Ecological validity of detecting data fabrication in experimental studies.

Chris HJ Hartgerink

18 March, 2016

# Study 1

To test the validity of statistical methods to detect data fabrication in summary results, we investigated summary results for four anchoring studies (Jacowitz & Kahneman, 1995). We selected the anchoring effect because it is a well-known psychological phenomenon and data from the Many Labs were available for inclusion in our study. The unit of analysis for this study is the set of summary statistics for the four anchoring studies from one respondent. Respondent is defined as either one of the Many Labs locations, or a researcher who participated in fabricating results.

## Methods

The four anchoring studies for which results were collected were (i) distance from San Francisco to New York, (ii) population of Chicago, (iii) height of the Mount Everest, and (iv) the number of babies born per day in the United States. Each of the four studies provided summary results for a two-factorial design, including a high- and low anchor condition and gender. These four anchoring studies were selected because quality data were available from the Many Labs project (Klein et al., 2014).

### Data collection.

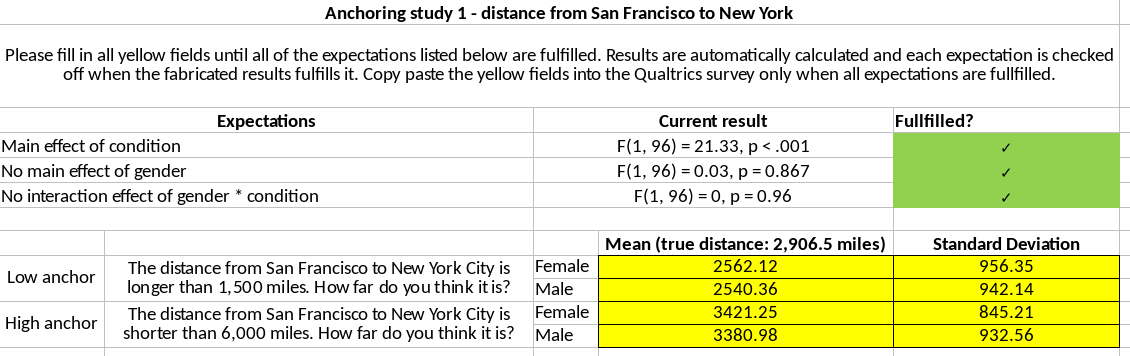
Thirty-six genuine datasets were collected from the publicly available Many Labs project (Klein et al., 2014, osf.io/pqf9r). The Many Labs (ML) project replicated several effects, including the anchoring effect, across thirty-six locations. ML replicated the anchoring effect with the four studies mentioned previously. Considering the size of the ML project, the transparency of research results, and minimal individual gain for fraud, we assumed these data to be genuine. For each of the thirty-six locations, means and standard deviations were collected for each of the four conditions in the four anchoring studies. In total, this yielded both 16 means per location and 16 variances per location. The raw data were cleaned according to the original analysis script from the ML project.

Thirty-six fabricated sets of summary results were collected for all four anchoring studies. We sampled 2,038 American psychology researchers who published a peer-reviewed paper in 2015, as indexed in the Web of Science (WoS). Psychology researchers were sampled to improve familiarity with the anchoring effect (Jacowitz & Kahneman, 1995), for which summary results were fabricated. U.S. researchers were sampled to ensure familiarity with the imperial measurement system, which is the scale of some of the anchoring studies. WoS was searched on October 13, 2015. In total, 2,038 unique corresponding e-mails were extracted from 2,014 papers (due to multiple corresponding authors).

The full sample frame was digitally approached to participate in this study on March XX, 2016 (invitation: osf.io/XXXX). The study took place via Qualtrics with anonimization procedures in place (e.g., no IP-address saved at all). The researchers were fully informed that the study would require them to fabricate data and that we conducted this study to test the validity of statistical methods to detect data fabrication. Participants were also informed they could stop at any time without providing a reason. If they wanted, participants received a $30 Amazon gift card as compensation for their participation for which they had to provide their email address. These email addresses were unlinked from email addresses upon completion of the study and sending out the gift cards.

Each respondent was instructed to fabricate summary statistics that fulfilled three hypotheses. That is, 4 (studies) × 2 (conditions) × 2 (sex) × 2 (mean or sd) = 32 statistics were fabricated. We instructed participants to fabricate results for the hypotheses (i) main effect of condition, (ii) no effect of sex, and (iii) no interaction effect between condition and sex. The fabricated summary statistics and their accompanying test results for these three hypotheses serve as the data that is used to test the detection of data fabrication.

In order to standardize the fabrication process to a minimal extent, we provided participants with a spreadsheet where the fabricated data had to be filled in. The fabricated data are checked against the provided hypotheses and the statistical test are provided instantaneously, given the fabricated means and standard deviations. Figure 1 depicts an example of this spreadsheet. Respondents were required to copy-paste the yellow cells into Qualtrics, to provide a standardized way of reporting the means and standard deviations.



**Figure 1.** Example of a filled in spreadsheet used in the fabrication process.

Upon completing the fabrication of the data, participants were debriefed. Several questions were asked about their statistical knowledge and approach to data fabrication, and they were reminded that data fabrication is widely condemned by professional organizations, institutions, and funding agencies alike. The full set of questions from the study are available at osf.io/XXXX.

Participation was rewarded with a $30 Amazon gift card and the fabricated results that were most difficult to detect received a bonus $50 Amazon gift card. If the participant wanted to receive a compensation and contend for the bonus $50, he/she had to enter an email to receive the reward. These email addresses were unlinked from individual responses upon sending the gift cards. Quotum sampling was applied to sample as many responses as possible for the available 36 rewards (i.e., not all respondents might request the gift card and count towards the quotum).

### Data analysis.

We included two methods to test for data fabrication and we combined the results of these two tests. These statistical tests compute the probability that, given sampling theory, the current data or an even more unlikely set would be observed. As a result, this probability can be seen as a *p*-value for the null hypothesis that the data is genuine (i.e., P[data|genuine]).

First we apply the "variance of variances" method (Simonsohn, 2013). This method tests whether the observed variances contain a reasonable amount of variation expected based on stochastic sampling processes. For example, if four independent samples all yield the variance of 2.22, this could be considered excessively consistent. More technically, the method first estimates the observed variation of the standardized standard deviations. Second, the method simulates how much variation in the standardized standard deviations can be expected given the sample sizes, means and variances. This yields an estimate of the probability that the observed lack of variation in the variances would be observed, or even less. This is in fact a *p*-value that tests the hypothesis that the observed variances are genuine and originate from a stochastic sampling distribution, under a set of distributional assumptions.

Second, we apply the reversed Fisher method, which tests whether there is an excessive amount of high *p*-values. The default Fisher method (Fisher, 1925) tests whether a set of *p*-values deviates from uniformity, where smaller *p*-values occur more frequently than larger *p*-values (i.e., whether there is evidence for an effect). However, the reversed Fisher method tests whether a set of *p*-values deviates from uniformity in such a way that there are more large *p*-values than small ones. As such, it tests whether a theoretically infeasible distribution of *p*-values is presented in a study, which is overly consistent with the null hypothesis. The reversed Fisher method is computed as

where *t* is the lowerbound (i.e., threshold) of the *k* number of *p*-values taken into account. The resulting *p*-value indicates the probability that the observed nonsignificant *p*-values are actually from a genuine null distribution.

Third, we combine the results of the "variance of variances" test and reversed Fisher method by testing for evidential value. To this end, we apply the default Fisher method. The Fisher method tests whether there is sufficient evidence for deviation from genuine data in the two separate tests and is computed as

Given that this is a combination test, we expect this to be a more powerful test than the "variance of variances" method or the reversed Fisher method alone. The three respondents with the highest *p*-value contained the least evidential value for deviating from genuine data and received an additional $50 Amazon gift card.

# References

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburg, United Kingdom: Oliver Boyd.

Jacowitz, K. E., & Kahneman, D. (1995). Measures of anchoring in estimation tasks. *Personality & Social Psychology Bulletin*, *21*, 1161–1166. Retrieved from <http://facweb.plattsburgh.edu/wendy.braje/students/psy205/JKarticle.pdf>

Klein, R. A., Ratliff, K. A., Vianello, M., Jr., R. B. A., Bahník, Š., Bernstein, M. J., … Nosek, B. A. (2014). Investigating variation in replicability. *Social Psychology*, *45*(3), 142–152. <http://doi.org/10.1027/1864-9335/a000178>

Simonsohn, U. (2013). Just post it: The lesson from two cases of fabricated data detected by statistics alone. *Psychological Science*, *24*(10), 1875–1888. <http://doi.org/10.1177/0956797613480366>