Proposed Overarching Analyses for the Decline Effect

# Study design

The overall design of the Decline Effects study involved four laboratories, each of which conducted four original experimental studies on research topics of interest. Each study had to involve a two-group, between-subjects manipulation. Multiple outcomes could be assessed, but labs had to designate a single focal outcome. For each original study, the same lab 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.

## Sample splits

Each confirmation study and replication study was conducted using a sample of 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.

Randomization to the first or second sample was 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 downloaded 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 were then sorted in an ascending order according to the random.org number assigned to each person. Respondents in the second 750 sample were also 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 were invited so that 750 completed interviews, with respondents passing the attention check(s), were finished collecting within two weeks of the first invitation sent.

After 750 respondents from the first 750 sample had completed the questionnaire and passed the attention check(s), the second 750 sample were 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 were allowed to be invited before the first 750 sample has finished collecting and had been closed for further collection.

## Observer effects

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 lab. 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.

## Blinding

Each original study was randomly assigned (blocking by lab) to a blinded or non-blinded condition. For blinded studies, data were collected but no analysis was conducted until all replications were completed. For non-blinded studies, data were analyzed upon completion of data collection, which could occur before further replication studies by other labs.

# 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).

## Data preparation and notation

We will compute the standardized mean difference between the two treatment groups, Cohen’s *d*, for each original study and each half-sample from the confirmation and replication studies. 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.

We will use the following notation to describe the planned analyses.

* Let denote the effect size estimate and let be the corresponding standard error from original study from originating lab .
* Let be the effect size estimate and let be the corresponding standard error, both from half of experiment in study from originating lab . Take for the original confirmation experiment and i = 1,…,4 for the subsequent replications in chronological order.
* Let be a predictor variable equal to ½ if half of experiment from study from originating lab was analyzed second and equal to -½ if it was analyzed first.
* Let be an predictor variable equal to ½ if half study from originating lab was blinded and equal to -½ if it was non-blinded.
* Let be the index of the lab that conducts replication of study *j* from originating lab *k*, for , , and . Thus if for self-replications, where replication of study from originating lab is conducted by the same lab that originated the study. Let be the index of the self-replication study, i.e., the replication for which .

## General estimation methods

All analyses will be conducted using the R statistical computing environment. In all of the meta-analytic models described below, we will estimate random effects using restricted maximum likelihood with the metafor package. In the event of non-convergence, variance components will be constrained to zero and then the model will be re-estimated. Standard errors for overall average effect sizes and for meta-regression coefficients will be calculated using cluster-robust standard errors (CR2-type), clustering by study, using the clubSandwich package. Hypothesis tests will be based on Satterthwaite-type small-sample corrections to account for the limited number of independent studies.

## Analysis 1: 750/750 split sample halves

This analysis involves testing time-based decline effects within each experiment, observer effects, and their interaction. We make three predictions:

### Time-based decline effects. 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. **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 were randomly assigned to analyze the first 750 or the second 750 sample first. **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 include an interaction term in the model of both observer effect order and data collection order.

To test these predictions, we will estimate a meta-regression model that includes terms for the sample half, the order of analysis, and their interaction. Let when and when . We will estimate the following meta-regression model based on the data from both halves of the confirmation and replication experiments from all 16 studies:

where is a fixed effect for each lab, representing the average effect size in studies originating from that lab, is the average change in effect size from first half to second half of the sample across experiments (the order of data collection effect), is the average difference in effect sizes between samples observed first and samples observed second (the order of observation effect), and 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 model also includes random effects for each study (, for ) and experiment nested within study (, for ). The sampling error term is assumed to have known variance . We will test the hypothesis to examine time-based decline effects, to examine observer effects, and = 0 to examine the interaction. Hypothesis tests for and will use test-wise alpha levels of to control the family-wise error rate. The test of will be treated as exploratory as there is no directional hypothesis.

As a specification check for the tests of time-based decline effects and observer effects, we will also estimate 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 effects, let denote the decline in effect sizes from the first half sample to the second half sample, with standard error calculated as . Let , so that ½ if the first half sample was analyzed first and ½ if the first half-sample was analyzed second. We will estimate the following meta-analytic model:

where the sampling error term is assumed to have known variance . The meta-regression coefficients have the same interpretation as in the previous model: is the average change in effect size from the first half to the second half of the sample and is the interaction between time-based decline effects and observer effects. Note that the random effects terms and now capture study-level and experiment-level variation in the *time-based decline effects*, rather than variation in the original effect size estimates.

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

with standard error given by . We will estimate the following meta-analytic mode:

where the sampling error term is now assumed to have known variance . The meta-regression coefficients have the same interpretation as in the original model: is the average difference in effect size from the half-sample analyzed first to the half-sample analyzed second and is the interaction between time-based decline effects and observer effects. The random effects terms and now capture study-level and experiment-level variation in the *observer order effects*, rather than variation in the original effect size estimates.

## Analysis 2: Confirmation versus Self-Replication

A second test of the decline effect will be to compare the effect size of a labs’ confirmation studies versus the corresponding self-replications. To test this within-study decline effect, we will analyze differences between the initial confirmation study and the same replication of the same effect by the same lab. A negative average change will be taken as evidence of within-lab decline effects. According to one theory of the decline effect, the decline is caused by a study being repeatedly run (i.e., an exposure effect). Thus, we predict that the more studies run between the confirmation study and the self-replication, the greater will be the decline effect. Finally, we would expect to see larger declines in non-blinded studies than in blinded studies.

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

with standard error given by

Let  be the grand-mean centered number of the self-replication. We will then estimate the following meta-regression model:

where the sampling error term is assumed to have known variance . Note that the random effect captures between-study variation in within-lab decline effects. In this model, represents the average within-study, within-lab difference between self-replication experiment and confirmation experiment; 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 represents the difference in within-study decline effects between blinded and non-blinded studies. There is no interaction term because the sample of studies would not be large to support testing one. The hypothesis test for will use of as it is the critical test of within-lab decline effects.

## Analysis 3: Slope across replications

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

To examine temporal decline effects, we will first aggregate effect size estimates across the half-samples from each confirmation study and each of the replication studies. Thus, let , with standard error . Using these aggregated effect size estimates, we will estimate the following meta-regression model:

where is a fixed effect for each lab, representing the average effect size in confirmation studies originating from that lab, is the average change in effect size for each successive replication study, is the average difference in effect sizes between blinded studies and unblinded studies, and represents the difference in slopes between blinded and unblinded studies (i.e., the interaction between the temporal decline effect and blinding). The model also includes random effects for each study intercept ( and study-specific slopes for the temporal decline effects (, both for . The random effects will be allowed to covary. The sampling error term is assumed to have known variance .

We will test the hypothesis to examine temporal decline effects and = 0 to examine whether temporal declines are moderated by blinding. Hypothesis tests for and will use test-wise alpha levels of to control the family-wise error rate. We will not test because it is not relevant to the theory of decline effects but include it in the model for proper estimation of the interaction term.

*Exploratory Anslysis*

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

This model elaborates upon the previous model by including fixed effects for the lab conducting each replication experiment. In the event of non-convergence, we will re-estimate the model after constraining random effects variance components to zero as necessary to achieve convergence. We will then test two hypotheses, pertaining to the originating lab effects and the replication lab effects. First, we will test the hypothesis to examine whether average effect sizes of the confirmation studies differ across labs. Second, we will examine whether average effect sizes vary depending on the lab conducting the replication study by testing the hypothesis . We will test these hypotheses using likelihood ratio tests (i.e., using model-based methods, rather than robust variance estimation) because of their greater power. We will report corresponding tests based on robust variance estimation methods (i.e., robust Approximate Hotelling’s tests) as sensitivity analyses. As a further sensitivity analysis, we will re-estimate the model after removing the occasion predictor, the blinding indicator, and their interaction, as well as the random slopes term, leaving:

We will then repeat the above hypothesis tests under the reduced model.