Pre-registered study protocol

Title: An assessment of analytic reproducibility at Psychological Science

Registration submitted by: Tom E. Hardwicke[a,*] & Michael C. Frank[b]

[a]Meta-Research Innovation Center at Stanford (METRICS), Stanford University, USA;
[b]Department of Psychology, Stanford University, USA

*tom.hardwicke@stanford.edu

Date submitted: October 17th 2017

Status: Data collection has not yet begun

Study rationale and aims

In response to growing concerns about the credibility of the psychological literature (Pashler & Wagenmakers, 2012), a number of concerted efforts have attempted to assess the reproducibility of published studies (e.g., Klein et al., 2014; Open Science Collaboration, 2015, Hagger & Chatzisarantis, 2016). In these replication projects, "reproducibility" is defined as the ability to obtain similar findings to a previous study after repeating the methods as closely as possible with a new sample of participants (i.e., *replicability*; Asendorpf et al., 2013). By contrast, the term "reproducibility" can also refer to the more fundamental ability to obtain exactly the same study findings by performing the same analysis procedures (i.e., *analytic reproducibility*; Stodden, 2015; Goodman et al., 2016). A recent National Science Foundation report determined that analytic reproducibility should be considered a "minimum necessary condition for a finding to be believable and informative" (Bollen et al., 2015).

Replication efforts are a vital component of the scientific enterprise, sifting signal from noise, and contributing to the progressive accumulation of evidence over time (Simons, 2014). However, typically these studies put the cart before the horse by attempting to verify a higher level in the reproducibility hierarchy (replicability) before a more fundamental level has been established (analytic reproducibility). If a study's findings cannot be reproduced using the same data and the same analyses, then for the sake of efficiency this should preclude any more resource-intensive attempt to replicate those findings in a new sample.

Unfortunately, the poor availability of raw research data in the field of psychology (e.g., Wicherts et al., 2006; Vanpaemel et al., 2015) complicates the routine assessment of analytic reproducibility. In a recent study, our team (Hardwicke et al., *in preparation*) capitalized on the introduction of a mandatory open data policy at the journal *Cognition* to identify studies with available and (in principle) reusable data that could be used in an assessment of analytic reproducibility. Focusing on only a subset of relatively straightforward analyses (ANOVAs, *t*-tests, correlations etc), reported in 35 articles, we encountered an initial reproducibility failure rate of 58%[1].

In the present study, we intend to investigate whether these findings extend to a corpus of articles enrolled in the "Open Science Badges" initiative introduced at the journal Psychological Science. In contrast to Cognition's mandatory open data policy, the Open Science Badges scheme involves authors voluntarily agreeing to share data, after which their article is adorned with an "Open Data Badge" to signal adoption of this open practice. Under both policies, an article's raw data should theoretically be deposited in an accessible online repository, or with the article's supplementary materials, enabling an independent third party to assess analytic reproducibility. The submission guidelines[2] of Psychological Science specifically state that authors can earn an:

"Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported result."

The guidelines also refer to sharing analysis scripts, although it is unclear if this is mandatory or merely encouraged:

"This includes annotated copies of the code or syntax used for all exploratory and principal analyses."

If they agree to participate, authors are asked to explicitly disclose[3] whether there is:

---

[1] Assessment of 11 of the 35 articles is still in progress and they are not included in this estimate. Additionally, the failure rate fell to 29% after assistance from the original authors was obtained.

[2] Retrieved (9th October, 2017) from: https://www.psychologicalscience.org/publications/psychological_science/ps-submissions#OPEN

[3] The submission guidelines say "The criteria for earning badges and the process by which they are awarded, along with answers to frequently asked questions, are described in the Open Practices document." The quote is taken from this document, retrieved (9th October, 2017) from: https://osf.io/tvyxz/wiki/2.%20Awarding%20Badges/

"…sufficient information for an independent researcher to reproduce the reported results."

In summary, if the Open Badges Scheme is having its intended impact, the success rate of analytic reproducibility should be 100%. Our study aims to evaluate this empirically. Specifically: does open data shared under the Open Sciences Badges scheme actually enable successful analytic reproducibility?

Sampling frame

We will focus on a sample of articles identified in a previous study evaluating the impact of the Open Science Badges initiative between January 2014 and May 2015 (Kidwell et al., 2016). Kidwell et al., identified 47 articles that received an open badge, 35 of which had *in principle* reusable data[4] (i.e., accessible, correct, complete, and understandable data).

Of these 35 articles, one team member (TEH) filtered out those which did not meet the criteria for a 'substantive finding supported by a straightforward analysis' (n = 9; see definition below), and those for which the data files could no longer be obtained (n = 1; contra Kidwell et al.). This resulted in a sample of 25 eligible articles. Despite being referred to in the author guidelines (see above), we found that only 6 of the eligible articles shared analysis scripts. Excluded article id codes and reason for exclusion are outlined in Appendix A.

Our operational definition of a 'substantive finding supported by a straightforward analysis' was adopted directly from Hardwicke et al. Specifically:

"*Reproducibility checks will be restricted to articles reporting behavioural data. For each article reporting behavioural data, [we] will attempt to identify a coherent subset of analyses that support one of the article's primary outcomes. The definition of a primary outcome will be to some extent subjective, but where possible we will select outcomes that are referred to in the article's abstract and or supported by a figure or table. For any given primary outcome, both descriptive and inferential statistics will be targeted for the reproducibility check. Only articles for which the primary outcome(s) is based on a relatively straightforward quantitative analysis will be considered (see below). The first eligible analysis reported in the article will be selected.*

[4] Thus, taking these articles into account, the overall analytic reproducibility success rate is already below 100%.

For the purposes of this investigation, a 'relatively straightforward' analysis is one that might typically be included in an introductory-level statistics textbook aimed at undergraduate psychology students (e.g., Field et al., 2012) and could therefore be comfortably performed by a competent graduate-level (or higher) psychology student or researcher. Only quantitative analyses will be considered. Examples include, mean, median, standard deviation, confidence intervals, standardized effect sizes, correlation, t-tests, and ANOVAs."

Design

This is a retrospective one-shot case study design.

Measured variables

After Hardwicke et al., we will leave a reasonable margin of error before considering any mismatch between reported outcomes and reproducibility check outcomes as a major error.

Specifically, we will use the following error classification scheme:

- **INSUFFICIENT INFORMATION ERROR**
- **MINOR NUMERICAL ERROR**
- **MAJOR NUMERICAL ERROR**
- **DECISION ERROR**

An *Insufficient Information Error* refers to any situation where the team cannot proceed with the reproducibility check because of ambiguous or absent information in the published article or associated files. We will attempt to resolve these issues by contacting the original authors.

When numerical discrepancies are detected, we will calculate the *percentage error* (PE) as below:

$$PE = \frac{|obtained - reported|}{reported} \text{x } 100$$

Based on the degree of PE, we will classify numerical errors as either:

- *Minor Numerical Error*: 0%> PE <10%
- *Major Numerical Error*: PE >=10%

Finally, when the reported *p* value falls on opposite side of the .05 boundary relative to the obtained *p* value, we will classify this as a *Decision Error*. Note that the classification of Decision Error will be made in addition to any observed Minor Numerical Error or Major Numerical Error.

After tallying up these four types of errors for an individual reproducibility check, we will decide on a final report outcome based on the following criteria:

- If there are any Insufficient Information Errors, Major Numerical Errors, or Decision Errors then the reproducibility check will be considered an overall **failure**.

- If there are only Minor Numerical Errors, or no discrepancies, then the reproducibility check will be considered an overall **success**.

We will complement quantitative error tallies with qualitative information about the types of errors derived from the Pilot/Co-pilot reproducibility reports. Note that final outcomes will be based on the final joint report.

Procedure

Overall workflow

An overview of the workflow for reproducibility checks is outlined in Figure 1. We will employ a data co-piloting model (Veldkamp et al., 2014; Hardwicke et al.) in which every reproducibility check involves at least two team members verifying each other's work.

Pilot/Co-pilot eligibility

Pilots/co-pilots will meet the following eligibility criteria:

- started or completed graduate level training in a scientific discipline, preferably psychology or related fields

- must have experience with, and be comfortable with, conducting the types of data analyses typically found in an introductory psychology statistics textbook (e.g., Field et al., 2012).

- must be able to run and document their reproducibility checks using *R* (and preferably *R* Markdown).

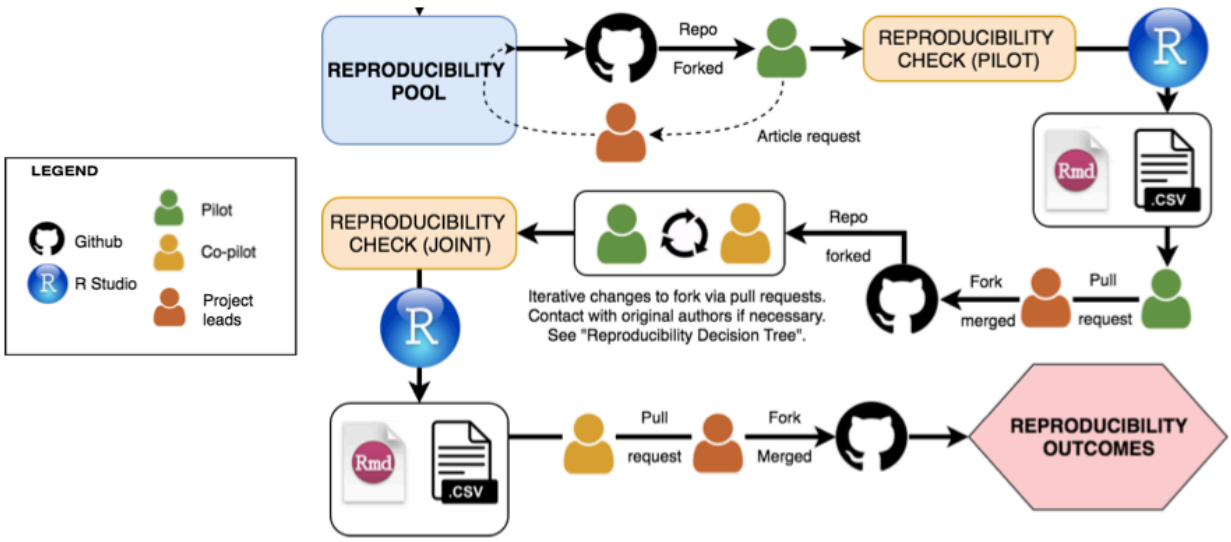

Figure 1. Overview of study workflow (after Hardwicke et al.). The workflow relies on two key infrastructure components: Github for version control and collaboration, and R Studio for running analysis and literate programming.

Conducting reproducibility checks

The principle aim of a reproducibility check is to reproduce the defined subset of target outcomes using the available data files, information provided in the original article, and any other additional documentation (e.g., codebook or analysis scripts). The Pilot/Co-pilot's role is *not* to attempt or suggest alternative analyses as that is outside the focus of this investigation.

When team members encounter ambiguous or absent information necessary for completing their reproducibility check, they should *not* engage in lengthy guesswork, but should report that there is an "Insufficient Information Error" (see error classification scheme above). These issues can be resolved by contacting the original authors and requesting additional information or clarifications.

Requesting author clarification/assistance

If Pilot and Co-Pilot conclude that the target outcomes are not reproducible for a given article, they will e-mail corresponding authors to request clarification/assistance (see Figure 2). The e-mail will contain the Pilot and Co-Pilot's preliminary report (in .Rmd and .html formats) documenting the reproducibility attempt, and a list of the key issues that require resolution. The e-mail will be sent to the article's corresponding author. If it is obvious that an e-mail address is inactive (e.g., the message rebounds) then the Pilot will conduct an Internet search

and attempt to identify a more recent contact e-mail (the pilot will spend no more than 5 minutes on this activity). If the corresponding author has not responded after two weeks, they will be contacted for a second time. No further attempts will be made to contact the corresponding author.

After contacting authors for assistance, a maximum time-limit of 2 months for resolution of the any issues will be observed. Any reproducibility issues that cannot be resolved within this time-period will be considered reproducibility failures for the purposes of the present investigation.

Literate programming

Reports will adopt a 'literate programming' approach that involves interleaving segments of analysis code with plain language documentation. Specifically, reports will be written in R Markdown format which includes R code and lightly formatted text ('markdown'). Pilots will work with a standardised template. Pilots will be encouraged to populate their reports with direct quotations and images extracted from the original article.

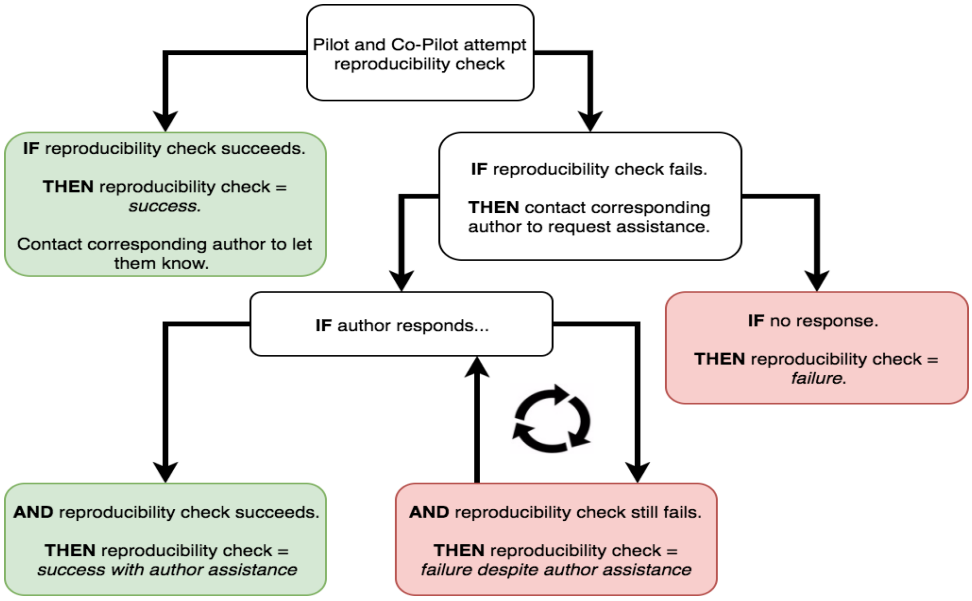

Figure 2. Decision tree for the Pilot/Co-Pilot joint report (after Hardwicke et al.).

Analysis plan

We will report descriptive statistics for the measured variables outlined below. Note that we do not intend to report on minor numerical errors as we do not consider them informative with regards to overall analytic reproducibility.

Table 2. Measured variables and analysis for reproducibility checks.

Measured variable	Analysis
Insufficient Information Errors	Total counts, proportion of articles containing at least one of these error types.
Major Numerical Errors	Total counts, proportion of articles containing at least one of these error types.
Decision Errors	Total counts, proportion of articles containing at least one of these error types.
Pilot and Co-Pilot summary (e.g., identifying possible reasons for a reproducibility failure)	Depending on the quantity and variety of reasons provided, we will attempt to (post-hoc) classify reasons for reproducibility failure into a succinct number of discrete categories and report the proportion of articles in each category.
Final reproducibility outcome (success, success with author assistance, failure, failure with author assistance)	Proportion of articles in each category. Initial failure rate (i.e., prior to author assistance) and final failure rate (i.e., after author assistance).
Subjective time-to-complete estimate (combined for all pilots)	Mean & standard deviation / median & interquartile range as appropriate based on data distribution.

Commitment to open practices

This research project will employ the following open practices:

- **Pre-registration:** The study protocol will be pre-registered on the Open Science Framework.

- **Open access**: Upon completion of the study, a manuscript will be written up and submitted to a preprint server. Subsequently, the manuscript will be submitted to a journal that has an open access option.

- **Open data and materials**: All study materials, data, and analysis scripts will be made openly available on the Open Science Framework upon completion of the study. This will include the verbatim reports submitted by Coders and Pilots.

References

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., et al. (2013). Recommendations for Increasing Replicability in Psychology. *European Journal of Personality*, *27*(2), 108–119.

Bollen, K., Cacioppo, J. T., Kaplan, R. M., Krosnick, J. A., & Olds, J. L. (2015). *Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science*. *National Science Foundation* (pp. 1–29).

Goodman, S. N., Fanelli, D., & Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Science Translational Medicine*, *8*, 1–6.

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., et al. (2016). A Multilab Preregistered Replication of the Ego-Depletion Effect. *Perspectives on Psychological Science*, *11*(4), 546–573.

Hardwicke, T. E., Mathur, M. B., Nilsonne, G., MacDonald, K., Banks, G. C., Kidwell, M. C., Clayton, E., Yoon, E. J., Mohr, A., Tessler, M. H., Lenne, R., Altman, S., Long, B., & Frank, M. C. (*in preparation*). Data availability, reusability, and analytic reproducibility at the journal Cognition: Evaluating the impact of a mandatory open data policy.

Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L.-S., et al. (2016). Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. *PLoS Biology*, *14*(5), 1–15.

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.

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence? *Perspectives on Psychological Science*, *7*(6), 528–530.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), 1–8.

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, *9*(1), 76–80.

Stodden, V. (2015). Reproducing Statistical Results. *Annual Review of Statistics and Its Application*, *2*(1), 1–19. http://doi.org/10.1146/annurev-statistics-010814-020127

Vanpaemel, W., Vermorgen, M., Deriemaecker, L., & Storms, G. (2015). Are we wasting a good crisis? The availability of psychological research data after the storm. *Collabra*, *1*, 1–5.

Veldkamp, C. L. S., Nuijten, M. B., Dominguez-Alvarez, L., van Assen, M. A. L. M., & Wicherts, J. M. (2014). Statistical Reporting Errors and Collaboration on Statistical Analyses in Psychological Science. *PLoS ONE*, *9*(12), e114876–20.

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, *61*(7), 726–728.

Appendix A. Articles excluded from Kidwell et al. sample.

Article id codes* by reason for exclusion are shown below.

Does not meet criteria for a 'straightforward and substantive analysis:
12-8-2014 PS, 21-7-2014 PS, 13-6-2014 PS, 16-2-2015 PS, 18-7-2014 PS, 3-2-2015 PS, 6-3-2015 PS, 16-8-2014 PS, 1-10-2013 PS, 19-4-2015 PS, 8-10-2014 PS, 15-2-2015 PS.

Link to data file not working:
8-11-2014 PS

* as assigned by Kidwell et al.