Ecological validity of detecting data fabrication in experimental studies.

Chris HJ Hartgerink, Jelte M Wicherts, Marcel ALM van Assen

August 2, 2016

Study 1

We tested the performance of statistical methods to detect data fabrication based on summary results with genuine and fabricated summary results of four anchoring studies (Tversky and Kahneman, 1974; Jacowitz and Kahneman, 1995). The anchoring effect is a well-known psychological heuristic that uses the information in the question as the starting point for the answer, which is then adjusted to yield a final estimate of a quantity. For example 'Is the percentage of African countries in the United Nations more or less than [10% or 65%]?'. These questions yield mean responses of 25% and 45%, respectively (Tversky and Kahneman, 1974), despite essentially posing the same factual question. A considerable amount of genuine datasets on this heuristic are freely available and we collected fabricated datasets within this study.

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) × 2 (male/female) factorial design. Throughout this study, the unit of analysis is a set of summary statistics (i.e., means, standard deviations, and test results) for the four anchoring studies from one respondent. Respondent is defined as researcher/lab where the four anchoring studies' summary statistics originate from. All materials, data, and analyses scripts are freely available on the OSF (`osf.io/b24pq`) and were preregistered (`osf.io/ejf5x`; deviations are explicated in this report).

Data collection

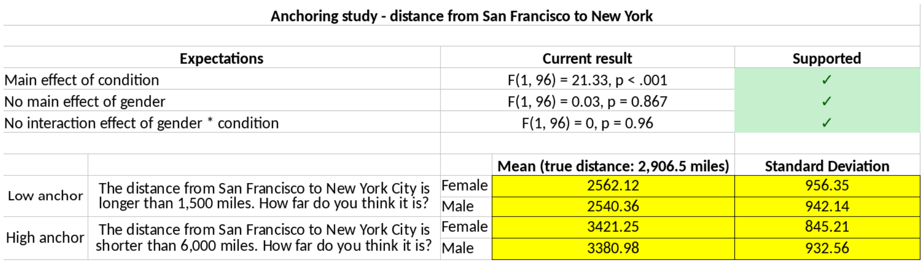
We downloaded thirty-six genuine datasets from the publicly available Many Labs (ML) project (`osf.io/pqf9r`; Klein et al., 2014). The ML project replicated several effects across thirty-six locations, including the anchoring effect in the four studies mentioned previously. Considering the size of the 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 computed, for each of the four conditions in the four anchoring studies across the thirty-six locations (i.e., $3 \times 4 \times 4 \times 36$). We computed these summary statistics from the raw ML data, which were cleaned using the original analysis scripts from the ML project.

Using quotum sampling, we collected thirty-six fabricated datasets of summary results for all four anchoring studies. The sampling frame consisted of 2,038 psychology researchers who published a peer-reviewed paper in 2015, as indexed in the Web of Science (WoS) with the filter set to the U.S. We sampled psychology researchers to improve familiarity with the anchoring effect (Jacowitz and Kahneman, 1995; Tversky and Kahneman, 1974), for which summary results were fabricated. We filtered for U.S. researchers to ensure familiarity with the imperial measurement system, which is the scale of some of the anchoring studies (note: we found out several non-U.S. researchers were included because this filter also retained papers with co-authors from the U.S.). 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).

A random sample of 1,000 researchers were digitally approached to participate in this study on April 25, 2016 (invitation: `osf.io/s4w8r`). The study took place via Qualtrics with anonimization procedures in place (e.g., no IP-addresses saved). We informed the participating researchers that the study would require them to fabricate data and explicitly mentioned that we would investigate these data with statistical methods to detect data fabrication. We also clarified to the respondents that they could stop at any time without providing a reason. If they wanted, respondents received a $30 Amazon gift card as compensation for their participation for which they had to provide their email address, after which they could win an additional $50 Amazon gift card if they were one of three top fabricators. The provided email addresses were unlinked from individual responses upon sending the bonus gift cards. The full text of the Qualtrics survey is available at `osf.io/w984b`.

Each respondent was instructed to fabricate 32 summary statistics (4 studies × 2 conditions × 2 sexes × 2 statistics [mean and sd]) that fulfilled three hypotheses. We instructed respondents 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. Respondents did not need to fabricate sample sizes, which were set to 25 per cell a priori. 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.

We provided respondents with a template spreadsheet to fill out the fabricated data, in order to standardize the fabrication process without restraining the participant in how they choose to fabricate data. Figure 1 depicts an example of this spreadsheet (original: `osf.io/w6v4u`). We requested the respondents to fill in the yellow cells with fabricated data, which includes means and the standard deviations for four conditions. Using these values, statistical tests are com-



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 template spreadsheet used in the fabrication process. Respondents fabricated data in the yellow cells, which were used to compute the results of the hypothesis tests. If the fabricated data confirm the hypotheses, a checkmark appeared in a green cell.

puted and shown in the ”Current result” column instantaneously. When these results confirm the hypotheses, a checkmark appears as depicted in the green cells. We required respondents to copy-paste the yellow cells into Qualtrics, to provide a standardized response format that could be automatically processed in the analyses.

Upon completing the fabrication of the data, respondents were debriefed. Several questions were asked about their statistical knowledge and approach to data fabrication. They were also reminded that data fabrication is widely condemned by professional organizations, institutions, and funding agencies alike.

We rewarded participation 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; one participant did not request a reward).

Data analysis

To detect data fabrication in a set of summary results, we first tested the standardized standard deviations (SDs) for data fabrication (Simonsohn, 2013) across the four anchoring studies. Such methods test whether the observed SDs contain a reasonable amount of variation, as expected based on random sampling processes. For example, if four independent samples all yield the variance 2.22, this could be considered excessively consistent when the probability that this amount of consistency (or more) is less than 1 out of 1000 in truly random samples. To compute this test, we first standardized the SDs for each of the four studies by computing

$$z_j = \sqrt{\frac{s_j^2}{MS_w}} = \sqrt{\frac{s_j^2}{\left(\frac{\sum_{j=1}^{k}(N_j-1)s_j^2}{\sum_{j=1}^{k}(N_j-1)}\right)}} \tag{1}$$

where z_j denotes the standardized SD in group j (MS_w is the simple arithmetic mean when sample sizes are equal for all cells, which is the case for the fabricated datasets). We tested different measures to detect data fabrication that utilize these standardized SDs (i.e., z_j). We included the variance of the standardized SDs (i.e., SD_z; Simonsohn, 2013) and tried out the max-min distance of the standardized SDs (denoted $max - min_z$) as an alternative measure. We compared the observed value for each measure with the expected distribution when the summary results are used to generate random samples. To this end, we simulated the expected distribution of standardized SDs and computed the expected distribution of each measure. This expected distribution was used to determine the p-value of the observed SD_z and $max - min_z$. We simulated the standardized variance for each of the j groups as

$$z_j^2 \sim \left(\frac{\chi^2_{N_j-1}}{N_j - 1}\right)/MS_w \tag{2}$$

These simulated values are used to compute the expected distribution of the SD_z and $max - min_z$ measures.

Second, we applied the reversed Fisher method to the 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 by testing for a right-skew p-value distribution, but we adjusted it here to test for results that are overly consistent with the null hypothesis and result in a left-skew distribution (see Figure 2). The original Fisher method is computed as

$$\chi^2_{2k} = -2\sum_{i=1}^{k}\ln(p_i) \tag{3}$$

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

$$\chi^2_{2k} = -2\sum_{i=1}^{k}\ln(1 - \frac{p_i - t}{1 - t}) \tag{4}$$

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.

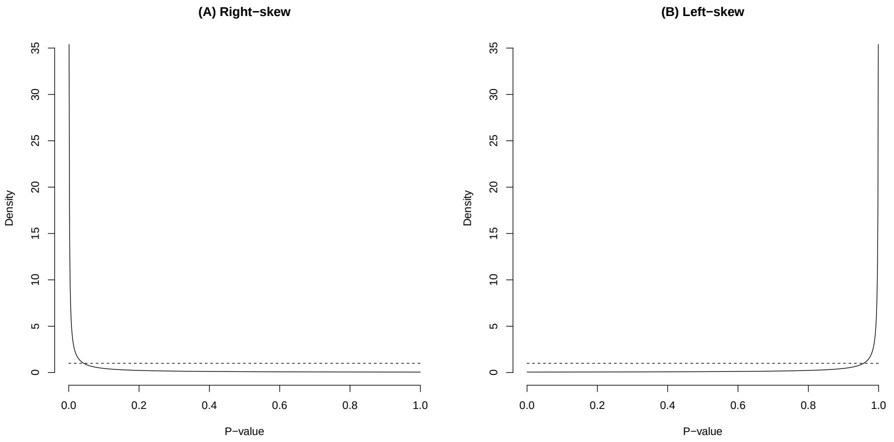


Figure 2: Conceptual representation of what the Fisher test inspects (Equation **??**; panel A) and our adjusted Fisher test inspects (Equation 4; panel B). Both panels test whether there is sufficient evidence that the solid line deviates from the dashed line, except that the type of deviation that the test is sensitive to is the exact opposite.

Results

References

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

Jacowitz, K. E. and Kahneman, D. (1995). Measures of anchoring in estimation tasks. *Personality & social psychology bulletin*, 21:1161–1166.

Klein, R. A., Ratliff, K. A., Vianello, M., Jr., R. B. A., Bahník, v., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., Hasselman, F., Hicks, J. A., Hovermale, J. F., Hunt, S. J., Huntsinger, J. R., IJzerman, H., John, M.-S., Joy-Gaba, J. A., Kappes, H. B., Krueger, L. E., Kurtz, J., Levitan, C. A., Mallett, R. K., Morris, W. L., Nelson, A. J., Nier, J. A., Packard, G., Pilati, R., Rutchick, A. M., Schmidt, K., Skorinko, J. L., Smith, R., Steiner, T. G., Storbeck, J., Swol, L. M. V., Thompson, D., ‘t Veer, A. E. v., Vaughn, L. A., Vranka, M., Wichman, A. L., Woodzicka, J. A., and Nosek, B. A. (2014). Investigating variation in replicability. *Social psychology*, 45(3):142–152.

Simonsohn, U. (2013). Just post it: The lesson from two cases of fabricated data detected by statistics alone. *Psychological science*, 24(10):1875–1888.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131.