**Proposed Overarching Analyses for Decline Effect**

***Analyses***

Three basic analyses will be run testing the presence of the decline effect. (1) The first analysis will test whether the effects statistically significantly increase or decrease depending on whether the effects belonged to the first or the second 750 half samples. (2) The second analysis will test whether the effect sizes of the originating lab’s self-replication study is statistically larger or smaller than the originating lab’s confirmation study. (3) The third analysis will test whether effects statistically significantly decrease or increase across all four waves of data collection (all 16 studies with all 5 confirmations and replications).

**We will compute the standardized mean difference between the two treatment groups, Cohen’s d, for each study. This will be the common effect size metric. Studies using a binary dependent variable will have the effect size coded from the marginal predicted probabilities and standard errors (or standard deviations). All effect sizes will be re-coded so that the predicted direction from the confirmation studies is positive. We do not yet know all of the analysis strategies used, so as a first stage we will align all of the effect sizes prior to running any analyses.**

***750/750 splits: Analysis***

When inviting participants to take part in each survey, participants will be randomly assigned to be invited to be a part of the first 750 half sample or the second 750 half sample. Randomly assigning participants to two different half samples will allow for the 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 will be randomly assigned to which half of the data they were collected in, it will allow for a causal test of time-based decline effects.

Randomization to the first or second sample will be accomplished by using random numbers obtained from the Random.org random integer generator (<https://www.random.org/integers/?mode=advanced/>).

Specifically, the survey firm, or the lab, will download random numbers from the random.org integer generator in batches of 10,000, with each integer having a random value between 1 and 10,000, using 1 column, decimal numeral system, and having “Generate your own personal randomization right now” checked.

Each number drawn was appended to one respondent in the full sample, until all respondents had been assigned one number each. Respondents who were assigned even random numbers were treated as belonging to the sample that will be invited first to complete the questionnaire (i.e., the first 750) and people who were assigned odd random numbers will be treated as belonging to the sample that will be invited after the first half sample has finished collecting (i.e., the second 750).

Respondents in the first 750 sample will then be sorted in an ascending order according to the random.org number assigned to each person.

Respondents in the second 750 sample will also be 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 750 sample respondents, enough respondents will be invited so that 750 completed interviews, with respondents passing the attention check(s), will be finished collecting within two weeks of the first invitation sent.

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

**We predict 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 will be randomly assigned to analyze the first 750 or the second 750 sample first. If observer effects cause the decline effect, then whichever 750 was analyzed first should yield  statistically significantly larger effect size estimates than the 750 that was analyzed second.

**We predict that the 750 that was analyzed first will have a larger effect size than the 750 analyzed second.**

Interaction

The general decline effects and observer-caused decline effects hypotheses are not mutually exclusive. Thus, we will also included an interaction term in the model of both observer effect order and data collection order.

Additional model terms

We will include fixed effects for both study and originating lab. IDs for studies were coded based on what study the effect sizes came from (e.g. a confirmation and all of it’s replications coded with the same number). Originating lab will also be coded for each study and treated as a fixed effect. Due to variation among labs in terms of implementation of the 750/750 splits, as well as fidelity of the labs’ survey companies, we predict that such an effect would be necessary.

**The analysis plan for the 750/750 splits will be the following:**

Observations will be the difference in effect sizes from each of the 750s, the 1st 750 ES subtracted from the 2nd 750 ES. This would mean that the constant would correspond to the decline in effect size from the 1st 750 to the 2nd. There will be an ‘firstanalyzed1st’ column that will be coded 1 if the 1st 750 was analyzed first and 0 if the 2nd 750 was analyzed first. This effect would correspond to an observer effect where analyzing the data in a different order changed the outcome. There will be a between-replications column coding for what study the effect sizes came from (named ‘experiment’) and a column named ‘labid’ with between-replications varying codes for what lab collected the study. There will also be a column for which team created the study, following the same lab codes as in the labid column ‘labelled ‘origin’. There will also be a column for the meta-analytic standard errors calculated by the following formula: sediff = sqrt(se\_1^2 + se\_2^2)

**Table 1.** Example of spreadsheet for analyzing the 750/750 splits.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Study Name | ES (Cohen's d) difference | firstanalyzed1st | Lab | experiment | labid | origin | sediff |
| Unemployment-  Confirmation | -0.2 | 1 | Stanford | 1 | 1 | 1 | .064 |
| Unemployment  Self-Replication | -0.3 | 0 | Stanford | 1 | 1 | 1 | .048 |
| Unemployment-  Replication1 | 0.7 | 0 | UVA | 1 | 2 | 1 | .039 |
| Unemployment-  Replication 2 | 2.5 | 1 | UCSB | 1 | 3 | 1 | .004 |
| Unemployment-  Replication 3 | -0.03 | 0 | UCB | 1 | 4 | 1 | .06 |
| Tumor - Confirmation | 1.5 | 1 | UCSB | 2 | 3 | 3 | .01 |
| Tumor - Self-replication | -1.2 | 0 | UCSB | 2 | 3 | 3 | .024 |

Thus, there will be a lab effect; a study-specific fixed effect; an experiment-specific fixed effect, and  is the sampling error with mean 0 and known standard deviation. As the first 750 ES will always be subtracted from the second 750 ES, the constant will be a measure of decline from the 1st 750 ES collected to the 2nd 750 ES collected.

**Analysis 2**

*Confirmation versus Self-Replication*

A second test of the decline effect that holds constant lab effects will be to compare the effect size of a labs confirmation versus its own self-replication.

Decline Effect

Data will again be the difference between effect sizes, with the confirmation ES subtracted from the self-replication ES. We will code a ‘time’ variable for whether the effect size came from the confirmation (=0) or the self-replication (=1). This will be a within-study test of decline, predicting a negative coefficient to indicate that a decline has happened between the confirmation and the self-replication. Data will be coded where by subtracting the original effect size from the self-replication effect, corresponding to a difference in effect size.

Exposure mechanism

There will also be a column for the number of studies that were run in between the confirmation and the self-replication. This will be a number between 0 and 3 to indicate the number of studies that had been run in between the confirmation and the self-replication. According to one theory of the decline effect, the decline is caused by a study being repeatedly run. Thus, we predict that the more studies that had been run, the greater the decline effect.

Blinding

A study that was blind was collected but the data was not seen or analyzed until all replications had been run. A decline effect could emerge for a number of reasons. If observer effects are driving the decline effect, we would expect to see larger declines in non-blinded studies than in blinded studies.

**The analysis plan for the Confirmation vs. Self-replication will be the following:**

Observations will be the confirmation effect size subtracted from the self-replication effect size, so there will be one column for the difference in effect sizes with negative signs corresponding to a decline effect. There will be a ‘numreps’ column that will be coded 0, 1, 2, or 3 based on the number of replications that were run in between the confirmation and the replication. There will be a ‘blind’ column for studies that were blind. Then there will be a between-replications column coding ‘id’ for what study the effect sizes came from (named ‘study’) and a column for the meta-analytic standard errors (see Table 2). Meta-analytic standard errors will be calculated as follows; SE(replication-confirmation) = √(SE\_replicaiton² + SE\_confirmation²).

**Table 2.** Example of spreadsheet for analyzing the Confirmation versus Self-replication effect sizes

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Study Name | ES (Cohen's d) difference | numreps | blind | Lab | id | se |
| Unemployment | -0.2 | 1 | 0 | Stanford | 1 | .064 |
| Tumor | 0.7 | 3 | 0 | UCSB | 2 | .039 |
| Cookie | 2.5 | 3 | 1 | Berkeley | 3 | .004 |
| Label | -0.5 | 3 | 1 | Stanford | 4 | .04 |
| Minimal Groups | 0.8 | 3 | 0 | UVA | 5 | .038 |

Thus, there will be a within-study effect of time (selfrep), two between-studies effect of the number of reps between confirmation and self-replication (numreps) and a blind dummy code for blinded studies (‘blind’). There will also be interactions between numreps, selfrep and blind, and a three-way interaction of self-rep and numreps and blind.

**Analysis 3**

*Slope of all Replications: Base Model*

The third test of the decline effect tests the average slope of the confirmation through replications for each study. This looks at the change of effect size over time as replications accumulate for a given study.

Decline Effect

Data will again be in long format, with five effect sizes per study (confirmation and four replications). We will use a level-and-shape model coding the confirmation study as 0 and the final replication as 1. This will be a within-study test of decline, predicting a negative coefficient to indicate that a decline has happened between the confirmation across all replications

Blinding

A study that was blind was collected but the data was not seen or analyzed until all replications had been run. A decline effect could emerge for a number of reasons. If observer effects are driving the decline effect, we would expect to see larger declines in non-blinded studies than in blinded studies. Observer effects may be the cause of the decline effect. If these are indeed operational, then we expect the slope to be moderated by whether a study was blind, with decline effects being stronger in non-blind studies and blinded studies to show little to no decline.

Lab

All labs will have the following codes: 1 = UCSB, 2 = UVA, 3 = Berkeley, 4 = Stanford. For the analysis we will code not only what lab ran which replication, but also which lab created the study. This is to take into account the fact that different labs may have different average effect sizes (different intercepts). We predict that the average slope will be the same for all labs (all labs experience the decline effect equally, all else equal). We also wish to test whether individual labs consistently get different effect sizes from the other labs.

**The analysis plan for the Decline Effect slope analysis will be the following:**

Observations will be the effect sizes from the confirmation and all four replications, reshaped into long format so there will be one column for the effect sizes (each study will contribute five rows per data collection, that is., the confirmation and the replications). There will be a ‘replication’ column that will be coded 0 for the confirmation and 1, 2, 3, 4 for the replications in order that they were assigned to be collected. There will be a ‘blind’ column for studies that were blind (1=blind study, 0=non-blind). Then there will be a between-replications column coding for what study the effect sizes came from (named ‘studyid’) and a column for the meta-analytic standard errors (see Table 3).

**Table 3.** Example of spreadsheet for analyzing the Confirmation versus Self-replication effect sizes

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Study Name | Effect size (Cohen's d) | replication | blind | labid | studyid | origlab | se |
| Label -  Confirmation | 0.2 | 0 | 1 | 4 | 1 | 4 | .064 |
| Label -  UCSB Replication | 0.3 | 1 | 1 | 1 | 1 | 4 | .048 |
| Label -  Berkeley Replication | 0.7 | 2 | 1 | 3 | 1 | 4 | .039 |
| Label - UVA Replication | .1 | 3 | 1 | 2 | 1 | 4 | .031 |
| Label -  Stanford Replication | .04 | 4 | 1 | 4 | 1 | 4 | .004 |
| Tumor -  Confirmation | 0.01 | 0 | 0 | 1 | 2 | 1 | .78 |
| Tumor -  Stanford  Replication | -0.03 | 1 | 0 | 4 | 2 | 1 | .6 |
| Tumor -  UVA  Replication | -0.5 | 2 | 0 | 2 | 2 | 1 | .04 |
| Tumor -  Berkeley  Replication | 0.8 | 3 | 0 | 3 | 2 | 1 | .038 |
| Tumor -  UCSB  Replication | 0.6 | 4 | 0 | 1 | 2 | 1 | .041 |

Thus, there will be two within-study effects, one of time (replication) and one of replicating lab. There will be two between-studies effects, one of the originating lab and a blind dummy code for blinded studies (‘blind’). There will also be interactions between blind and time. The analysis of change will be a level-and-shape model, meta-analytically weighted in a meta-regression with robust variance estimation (Hedges et al., 2010). Intercepts and slopes will be allowed to vary with a slope by intercept covariance. The intercept will be regressed on the originating lab (dummy coded as 3 separate variables with UCSB as the reference variable).