Ecological performance of detecting data fabrication with summary statistics

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

October 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) $\times$ 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. For current purposes, a 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 (`https://osf.io/b24pq`) and were preregistered (`https://.io/ejf5x`; deviations are explicated in this report).

Data collection

We downloaded thirty-six genuine datasets from the publicly available Many Labs (ML) project (`https://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. 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). 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 approached via e-mail to participate in this study on April 25, 2016 (invitation: `https://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 if they were willing to enter their email address. 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 `https://osf.io/w984b`.

Each respondent was instructed to fabricate 32 summary statistics (4 studies $\times$ 2 conditions $\times$ 2 sexes $\times$ 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 fabri-

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.

cated data, in order to standardize the fabrication process without restraining the participant in how they chose to fabricate data. Figure 1 depicts an example of this spreadsheet (original: `https://osf.io/w6v4u`). We requested 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 computed and shown in the "Current result" column instantaneously. If these results confirmed the hypotheses, a checkmark appeared as depicted in Figure 1. 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. Respondents answered several questions about their statistical knowledge and approach to data fabrication and finally we reminded them 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.

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. This method tests 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 probability, we first standardized the SDs for each of

the four studies with

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

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.

Testing the standardized SDs for potential data fabrication can be done either for each study separately or all studies combined; the test can also be done under different assumptions of population variances across conditions. The assumptions of population variance can either by that all SDs originate from the same distribution (as in Simonsohn, 2013), the SDs within a factor are from the same distribution, or each group comes from its own distribution. We preregistered the method that assumes the SDs are drawn from the same distribution for the various conditions (i.e., homogeneous SDs) and are tested across all studies. However, upon conducting the analyses, we decided homogenous SDs are not unequivocal and included computations where the SDs for the low anchor and high anchor are from different distributions (i.e., heterogeneous SDs). Additionally, the signal for data fabrication across the four anchoring studies might result in different studies cancelling each other out, so we also included analyses where each study was analyzed separately.

Second, we applied the reversed Fisher method to detect data fabrication to the nonsignificant p-values twice: once for the results of gender effects hypothesis in each study and once for the results of the interaction effect hypothesis for each study. 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. The theoretical idea behind this method is that researchers who fabricate nonsignificant data might forget to fabricate a uniform p-value distribution, given the frequent misinterpretation of p-values (e.g., as the probability of an effect, Goodman, 2008; Altman and Bland, 1995).

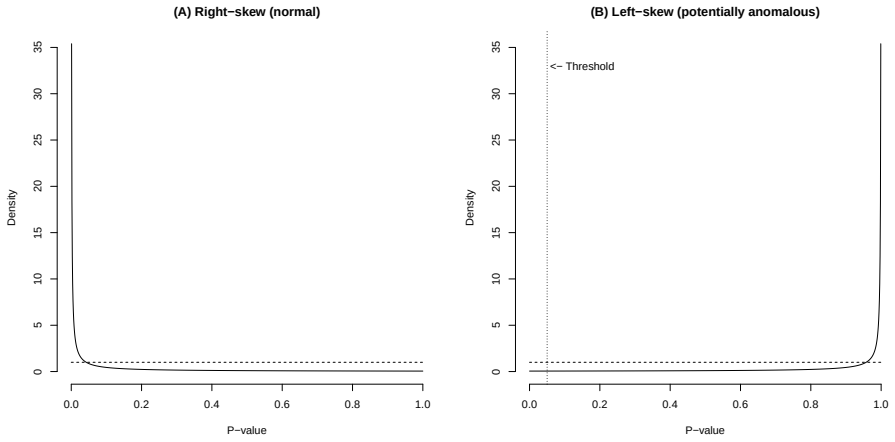

Figure 2: Conceptual representation of what the Fisher test inspects (Equation 3; panel A) and what the adjusted Fisher test inspects (Equation 4; panel B). Both panels test whether there is sufficient evidence that the solid line deviates from the dashed line, except that the type of deviation that the test is sensitive to is the exact opposite.

Finally, we combined the aforementioned methods to detect data fabrication with the Fisher method. This included the SD_z measure across all studies and the Fisher test (Equation 4) of the gender hypothesis test and the interaction test. We expected this combination test of the three individual tests for data fabrication to be more powerful than the individual tests, given that these tests inspect different manifestations of data fabrications. Based on the results of the

combined test results, the three least detectable data fabricators were selected. The three respondents with the highest p-values on this combined method to detect data fabrication contained the least evidential value for deviating from genuine data and received an additional $50 Amazon gift card.

For each of these four tests to detect data fabrication (SD_z, Fisher test for the gender and interaction hypotheses, combined methods) we carried out sensitivity and specificity analyses using ROC-curves. Analyses were conducted with the `pROC` package in `R` (`v1.8`; Robin et al., 2011). ROC-analyses indicate the sensitivity (i.e., True Positive Rate [TPR]) and specificity (i.e., True Negative Rate [TNR]) for various decision criteria (e.g., $\alpha = 0, .01, .02, ..., .99, 1$). With these ROC-curves, informed decisions about optimal alpha levels can be made based on various criteria. In this case, we determine the optimal alpha level by finding that alpha level for which the combination of TPR and TNR were highest. For example, if $\alpha = .04$ results in $TPR = .30$ and $TNR = .70$, but $\alpha = .05$ results in $TPR = .5$ and $TNR = .5$, .05 was chosen as an optimal decision criterion based on the sample.

Results

The collected data included 36 genuine data from Many Labs 1 (`https://osf.io/pqf9r`; Klein et al., 2014) and 39 fabricated datasets(`https://osf.io/e6zys`; 3 participants did not participate for a bonus).

Figure 3 shows a group-level comparison of the genuine- and fabricated p-values and effect sizes (r). Such group-level comparisons provide an overview of the differences between the genuine- and fabricated data (see also Akhtar-Danesh and Dehghan-Kooshkghazi, 2003). These distributions indicate little group differences between genuine- and fabricated data when nonsignificant effects are inspected (i.e., gender and interaction hypotheses) whereas there seem to be large group differences when we required subjects to fabricate significant data (i.e., condition hypothesis). From our own experience, and anecdotal evidence elsewhere (Bailey, 1991), large effects have previously raised initial suspicions. These data corroborate the idea that extremely large effect sizes (e.g., $r > .95$) might prove to be an easy-to-implement flag for potentially anomalous data (it is wise to seek for alternative explanations after flagging, however). Considering this, we also investigated how well effect sizes perform in detecting data fabrication (not preregistered). In the following sections, we investigate the performance of such statistical methods to detect data fabrication on an respondent-level basis.

Performance of using SDs to detect data fabrication

We applied two different operationalizations to inspect for data fabrication in the genuine- and fabricated datasets based on variance in the SDs. The SD_z method (Simonsohn, 2013) inspects whether the variance of the SDs themselves varies sufficiently, and we tested another method that inspects the range of the SDs varies sufficiently. These two methods operate similarly, except for the

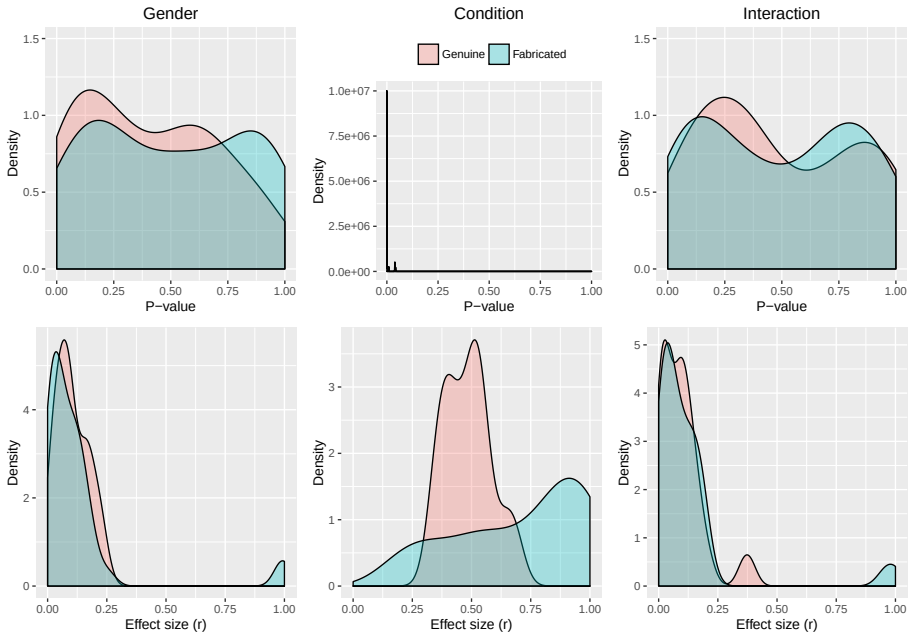

Figure 3: Overlay of density distributions for both genuine and fabricated data, per effect and type of result. We instructed respondents to fabricate nonsignificant data for the gender and interaction effects, and a significant effect for the condition effect.

final computation. Other operationalizations are possible and we invite others to reuse the data from this study (`https://osf.io/b24pq`).

Table 1 indicates that both methods show similar performance when inspected with the Area Under the Receiver Operating Curve (AUROC), but with a preference for SD_z. This preference is due to the variance of variances outperforming the range of variances when all studies are combined in one test and its past application (Simonsohn, 2013). Throughout the rest of this section, we will only discuss the results of SD_z further.

The AUROC indicates the probability that a randomly drawn fabricated dataset is classified as fabricated before a randomly drawn genuine dataset is classified as fabricated (Hanley and McNeil, 1982). In other words, if AUROC = .5 this indicates that correctly classifying a randomly drawn dataset in this sample is equal to a coin flip. For this setting, we will regard any AUROC $< .6$ as plainly insufficient for detecting data fabrication, $.6 \leq$ AUROC $< .7$ as failed, $.7 \leq$ AUROC $< .8$ as sufficient, $.8 \leq$ AUROC $< .9$ as good, and $.9 \leq$ AUROC ≤ 1 as excellent.

The AUROC for the variance of variance analyses is insufficient or sufficient depending on the assumption regarding population variances. Table 1

Method	AUROC SD_z	AUROC $maxmin_z$
Homogeneous, all studies combined	0.423	0.303
Homogeneous, study 1	0.368	0.374
Homogeneous, study 2	0.421	0.446
Homogeneous, study 3	0.510	0.521
Homogeneous, study 4	0.540	0.542
Heterogeneous, all studies combined	0.770	0.756
Heterogeneous study 1, low anchor condition	0.643	0.643
Heterogeneous study 1, high anchor condition	0.438	0.438
Heterogeneous study 2, low anchor condition	0.751	0.751
Heterogeneous study 2, high anchor condition	0.612	0.612
Heterogeneous study 3, low anchor condition	0.668	0.668
Heterogeneous study 3, high anchor condition	0.651	0.651
Heterogeneous study 4, low anchor condition	0.796	0.796
Heterogeneous study 4, high anchor condition	0.555	0.555

Table 1: Area Under the Receiver Operating Curve (AUROC) for the two statistical methods used to detect data fabrication based on variances. Homogeneous: assumes one population variance underlying all groups. Heterogeneous: assumes separate population variance per anchoring condition.

indicates that AUROC = .423 when homogeneous population variances are assumed, whereas AUROC = .770 when heterogeneous population variances are assumed. For the anchoring studies, homogeneous population variances across conditions seem unreasonable and greatly affect the performance of the statistical method. The method that incorporates heterogeneous population variances (i.e., one per anchoring condition) greatly improves the performance of the variance of variances analysis.

Further inspecting the heterogeneous variance of variances analysis indicates that no false positives occur until $\alpha = 0.13$.

Performance of using p-values to detect data fabrication

We asked researchers to fabricate data for nonsignificant effect sizes after inspecting whether the genuine data resembled a uniform distribution. Consequently, However, as Table 2 indicates, the auroc for these methods are hardly better than chance. In other words, for effects that are specifically fabricated to be nonsignificant, researchers might not fabricate such large p-values to be suspicious.

Method	AUROC
Fisher test gender hypothesis	0.521
Fisher test interaction hypothesis	0.535

Table 2: Area Under the Receiver Operating Curve (AUROC) for PLACEHOLDER

Performance of combining SDs and p-value methods to detect data fabrication

Method	AUROC
Combined Fisher test (3 results, SD_z homogeneous)	0.602
Combined Fisher test (3 results, SD_z heterogeneous)	0.736
Combined Fisher test (6 results, SD_z homogeneous)	0.643
Combined Fisher test (6 results, SD_z heterogeneous)	0.643

Table 3: Area Under the Receiver Operating Curve (AUROC) for PLACEHOLDER

Performance of large effect sizes to detect data fabrication

Discussion

References

Akhtar-Danesh, N. and Dehghan-Kooshkghazi, M. (2003). How does correlation structure differ between real and fabricated data-sets? *BMC Medical Research Methodology*, 3(1):1–9.

Altman, D. G. and Bland, J. M. (1995). Statistics notes: Absence of evidence is not evidence of absence. *BMJ*, 311(7003):485–485.

Bailey, K. R. (1991). Detecting fabrication of data in a multicenter collaborative animal study. *Controlled Clinical Trials*, 12(6):741 – 752. Large effects were on of the suspicions.

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

Goodman, S. (2008). A dirty dozen: Twelve p-value misconceptions. *Seminars in Hematology*, 45(3):135 – 140. Interpretation of Quantitative Research.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36. PMID: 7063747.

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.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). pROC: an open-source package for R and s+ to analyze and compare ROC curves. *BMC bioinformatics*, 12:77.

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.

SessionInfo

```
> sessionInfo()

R version 3.3.1 (2016-06-21)
Platform: x86_64-redhat-linux-gnu (64-bit)
Running under: Fedora 23 (Workstation Edition)

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] stringr_1.0.0   plyr_1.8.4      car_2.0-19      httr_1.2.0
 [5] xtable_1.8-2    gridExtra_2.2.1 ggplot2_2.1.0   latex2exp_0.4.0
 [9] foreign_0.8-66  pROC_1.8

loaded via a namespace (and not attached):
 [1] Rcpp_0.12.6      MASS_7.3-45      R6_2.1.2         grid_3.3.1
 [5] gtable_0.2.0     magrittr_1.5     scales_0.4.0     stringi_1.1.1
 [9] tools_3.3.1      munsell_0.4.3    colorspace_1.2-6 nnet_7.3-12

>
```