Topics for Papers from the Decline Effect

Prepared by: John Protzko, Matt Berent, Jon Krosnick

**Core paper (3 pre-registered tests)**

*Is there a Decline Effect?*

The overall design of the Decline Effect 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).

**Additional Tests of the Decline Effect**

In addition to the main analysis using the three tests of the decline effect, a number of additional papers can be written testing the decline effect in additional ways. Below we highlight some of those studies and subsequent papers.

*Does the decline effect occur just before running a confirmatory test?*

In the decline effect project, labs ran pilot studies testing effects as many times as they wished before nominating that study to be replicated by other teams. These *pilot* studies were run either as studies after a replication using the extra time given or as independent data collections. When conducted as independent data collections, *pilot* studies usually included fewer participants than the 1,500 replication studies. One cause for decline effect in published studies might be that the first studies are underpowered from using too few participants. To test this, we will compare the effect size of the first study ran as part of the project (the confirmation study) to the pilot study that was run just before it (called “C-1”). It may be the case that as part of the nominating of a study from C-1 to confirmation the decline effect occurred.

Data will be the difference between effect sizes, with the ES from the last *pilot* run before the confirmation study subtracted from the confirmation study ES. We will code a ‘time’ variable for whether the effect size came from the *pilot* (=0) or the confirmation (=1). This will be a within-study test of decline, predicting a negative coefficient to indicate that a decline has happened between the *pilot* and the confirmation.

*Does the number of times a pilot study was tested before confirmation cause a decline effect?*

The central theory of the decline effect is that the number of times a study is conducted causes a decrease in the magnitude of effect each time. This is not the only possibility for decline effects, but was the central conceptualization during the decline effect project. The project imposed no limitations for how many times a treatment effect was allowed to be piloted before it could be nominated for replication. Hence, the number of pilot data collections varies for each study. We will test this conception of the decline effect by testing the change from C-1 to confirmation as a function of the number of times the study was piloted prior to confirmation. The theory of the decline effect would predict that the more times a study was tested, the smaller the change from C-1 to confirmation. This occurs because the bulk of the decline had already occurred during the numerous pilots before C-1.

*Does excluding the confirmation study alter the findings of the slope of the decline effect?*

In test #3 of the decline effect, each replication was run in complete counterbalanced order. What was not counterbalanced, however, was the placement of the confirmation study. In all waves of data collection, the originating lab tested their own confirmation first, followed by the replications. This was done to re-create what is normally done in science, where a lab comes up with an idea, tests it, and then other labs may try to replicate it. Thus, by removing the confirmation study from the slopes analysis of effect sizes over time, we may come to a different understanding of the decline effect than when including the confirmation studies.

To test this, we will re-run the slopes analysis with just the replications, excluding the confirmation study.

*Does including the pilot studies to the slopes analysis alter the decline effect?*

It may be the case that the decline effect occurred during the pilot phase to a greater extent than it did during the replication phase of the project. This would be consistent with theories of the decline effect that posit the right conditions for discovery only in the first stages of inquiry, with subsequent studies showing weaker effects due to the misalignment of ideal conditions. Thus, we will run a separate analysis of the slopes analysis of the decline effect including the pilot studies for each confirmation that was run.

**Investigations of Additional Effects Using Study Data**

The richness of data collected to test the Decline Effects affords numerous opportunities

to test additional hypotheses about empirical research.

*What are the effects of sample size on p-hacking?*

Researchers may increase the probability of observing significant treatment effects through a variety of seemingly justifiable practices. These practices include, but are not limited to: (1) Eliminating outliers from data on the grounds that outliers disproportionately influence statistical results; (2) Introducing control variables into analyses on the grounds that the controls eliminate variance that masks true population effects; and (3) Eliminating respondents who fail to meet some criteria, such as a manipulation or attention check, on the grounds that such individuals do not adequately represent the population about which inference is desired. The proposed study will examine the extend to which sample size moderates the effect these different activities have on the probability of observing significant effects when analyses conducted in the absence of these activities generated no significant effects.

For these analyses, subsamples will be created by randomly assigning participants from the total sample to the subsamples. The total sample will be divided into as many subsamples of 13 participants as possible, then again into as many subsamples of 25 participants as possible. Additional divisions of the total sample will be conducted to create subsamples of 50 participants, 100 participants, and 200 participants.

Several models will be tested within each subsample. One model will test effect of the treatment on a study’s main dependent. Another will test the same model after eliminating outliers. A third will test the effect after eliminating all participants who failed the attention check. A fourth will test the effect after eliminating respondents who fail the manipulation check. The final model will test the effect after controlling for as many demographic variables as possible.

The results of the model tests will be entered as data in an additional analysis. Each row of data will include the effect size for the treatment effect on the dependent variable, the size of the sample used to test the model, and dummy variables to identify if the model (1) eliminated outliers, (2) eliminated respondents who failed manipulation checks, (3) eliminated respondents who failed attention checks, (4) included demographic control variables. The model tested will estimate the interaction effects of sample size and the four dummy variables on effect sizes after controlling for all main effects.

*Hypothesis 1*: The effect of eliminating outliers on effect size will be larger for small samples than large samples.

*Hypothesis 2*: The effect of eliminating respondents who fail manipulation checks on effect size will be larger for small samples than large samples.

*Hypothesis 3*: The effect of eliminating respondents who fail attention checks will be larger for small samples than large samples.

*Hypothesis 4*: The effect of introducing demographic controls on effect size will be larger for small samples than large samples.

*What are the observed effects of sample size on significance testing?*

The probability of observing an effect in a sample is influenced by the size of the sample. As sample size increases, the probability of observing a “real effect” (i.e., an effect that exists in the population) increases. Presumably, sample size is unrelated to the probability of observing an effect in a sample that does not exist in the population. However, decreasing sample size may increase the probability that any sample data violates assumptions on which statistical probability tests are based. Many statistical tests assume normally distributed data, and violations of this assumption are more likely in smaller samples from populations in which data are not normally distributed. As violations of statistical assumptions increase, the chances of observing effects in a sample that do not exist in the population may increase. This suggests that: (1) Effects that exist in the population are more likely to be observed in large samples than small samples; and, (2) Effects that do not exist in the population are more likely to be observed in small samples than large samples.

A series of analyses will test the effects of sample size on the probability of observing effects in research. For these analyses, each study’s total sample will be divided into subsamples of equal size.

Some subsamples will be created by segmenting the total sample based on the temporal sequence in which participants completed the study. For example, one subgroup will include the first 200 participants who completed the study (participants 1 through 200), another will include the next 200 participants who completed the study (participants 201 through 400), the next will include participants 401through 600, and so on until the maximum number of subsamples with 200 participants has been created. This process will be repeated for subsamples of 100 participants, 50 participants, 25 participants, and 13 participants.

Additional subsamples will be created by randomly assigning participants from the total sample to the subsamples. As in the previous method, the total sample will be divided into as many subsamples of 13 participants as possible, then again into as many subsamples of 25 participants as possible. Additional divisions of the total sample will be conducted to create subsamples of 50 participants, 100 participants, and 200 participants.

For each subsample, statistical models will estimate the effect of a treatment variable on each demographic available with the data. The percent of all models of equal size and subsampling method that generated a significant effect on a demographic variable will be used as a dependent variable in a subsequent analysis. In the subsequent analysis, subsample size and subsampling method will be used to predict percent of significant effects.

An additional set of analyses will be limited to those pilot studies that produced no significant treatment effects. These analyses will test for the effects of the treatment variable on the on the dependent variable using subsamples. The percent of all models of equal size and subsampling method that generated a significant treatment effect will be entered in a model in which they are predicted by sample size and subsampling method.

A final set of analyses will be limited to those studies that produced a significant treatment effect. These analyses will test for the effects of the treatment variable on the on the dependent variable using subsamples of the same size and subsampling method. The percent of all models of equal size and subsampling method that generated a significant treatment effect will be entered in a model in which they are predicted by sample size and subsampling method.

*Hypothesis 1*: Samples with less than 30 participants will generate a higher percent of significant treatment effects on demographics than samples with more than 30 participants.

*Hypothesis 2*: Among studies that found no significant effect among the total sample, samples with less than 30 participants will generate a higher percent of significant treatment effects on dependent variables than samples with more than 30 participants.

*Hypothesis 3*: Among studies that found significant effects among the total sample, samples with less than 30 participants will generate a lower percent of significant treatment effects on dependent variables than samples with more than 30 participants.

*What is the reproducibility rate of moderators and mediators?*

Moderation, the discovery of conditional causal effects where the effect of a treatment on a variable is different for different people, is a central element of scientific investigation. In the psychological sciences, it is most likely that every single treatment effect is in some way conditional on individual differences of people. The investigation of moderation, however, is often performed post-hoc after omnibus average treatment effects are not found. In 100 replications of some of the most impactful psychological research, 46% of main effects but only 22% of interactions replicated (OSC, 2015). Furthermore, interactions carry additional assumptions, including linearity and common support across the interaction space that are often untested; estimates of the reproducibility when taking these assumptions into account has shown a reproducibility rate of interactions of approximately 10% in some fields (Hainmuller et al., 2018). Each of these investigations, however, only had a single replication of a previous study. Thus, we are poised to provide a more powerful test of the reproducibility rate of moderation by looking at the reproducibility of moderations from confirmation studies through the four replications within the studies in the Decline Effect. This will provide a more powerful test of the reproducibility of moderation than has been seen in previous investigations. We will test this by exploring moderations in confirmation studies. Then, we will test to see whether those moderations can replicate in the four subsequent replications of the confirmatory study.

*What is the reproducibility rate of mediation?*

Several of the replication studies included mediator variables. While mediator variable data were collected in each study, the primary analyses of the decline effect do not include the mediator data. Rather, tests of the decline effect have focused exclusively on the effects of treatment variables on dependent variables. The proposed analyses will examine the consistency with additional variables mediate the treatment effects on dependent variables across studies.

These analyses will use data for all studies in which the confirmation study tested whether or not a variable mediated the treatment effect on the dependent variable. Every mediation analysis conducted by the group that the produced the confirmation study will be replicated separately for each replication study produced by the other groups. The difference between the treatment effect on the dependent variable for models without the moderator variable and models with the moderator variable will be used as data in a subsequent analysis. The subsequent analysis will compare the average difference for confirmation studies to each of the replication studies.

*Hypothesis 1*: The average difference between models with and without the mediator variable will be the same for the confirmation study and the replication studies.

*Does restricting studies to participants who passed an attention check alter the findings?*

Attention checks are a common and integral part to social research, especially experimental research. Attention checks can take the form of simple ‘check this box’-style questions, to more complicated questions about the study contents (e.g. what did you just read). The most common practice is to remove participants who do not pass the attention checks, decreasing overall sample size. Little is actually known, however, to what extent attention checks alter the results of experiments. Thus, we propose to test the effect of removing participants who fail attention checks. This will be done with be a meta-analysis of the interactions of treatment effect moderated by whether participants passed the attention check on the individual DVs from each comprising study. A table of preliminary results from one lab can be seen in Table 1.

Table 1.*Predicting the main dependent variable with the treatment, succeeding on the attention check, and the interaction between them.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Replicating lab | Wave and experiment | *b*treatment\*pass | *SEb* | Pass  attention check % |
| Stanford | Wave 1 - Labels - Pilot | 0.22\*\*\* | 0.07 | 85.68% |
| Stanford | Wave 1 - Labels - Confirmation | 0.11 | 0.07 | 87.47% |
| Stanford | Wave 1 - Labels - Self-replication | 0.10 | 0.08 | 89.44% |
| Stanford | Wave 1 - Cookies (UCB attention check) | Results not shared yet | - | - |
| Stanford | Wave 1 - Cookies (Stanford attention check) | Results not shared yet | - | - |
| Stanford | Wave 1 – Tumor | Results not shared yet | - | - |
| Stanford | Wave 1 - Minimal Groups | -0.01 | 0.03 | 93.35% |
| Stanford | Wave 2 - Ads - Pilot | 0.01 | 0.04 | 85.22% |
| Stanford | Wave 2 - Ads - Confirmation | 0.09+ | 0.05 | 89.00% |
| Stanford | Wave 2 - Ads - Self-replication | 0.02 | 0.05 | 88.85% |
| Stanford | Wave 3 - Ostracism - Pilot | 0.03 | 0.03 | 89.97% |
| Stanford | Wave 3 - Ostracism - Confirmation | 0.02 | 0.04 | 91.19% |
| *Note:*Only participants who completed the questionnaire and completed only once are included. | | | | |
| + *p* < 0.1, \* *p* < 0.05, \*\* *p* < 0.01, \*\*\* *p* < 0.001 | |  |  |  |

*What is the reproducibility of different variable selection algorithms for covariate selection across replications?*

A number of advanced estimation procedures have been invented in recent years whose purpose is to increase model fit by taking a large number of potential covarying variables and break them down to a subset of ‘influential’ variables. These procedures have been as mundane as backwards, forward, and stepwise selection procedures in SPSS to advanced procedures such as LASSO estimation and other ridge estimation properties, Random Forests, Bayes Additive Regression Trees, and others.

These estimators have been tested in isolation for recapturing causal effects in large simulated datasets, but not for covariate selection on reproduced real-world data. Thus, the Decline Effect project and dataset will be useful for testing variable selection algorithms against one another in not only primary data but also for predicting which covariates would be influential in replications.

This is crucially important because not all studies can be easily replicated. Some studies use information involving historical changes (such as the broadcast of Sesame Street, Starvation effects during World War II, Nationwide Income Tax experiments, etc.) that cannot be reproduced. This is the precise place knowing which variable selection algorithms, if any, can select the reproducible covariates in a study when reproducibility cannot be tested.

Thus, for every study that included covariate information as part of its data collection, we will first run a meta-analysis of all covariate effects across the confirmation and four replications to determine which covariates, replicably, ‘matter’. Then, we will test a number of variable selection algorithms first in the confirmation data, then in each of the replications, to see which ones can reproduce the ‘right’ covariates.