Stat215B Final Project Proposal

John Haberstroh, John Semerdijan, Lindsey Osimiri

April 1st, 2016

1 Introduction

The Open Science Collaboration (OSC) published a landmark study in Science this past August which detailed the ability of a consortium of scientists to reproduce prior published work. This study was motivated by the commonly-acknowledged, yet insufficiently-characterized, lack of reproducibility in basic science research. The study done by the OSC is large-scale, spans multiple journals and research areas within psychology, and provides all of its data as open source for future collaboration and critique. In this study, they saw that the percentage of successful replication of studies fell between 35 and 47%, which is much smaller than the expected 78.5-89% replication as determined by the published p-values and effect sizes. A recent critique of these results, published in Science in March 2016, argues that there are many errors in the statistical analysis done in this study which, if accounted for, would greatly increase the replication efficiency. In our final project, we seek to investigate the claims made by both the OSC and their critics, and decide which side of the debate we support.

2 Previous Work

In this study, the OSC sampled experiments from articles published in 2008 in three highly-regarded psychology journals. The articles were chosen simply by which were published first in each journal; over 100 articles were chosen in this manner. Each replication team chose only one experiment from their chosen study to replicate. For each experiment, a single result was chosen as the de facto estimator of reproducibility before the replication experiment was carried out. This key result was a joint decision between each replication team and the original authors of each study, which enabled replication teams to pick a representative aspect of the experiment.

The OSC's aggregate analysis of the key results generated in the replication experiments addressed three key issues:

1. whether the replication experiments produce expected results

2. whether additional data would change the identified results of the original experiments

3. whether there are factors strongly correlated with reproducibility

To address the first issue, the authors compared p-values and effect sizes of the original and replication experiments. To address the second issue, the authors conducted a meta-analysis on a dataset consisting of the results from both the original and replication experiments. Finally, for the last issue, the authors selected six indicators of the original study (original P value, original effect size, original sample size, importance of the effect, surprising effect, and experience and expertise of original team) and seven indicators of the replication study (replication P value, replication effect size, replication power based on original effect size, replication sample size, challenge of conducting replication, experience and expertise of replication team, and self-assessed quality of replication), and performed correlations between these indicators and the estimates of reproducibility. In sum, the authors found that a far smaller percentage of experiments than expected had reproducible effects as judged by P value significance (35/100 vs 89/100) and effect size (47.4 vs 78.5). Additionally, they found that the meta-analyses resulted in many (32) of the key results having a CI that included 0. Finally, the correlation of indicators showed that some were strongly correlated to reproducibility, but none would be an accurate predictor.

3 Proposed Work

3.1 Error

The first criticism of the Open Science Collaboration (OSC) study by critic Gilbert et al. was that their results were skewed by a failure of some participating labs to faithfully replicate the original experiments. To demonstrate their point, they compare the OSC to a similar study, the Many Labs Project (MLP). As a matter of realistic practice, neither MLP nor OSC experimenters were able to properly follow the protocol of every study under question. However, the MLP and the OSC approached the problem of replication differently: while the MLP asked 36 labs to independently replicate 16 original psychology studies, the OSC asked 100 labs to independently replicate one of 100 different studies. With a standard rejection p-value of .05, the OSC claim that they expect a failure to replicate in only 5% of the replications. Unfortunately, these p-values are not externally valid for the reproduced statistics on dissimilar populations. On the other hand, the MLP setup allows us to separate the effect of error due to both sampling from the error due to an unacceptable protocol ("infidelity"). Gilbert et al. conclude that the "infidelities" allowed by OSC participants unfairly decreased the replication rate.

We intend to reproduce the analysis of both the OSC and the Gilbert et al. publications. Fortunately, the OSC authors provide both code and data online. To control for external validity, we would like to look at the subset of studies that properly followed the sampling strategy proposed by the original 100 researchers, thereby removing studies with "infidelities", and compare their results to those of MLP. By controlling for well-executed studies, we now only have to consider the impact of sampling error between replications.

3.2 Power

Statistical power is the second major criticism of the OSC replication study by Gilbert et al.. Only 47% of the OSC replications were successful while the MLP group was able to replicate 85% of the 16 studies which used pooled data from 36 labs. Gilbert et al. argue that if the MLP researchers had used the OSC approach – to replicate each study only once, stead of 36 times – they would have replicated only 34%.

We plan to verify the above conclusion by Gilbert et al. and also determine which replication method is appropriate given the pivotal question about the state of reproducibility across psychology research. In practical terms, given some limited resources, is it more important to reproduce the results of landmark papers – i.e. the focus of MLP – or is it more critical to sample randomly across psychology research studies and assess their validity through a single replication as in the OSC? We believe these strategies answer slightly different questions and we will attempt to determine which is most faithful to the question of replication.

We will also apply shrinkage to the relevant parameters in the MLP studies, since each study results in 36 estimates. We can draw comparisons between the original parameter estimates, the MLE estimates from the 36 replications, and the shrunken values.

3.3 Bias

Another issue alleged by critics is that there is bias in the estimate of the replication ability due to the assumption that infidelities in study design could be treated as an exogenous source of noise. In the critics' analysis of one covariate, whether the replication study was approved by the original authors, they find that approved studies were four times more likely to replicate than studies that were not approved. Only 69% of the replication studies were approved, so the critics argue that not controlling for this difference unfairly biases OSC's estimate of reproducibility towards being much less than it should be. In our analysis of this critique, we will first verify that the critics' analysis is correct, and that endorsed studies do have a much higher replication probability. Additionally, we will seek to find other covariates that may have biased the results that were not accounted for in the original analysis, like the difference in self-reported expertise between the replication team and the original team. The OSC recorded dozens of covariates that may relate to observed results, but only showed the effect of 13 of these covariates on reproducibility. We will examine the other covariates more thoroughly.