as replications accumulate for a given study. We further examined whether the change in effect size over time was moderated by whether the study (confirmation and replications) was blinded. If observer effects are the cause, we would expect the slope of the effect sizes to be moderated by whether a study was blind, with change being stronger in non-blind studies and blinded studies showing little or no change.

Thus, for a given study, coding the confirmation as 0 for the y-intercept and each replication in order as 1, 2, 3, and 4; fitting an individual-level growth curve to the change in effect size across replications for a given study would return a slope. Negative slopes would indicate a decline in effect sizes across replications, positive slopes would indicate an incline in effect sizes across replications, and a flat slope would indicate a stable effect.

We then tested the aggregate of these slopes for the 16 high-powered studies. As the order in which the replications were run was randomized via a Latin square design, if one lab consistently showed weaker effect sizes than the other labs, that effect would be spread out across replication orders, neutralizing any statistical leverage on the slope.

In this meta-regression model, we also included a dummy variable for whether the study was blinded. The prediction from this test was that blind studies would show a weaker or no decline in effect sizes while normal studies would show stronger decline[8]. The model also includes random effects for each confirmation or replication attempt of each study. The random effects were allowed to covary within study according to an auto-regressive structure.

Pre-registered Analyses of Declining Effect Sizes

We had three main analyses to test the replicability of new findings as well as testing declining effects. The first was that effect sizes would decline *within* laboratories between confirmatory study and self-replication. The second analysis concerns the 750/750 randomized splits. The third is the nature of effect sizes across laboratories and replications.

General estimation methods

All analyses were conducted using the R statistical computing environment (Version 4.2.1)[9]. All analyses used meta-analytic random effects models estimated using restricted

maximum likelihood with the metafor package (Version 2.1.0)[10]. Standard errors and confidence intervals for all analyses were calculated using cluster-robust standard errors (CR2-type), clustering by study, using the clubSandwich package (Version 0.3.5)[11]. Hypothesis tests were based on Satterthwaite-type small-sample corrections to account for the limited number of independent studies.

Section 8. Supplementary Information References

S1. Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716. https://doi.org/10.1126/science.aac4716

S2. Klein, R., Ratliff, K., Vianello, M., Adams Jr, R., Bahník, S., Bernstein, M., ... & Cemalcilar, Z. (2014). Data from investigating variation in replicability: A "many labs" replication project. *Journal of Open Psychology Data*, 2(1). https://doi.org/10.5334/jopd.ad

S3. Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams Jr, R. B., Alper, S., ... & Batra, R. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, 1(4), 443-490. https://doi.org/10.1177/2515245918810225

S4. Alogna, V. K., Attaya, M. K., Aucoin, P., Bahník, Š., Birch, S., Birt, A. R., ... & Buswell, K. (2014). Registered replication report: Schooler and Engstler-Schooler (1990). *Perspectives on Psychological Science*, 9(5), 556-578.

S5. Ebersole, C. R., Mathur M., et al. (in principle acceptance). Many Labs 5: Testing pre-data collection peer review as an intervention to increase replicability. *Advances in Methods and Practices in Psychological Science*.

S6. Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. SAGE publications, Inc.

S7. Sánchez-Meca, J., Marín-Martínez, F., & Chacón-Moscoso, S. (2003). Effect-size indices for dichotomized outcomes in meta-analysis. *Psychological methods*, 8(4), 448.

S8. Protzko, J., & Schooler, J. W. (2017). Decline effects: Types, mechanisms, and personal reflections. *Psychological science under scrutiny: Recent challenges and proposed solutions*, 85-107. https://doi.org/10.18637/jss.v036.i03

S9. R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

S10. Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3), 1-48. URL: http://www.jstatsoft.org/v36/i03/

S11. Pustejovsky, J. E. (2019). clubSandwich: Cluster-Robust (Sandwich) Variance Estimators with Small-Sample Corrections. R package version 0.3.5. https://github.com/jepusto/clubSandwich

S12. Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic bulletin & review*, 16(2), 225-237.

S13. Schooler, J. (2011). Unpublished results hide the decline effect: Some effects diminish when tests are repeated. *Nature*, *470*(7335), 437-438. https://doi.org/10.1038/470437a