Pre-Registration of: Predicting Many Labs Replications from Original Paper Statistics

Date: 2015-07-10

Researcher:
Erika Salomon
PhD Candidate
University of Illinois at Urbana-Champaign
ecsalomon@gmail.com
@ecsalomon

Background & Rationale

Psychologists have started using metrics based on test statistics or p-values (e.g., p-curve[1], R-index[2]) to make claims about the evidential value, robustness, or plausibility of published research findings. At the same time, large-scale replication projects (e.g., Many Labs 1[3], Registered Replication Reports at *Perspectives on Psychological Science*[4], and a special issue of *Social Psychology*[5]) have begun directly testing the replicability of psychological research.

This study examines whether, and to what degree, the test-statistic-based metrics predict replication results. Is it possible to infer from the test statistics in a paper whether (or how well) its results will replicate?

Methods and Materials

This project will investigate whether (and how well) test-statistics-based summaries of individual papers predict outcomes of replications of studies in those papers. Study 1 will be based on those papers whose studies were replicated in Many Labs 1[6] and 3[7].

Predictor Variables

Papers whose studies are replicated in Many Labs 1 and Many Labs 3 will be summarized using the following metrics:

- ***p*-curve[8]:** the Z-scores for the tests of evidential value and inadequate evidential value

[1]Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). *P*-Curve: A Key to the File-Drawer. *Journal of Experimental Psychology: General, 143*, 534–547. http://doi.org/10.1037/a0033242

[2]Schimmack, U. (2014, Dec 13). The R-Index for 18 Multiple Study Articles in *Science* (Francis et al., 2014). *Replication Index* [Blog]. Retrieved from http://web.archive.org/web/20150710004211/https://replicationindex.wordpress.com/2014/12/13/the-r-index-for-18-multiple-study-articles-in-science-francis-et-al-2014/

[3]Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., et al. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology, 45*, 142–152. http://doi.org/10.1027/1864-9335/a000178

[4]Association for Psychological Science. (2014). *Registered Replication Reports*. Retrieved from http://web.archive.org/web/20150710005507/http://www.psychologicalscience.org/index.php/replication

[5]See the editors' introduction here: Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology, 45*, 137–141. http://dx.doi.org/10.1027/1864-9335/a000192

[6]Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., et al. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology, 45*, 142–152. http://doi.org/10.1027/1864-9335/a000178

[7]Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulsborstad, H. M., Allen, J. M., Banks, J. B., et al. (2015). Many Labs 3: Evaluating participant pool quality across the academic semester via replication (Version 1). Retrieved from https://osf.io/s59bg/

[8]Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). *P*-Curve: A Key to the File-Drawer. *Journal of Experimental Psychology: General, 143*, 534–547. http://doi.org/10.1037/a0033242

- ***R*-index**[9]**:** the R-index metric
- **Test of insufficient variance (TIVA)**[10]**:** The variance estimate
- **Correlation between effect size and sample size:** The correlation coefficient

All of these values will be estimated using the p-checker tool[11], which accepts as input the original test statistic and degrees of freedom for a given statistical test (e.g., $t(31) = 2.06$). Data will be collected for the test of the *primary* hypothesis for each study in a given paper. Based on recommendations from the p-checker app and the p-curve guide[12], the following guidelines will apply:

- For each study, select the critical test of the focal hypothesis for that study.
- For interactions:
  - **Attenuation** (effect is smaller under one level of the moderator): Use test of highest order interaction or difference in linear trends
  - **Sign Change** (effect reverses direction under one level of the moderator): Use the tests of the simple effects (2x2), the tests of lower order interactions (2x2x2, where attenuation reverses under second moderator), or separate linear trends, NOT the interaction
- 3-cell designs:
  - **High/medium/low:** linear trend
  - **Treatment vs 2 different controls:** Treatment vs. control 1 contrast, Treatment vs. control 2 contrast - BUT NOTE THE THIRD ONE
  - **2 Treatments vs 1 control:** Treatment 1 vs. control contrast, Treatment 2 vs. control contrast - BUT NOTE THE THIRD ONE
- Do not use inexactly reported test statistics (e.g., $F < 1$)

Outcome Variables

Replication results will be operationalized in two ways:

1. As a **dichtomous** outcome, indicating whether the average Many Labs effect was significantly different from 0 (coded as `1`) or not (coded as `0`)
2. As a **continuous** outcome, indicating the *difference* in results from the original study. All original and replication test statistics will be converted to Cohen's ds using the p-checker app. The difference in Cohen's ds between the replication and original results will be taken, with **negative** values indicating that the replication effect was *smaller* than the original and **positive** values indicating that the replication effect was larger than the original.

Analysis Plan

Primary Analyses

Because the predictor variables are all transformations of the same data (test statistics), I expect them to be correlated and thus to introduce heteroscedasticity when included in a model together. Thus, the primary analyses will be:

[9] Schimmack, U. (2014, Dec 13). The R-Index for 18 Multiple Study Articles in *Science* (Francis et al., 2014). *Replication Index* [Blog]. Retrieved from http://web.archive.org/web/20150710004211/https://replicationindex.wordpress.com/2014/12/13/the-r-index-for-18-multiple-study-articles-in-science-francis-et-al-2014/

[10] Schimmack, U. (2014, Dec 30). The Test of Insufficient Variance (TIVA): A New Tool for the Detection of Questionable Research Practices. *Replication Index* [Blog]. Retrieved from http://web.archive.org/web/20150710041437/https://replicationindex.wordpress.com/tag/test-of-insufficient-variance/

[11] Schönbrot, F. (2014). P-checker: The one-for-all p-value analyzer [Software]. Retrieved from http://shinyapps.org/apps/p-checker/

[12] Simonsohn, U., Nelson, L., Simmons, J. (2015, March 2). *Official User's Guide to the* P-*Curve*. Retrieved from http://www.p-curve.com/guide.pdf

1. Logistic regression models predicting the dichotomous outcome (sgnificant replication or not) from each of the predictor variables considered individually
2. Pearson correlation coeffcients of the relationship between each of the predictor variables and the continuous outcome (difference in Cohen's *d*s between original and replication)

In addition, most of the Studies in Many Labs 1 significantly replicated, whereas most of the studies in Many Labs 3 did not significantly replicate. Thus, there will be little variation within each of these sets on the dichotomous outcome. For this reason, **the two sets of replications will be pooled for each analysis**.

Four of the effects included in Many Labs 1 are from a single paper. To give this paper equal weight to the others, it will be entered only once into each model. Because all of the effects significantly replicated, its value on the dichotomous outcome measure will be recorded as `1`. For the continuous measure, the *average* of the differences in Cohen's *d*s for the four effects will be taken.

Additional Analyses

Several additional analyses will be performed to explore the data set. These include but are not limited to:

- Estimating correlations among the predictor variables
- Regressing the outcomes from multiple predictor variables
- Analyzing the data separately for each Many Labs project
- Alternative ways of treating Many Labs outcomes from a single paper