# Ecological validity of detecting data fabrication in experimental studies.

*Chris HJ Hartgerink*

*25 April, 2016*

## Study 1

To test the validity of statistical methods to detect data fabrication in summary results, we investigated summary results of four anchoring studies (Jacowitz & Kahneman, 1995; Tversky & Kahneman, 1974). We selected the anchoring effect because it is a well-known psychological phenomenon and many genuine data sets on the effect are available from the Many Labs project (Klein et al., 2014). The unit of analysis for this study is the set of summary statistics (i.e., means, standard deviations, and test results) for the four anchoring studies from one respondent. Respondent is defined here as either one of the Many Labs locations, or a researcher who fabricated the four anchoring studies' summary statistics.

### 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 2 (low/high anchoring) $\times$ 2 (male/female) factorial design.

#### 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 across thirty-six locations, including the anchoring effect in 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, sample sizes, means, and standard deviations (four each) were collected for each of the four conditions in the four anchoring studies across the thirty-six locations (i.e., 48 $\times$ 36). These summary statistics were collected from the raw ML data, which were cleaned using the original analyses scripts from the ML project to ensure that the data did not contain improper responses.

Thirty-six fabricated sets of summary results were also 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; Tversky & Kahneman, 1974), 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 April XX, 2016 (invitation: osf.io/s4w8r). 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 32 summary statistics (4 studies × 2 conditions × 2 sexes × 2 statistics [mean and sd]) that fulfilled three hypotheses. Respondents did not need to fabricate sample sizes, which were set to 25 per cell a priori. 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 to examine the properties of tools to detect 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 out. Figure 1 depicts an example of this spreadsheet. The yellow cells are those that the respondent is requested to fabricate and include both the means and the standard deviations. Using these values, statistical tests are computed and shown in the "Current result" column instantaneously. When these results corroborate the given expectations, a checkmark appears as in the green cells in Figure 1. We required respondents to copy-paste the yellow cells into Qualtrics, to provide a standardized response format.

**Anchoring study - distance from San Francisco to New York**

| Expectations | | Current result | | Supported |
|---|---|---|---|---|
| Main effect of condition | | F(1, 96) = 21.33, p < .001 | | ✓ |
| No main effect of gender | | F(1, 96) = 0.03, p = 0.867 | | ✓ |
| No interaction effect of gender * condition | | F(1, 96) = 0, p = 0.96 | | ✓ |
| | | | | |
| | | | **Mean (true distance: 2,906.5 miles)** | **Standard Deviation** |
| Low anchor | The distance from San Francisco to New York City is longer than 1,500 miles. How far do you think it is? | Female | 2562.12 | 956.35 |
| | | Male | 2540.36 | 942.14 |
| High anchor | The distance from San Francisco to New York City is shorter than 6,000 miles. How far do you think it is? | Female | 3421.25 | 845.21 |
| | | Male | 3380.98 | 932.56 |

**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/w984b.

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.

To detect data fabrication in a set of summary results, we first test the variance of fabricated standard deviations (SDs; Simonsohn, 2013) across the four anchoring studies. This method tests whether the observed SDs contain a reasonable amount of variation, as 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 when the probability that this amount of consistency (or more), given genuine samples, is less than 1 out of 1000 (this is simply an example). To compute this test, we first standardize the SDs for each of the four studies by computing

$$\tilde{s}_j = \frac{s_j^2}{MS_w} = \frac{s_j^2}{\frac{\sum(N_j-1)s_j^2}{\sum(N_j-1)}}$$

where $\tilde{s}_j$ denotes the standardized SD in group $j$. Note that $MS_w$ is the simple arithmetic mean when sample sizes are equal for all cells. We test several different measures to detect data fabrication that utilize these

standardized SDs (i.e., $\tilde{s}_j$), including the max-min distance (denoted $\tilde{\sigma}_{max-min}$) and the variance of the standardized SDs (i.e., $\tilde{\sigma}_{sd}$; Simonsohn, 2013). We compare the observed value for each $\tilde{\sigma}$ measure with the expected distribution of outcomes when the data would be the result of random sampling processes. To this end, we simulate the expected distribution of standardized SDs and compute the expected distribution of each $\tilde{\sigma}$. This expected distribution is used to determine the $p$-value of the observed $\tilde{\sigma}$ value. We simulate the standardized SD for each of the $j$ groups as

$$\tilde{s}_j^2 \sim \left[\frac{\chi^2_{N_j-1} s_j^2}{N_j-1}\right] / MS_w$$

These simulated values are used to compute the expected distribution of all $\tilde{\sigma}$ measures.

Second, we apply the reversed Fisher method to the fabricated nonsignificant $p$-values twice, once for the gender effects hypothesis and once for the interaction effects hypothesis, in order to detect data fabrication. The Fisher method (Fisher, 1925) tests for evidence of an effect in a set of $p$-values and has previously been used as a meta-analytic method (Hong & Breitling, 2008), but we adjust it here to test for results that are overly consistent with the null hypothesis. The original Fisher method is computed as

$$\chi^2_{2k} = -2 \sum \ln(p_i)$$

and tests for right-skew in a set of $p$-values, but we adjust it to the following

$$\chi^2_{2k} = -2 \sum \ln(1 - \tfrac{p_i - t}{1-t})$$

where it now tests for left-skew (i.e., more larger $p$-values than smaller $p$-values) across the $k$ number of $p$-values that falls above the threshold $t$. We set this threshold to .05 in order to include only nonsignificant test results.

Finally, we combined the results of these three individual tests for data fabrication using the Fisher method. We expect this combination test of the three individual tests for data fabrication to be a more powerful than the individual tests. Based on the results of the combined test results, the three 'best' data fabricators are selected. The three respondents with the highest $p$-values contain the least evidential value for deviating from genuine data and receive an additional $50 Amazon gift card.

For each of these three tests to detect data fabrication we carry out sensitivity and specificity analyses using ROC-curves in order to determine optimal alpha levels. This analysis helps determine the classification performance of these three tests. In order to determine the optimal alpha level, we varied the alpha level from .000001 through .1 and assessed the classification performance of these methods to detect data fabrication. The optimal alpha level is determined by finding that alpha level for which both the true positive classification rate and the true negative classification rate are highest. For example,

# References


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

Hong, F., & Breitling, R. (2008). A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics*, *24*(3), 374–382. http://doi.org/10.1093/bioinformatics/btm620

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

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. http://doi.org/10.1126/science.185.4157.1124