Ecological performance of detecting data fabrication with summary statistics

Chris HJ Hartgerink, Jan G Voelkel, Jelte M Wicherts, Marcel ALM van Assen
Preregistration conducted by first author

March 3, 2017

Any field of empirical inquiry is faced with cases of scientific misconduct at some point, either in the form of fabrication, falsification or plagiarism (FFP). Psychology was faced with Stapel; medical sciences were faced with Poldermans and Macchiarini; life sciences were faced with Voignet. These are just a few examples of cases in the last decade. Overall, an estimated 2% of all scholars have admitted to falsifying or fabricating research results at least once (Fanelli, 2009) and this is likely to be an underestimate due to socially desirable responses. The detection rate is likely to be even lower; for example, only around a dozen cases are discovered in the United States and the Netherlands, despite covering several hundreds of thousands of researchers. At best, this amounts to a detection rate far below 1% of those 2% who admit to fabricating data — the tip of a seemingly much larger iceberg.

In order to stifle attempts at data fabrication, improved detection of fabricated data is considered to deter such harmful attempts. Although deterrence theory dates back to the middle of the 17th century (Hobbes, 1651), its implementation has not occurred across the different forms of scientific misconduct equilaterally. Basically, deterrence theory stipulates that with increased risk of detection, the utility of scientific misconduct (for this context) will decrease and therefore fewer people will engage in such behaviors. This principle of deterrence has been implemented with plagiarism scanners, a development that already started a long time ago (e.g., Parker and et al., 1989). However, increased deterrence of fabrication and falsification by improved detection mechanisms has not been as widely implemented.

In the last decade, detecting image manipulation has become one of the few forms of detecting scientific misconduct other than plagiarism. The Journal of Cell Biology scans each submitted image for potential manipulation (The Journal of Cell Biology, 2015), which greatly increases the risk of detecting (blatant) image manipulation. More recently, algorithms have been developed to automate the scanning of images for (subtle) manipulations (Koppers et al., 2016). These developments in detecting image manipulation have increased detection risk during the pre-publication and post-publication phase by improving detection mechanisms and increasing the understanding of how images might be manipulated. Moreover, their application also helps researchers systematically evaluate research articles to estimate the extent of the problem of image manipulation (4% of all papers are estimated to contain manipulated images; Bik et al., 2016).

Statistical methods can provide one way to improve detection of data fabrication in empirical research. Humans are notoriously bad at understanding and estimating probabilities (e.g., Tversky and Kahneman, 1974, 1971), which could manifest itself in the fundamentally probabilistic data they try to fabricate. That researchers do not understand probabilistic processes also presents itself in the interpretation of genuine research data (Hoekstra et al., 2006; Sijtsma, 2015; Goodman, 1999). When data are fabricated, probabilistic principles are easily violated if these principles are forgotten at the univariate level, bivariate level, trivatiate level, or beyond (Haldane, 1948). Based on such a theoretical framework, statistical methods that investigate whether the reported data are actually plausible under theoretically probabilistic processes can be used to detect potential data fabrication.

The application of such statistical methods to detect data fabrication has occurred in several cases in recent years and has potential for future application beyond a case-basis. For example, problems with papers by Fuji were highlighted with statistical methods (Carlisle, 2012; Carlisle et al., 2015), resulting in 183 retractions (Oransky, 2015). In this case, baseline measures across randomized groups were examined for too little variation. Random assignment should introduce a certain degree of random error that is might be missed by a human fabricator, misestimating the probabilistic process that generates such error. Another two cases are those of Sanna and Smeesters, where fabricated data were also detected with statistics (Simonsohn, 2013). These cases inspected the variance of variances (i.e., the second level, or meta, variance). Once again, too little variation was what revealed problems in these data. These methods, although developed in a case-setting, need not be limited to cases. The application of such methods can be (semi-)automated if data are available in a machine-readable format that one of the statistical methods can be applied to. An example of such a potential case for mass application of using statistics to detect (potential) data fabrication is in the ClinicalTrials.gov database, where baseline measures across randomized groups are readily available for download and subsequent analysis (Hartgerink and George, 2015).

Nonetheless, considering the potential harm of applying statistical methods to flag potentially problematic results, it needs to be sorted out whether such methods have diagnosticity that actually makes it responsible to apply them. We hardly know how researchers might go about fabricating data. Cases such as Fuji, Smeesters, and Sanna provide some insights, but are highly pre-selected (i.e., those who got caught/confessed) and as such, systematically biased. Relatively extensive descriptions in rare and partial autobiographical accounts provide little insight into the actual data fabrication process, except for the setting where it might take place (e.g., late at night when no one is around; Stapel, 2014). Additionally, the performance of methods to detect data fabrication is highly dependent on the unknown prevalence of data fabrication and the power to actually to detect data fabrication. Given that we do not know how researchers might fabricate data, the diagnosticity of these methods cannot realistically be simulated.

The effectiveness of detection mechanisms and their consequences, hence their expected deterrence, is exacerbated by the increased usage of public online discussion platforms such as PubPeer (`https://pubpeer.com`). PubPeer serves as an ”online journal club” where anyone can discuss articles. Authors are notified when someone leaves a response, providing them with the possibility to respond. Such a platform allows for public discussion of the paper, including discussion of reanalyses, methodology, availability of materials, etc. When detection mechanisms are freely available to use, they can lead to (a surge of) comments when applied by users. Recently, in 2016, one user (the main author of this paper) used ‘statcheck‘ software to report potential statistical reporting inconsistencies for 50,000 psychology articles, which led to a large discussion about (automated) online comments (Baker, 2016). The impact of such new possibilities is not to be underestimated, although its potential to contribute to the scientific discussion should also not be. Nonetheless, the ‘statcheck‘ software was well validated prior to this application (Nuijten et al., 2015) and the same principle applies to the application of statistical methods to detect (potential) data fabrication.

Throughout this paper, we inspect statistical methods to detect data fabrication that can be applied to (1) summary results or (2) raw data. Even though the data available look different depending on the structure of the study, there are certain common characteristics of results and the underlying raw data that can be inspected. For example, summary results frequently include means, standard deviations, test-statistics, and *p*-values. Raw data frequently contain at least some variables measured at a interval- or ratio scale (Stevens, 1946).

We present a set of studies that directly test the validity of statistical methods to detect data fabrication. To this end, Study 1 inspected the performance of statistical methods to detect data fabrication when using only summary results (e.g., means and standard deviations) as typically reported in empirical research articles. Study 2 inspected the performance of statistical methods aimed at detecting data fabrication in raw data (i.e., the data underlying summary results). These two studies provide a first indication of how applicable and effective statistical methods are to detect data fabrication in practice, with actual researchers fabricating actual data.

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. 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 a preregistration is available at `https://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 (`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`.

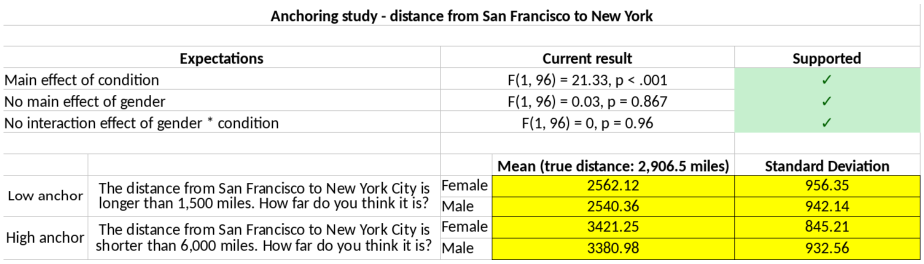
Each respondent was instructed to fabricate 32 summary statistics (4 studies $\times$ 2 conditions $\times$ 2 sexes $\times$ 2 statistics [mean and sd]) that fulfilled three hypotheses. We instructed respondents to fabricate results for the hypotheses (i) main effect of condition, (ii) no effect of sex, 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 fabricated 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)



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 of Study 1. 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 (one of four template spreadsheets available at `https://osf.io/w6v4u/`).

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

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

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).

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.

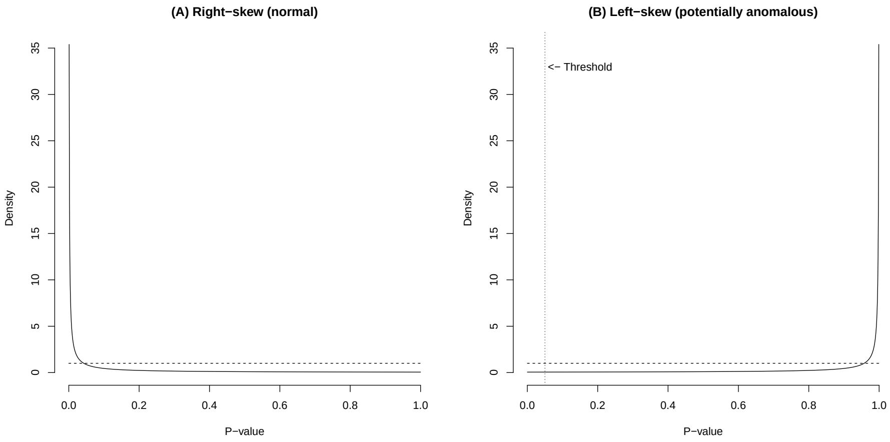


Figure 2: Conceptual representation of what the Fisher test inspects (Equation 3; panel A) and what the adjusted Fisher test inspects (Equation 4; panel B). Both panels test whether there is sufficient evidence that the solid line deviates from the dashed line, except that the type of deviation that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

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. 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

Discussion

Study 2

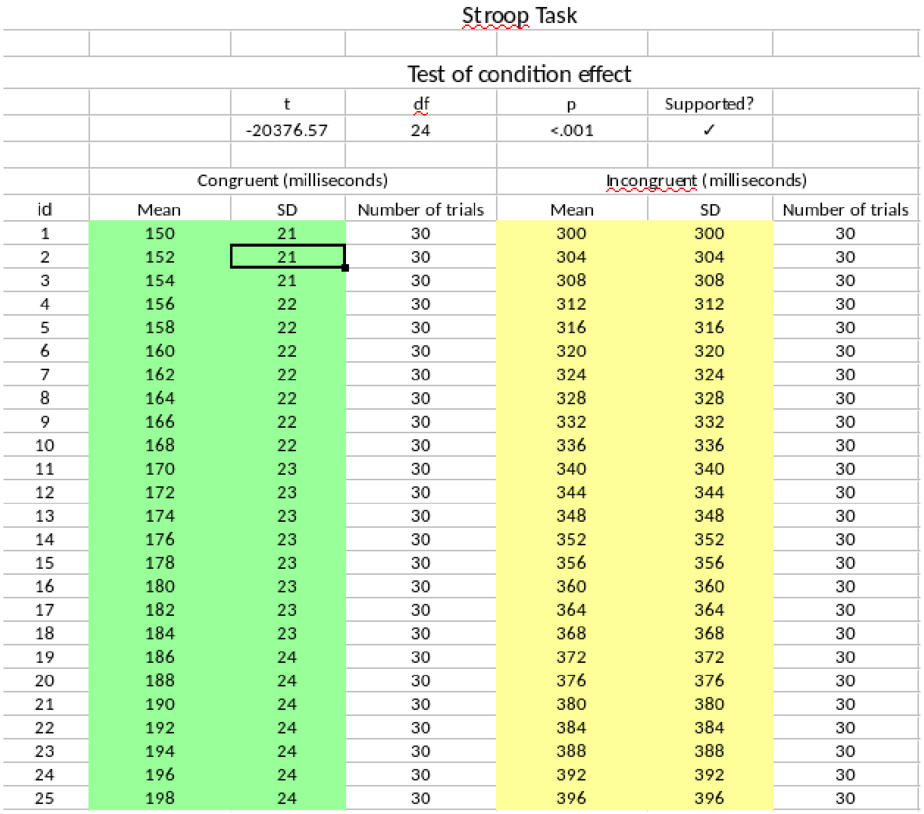
In Study 2 we investigated detecting data fabrication in raw data as an extension of Study 1, which presented results of detecting data fabrication in summary results. In essence, the procedure is similar: we asked actual researchers to fabricate data that they thought would go undetected. For Study 2 we included a face-to-face interview to qualitatively assess how data fabrication occurs. A preregistration of this study occurred during the seeking of funding (Hartgerink et al., 2016) and during data collection (`https://osf.io/XXXX`).

To test the validity of statistical methods to detect data fabrication in raw data, we investigated raw data of a Stroop experiment (Stroop, 1935). In the Stroop task, participants are asked to determine the color a word is presented in (i.e., word colors), but the word also reads a color (i.e., color words). The presented word color (i.e., 'red', 'blue', or 'green') can be either presented in the congruent color (e.g., 'red' presented in red) or an incongruent color (i.e., 'red' presented in green). The dependent variable in the Stroop task is the response latency (in this study milliseconds are used). Participants in actual studies are typically presented with a set of these, where the mean and standard deviation per condition serves as the raw data. The Stroop effect typically is computed as the difference in mean response latencies between the congruent and incongruent conditions.

Methods

Twenty-one genuine datasets on the Stroop task were collected from the Many Labs 3 project (`https://osf.io/n8xa7/`; Ebersole et al., 2016). Many Labs 3 (ML3) includes 20 participant pools from universities and one online sample (the original preregistration mentioned 20 datasets, accidentally overlooking the online sample; Hartgerink et al., 2016). Using the original raw data and analysis script from ML3 (`https://osf.io/qs8tp/`), we computed the mean (M) and standard deviation (SD) for the participant's response latencies in both the within-subjects conditions of congruent trials and incongruent trials. These also formed the basis for the template of the data that needed to be fabricated by the participants (see also Figure 3). The Stroop effect was calculated as a t-test of the difference ($H_0 : \mu = 0$).

We collected twenty-eight faked datasets on the Stroop task experimentally in a two-stage sampling procedure. First, we invited 80 Dutch and Flemish psychology researchers who published a peer-reviewed paper on the Stroop task between 2005-2015 as available in the Thomson Reuters' Web of Science database. We selected Dutch and Flemish researchers to allow for a face-to-face interview on how the data were fabricated. We chose the period 2005-2015 to prevent a drastic decrease in the probability that the corresponding author would still be addressable via the given email. The database was searched on October 10, 2016 and 80 unique e-mails were retrieved from 90 publications. Only two of these 80 participated in the study; we subsequently implemented a second sampling stage where we collected e-mails from all PhD-candidates, teachers, and professors of psychology related departments at Dutch universities. This resulted in 1659 additional unique e-mails that we subsequently invited to participate in this study. Due to a malfunction in Qualtrics' quotum sampling, we



Stroop Task

Test of condition effect

	t	df	p	Supported?
	-20376.57	24	<.001	✓

	Congruent (milliseconds)			Incongruent (milliseconds)		
id	Mean	SD	Number of trials	Mean	SD	Number of trials
1	150	21	30	300	300	30
2	152	21	30	304	304	30
3	154	21	30	308	308	30
4	156	22	30	312	312	30
5	158	22	30	316	316	30
6	160	22	30	320	320	30
7	162	22	30	324	324	30
8	164	22	30	328	328	30
9	166	22	30	332	332	30
10	168	22	30	336	336	30
11	170	23	30	340	340	30
12	172	23	30	344	344	30
13	174	23	30	348	348	30
14	176	23	30	352	352	30
15	178	23	30	356	356	30
16	180	23	30	360	360	30
17	182	23	30	364	364	30
18	184	23	30	368	368	30
19	186	24	30	372	372	30
20	188	24	30	376	376	30
21	190	24	30	380	380	30
22	192	24	30	384	384	30
23	194	24	30	388	388	30
24	196	24	30	392	392	30
25	198	24	30	396	396	30

Figure 3: Example of a filled in template spreadsheet used in the fabrication process for Study 2. Respondents fabricated data in the yellow cells and green cells, which were used to compute the results of the hypothesis test of the condition effect. If the fabricated data confirm the hypotheses, a checkmark appeared. This template is available at `https://osf.io/2qrbs/`.

oversampled, resulting in 28 participants instead of the originally intended 20 participants.

Each participant received instructions on the data fabrication task via Qualtrics but was allowed to fabricate data until the face-to-face interview took place. In other words, each participant could take the time they wanted/needed to fabricate the data as extensively as they liked. Each participant received downloadable instructions (`https://osf.io/7qhy8/`) and the template spreadsheet via Qualtrics (see Figure 3). The interview was scheduled via Qualtrics with JGV, who blinded the rest of the research team from the identifying information of each participant and the date of the interview. All interviews took place between January 31 and March 3, 2017. To incentivize researchers to participate, they received €100 for participation; to incentivize them to fabricate (supposedly) hard to detect data they could win an additional €100 if they belonged to one out of three top fabricators. The contents of the interview will be transcribed for further research on qualitatively assessing how researchers might fabricate experimental data.

The procedure to evaluate the genuine- and fabricated data included digit analyses, variance analyses, comparison of true- and fabricated multivariate associations, and combining results of several of the aforementioned methods with the Fisher method. Figure 4 shows how the various methods are combined. In short, the digit analyses include Benford's law (Benford, 1938) and terminal digit analysis (Mosimann et al., 1995, 2002); analyzing the variance of the standardized SDs (similar to Study 1; Simonsohn, 2013); the multivariate associations compare the observed correlations between the variables in the fabricated data with the observed correlations between the same variables in the genuine data (the multivariate comparison is only done for fabricated data). The Fisher method (Equation 3) is used to combine the results of the terminal digit analyses, the analyses of the standardized SD variance, and the multivariate associations to determine the hardest to detect fabricated datasets.

The digit analyses are based on a simple enough premise: test whether the first (i.e., leading) or final (i.e., terminal) digits follow an a priori specified distribution. Benford's law (Benford, 1938) stipulates that the leading digit follows the distribution of

$$P(d) = log_{10}(\frac{d+1}{d}) \tag{5}$$

where d is the leading digit and $P(d)$ denotes the probability of d (e.g., $P(1) = 0.301$). This expected distribution can be tested with a χ^2-test ($df = 8$) of the tabular count of leading digits. Similarly, terminal digit analysis (Mosimann et al., 1995, 2002) expects that the final digit contains the most measurement error, hence, is uniformly distributed if sufficient digits are available. As a rule of thumb, three digits is sufficient to apply terminal digit analysis (Mosimann et al., 1995, 2002). The observed terminal digit count can be tested against the expected uniform distribution with a χ^2-test as well, although with $df = 9$ instead of $df = 8$ because a leading digit can never be zero.

We analyzed the variance of the standardized SDs for each condition in

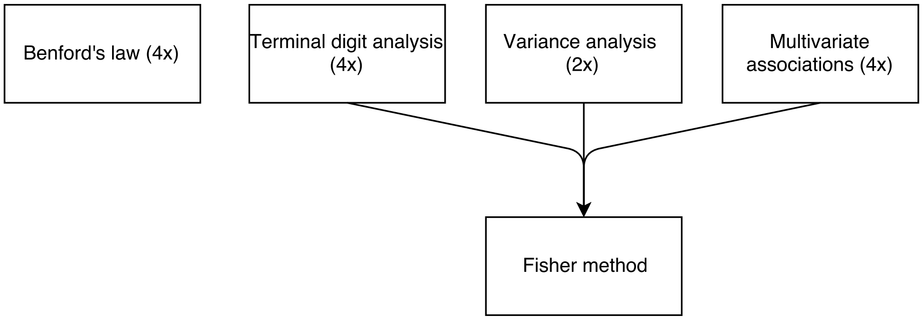


Figure 4: The applied statistical methods to test for data fabrication in Study 2, depicting those that are combined into an overall test for data fabrication with the Fisher method. Benford's law is excluded from the overall tests because of an expected lack of utility. Original from Hartgerink et al. (2016) under CC-BY.

each dataset (i.e., two tests per dataset). Whereas in Study 1 the test was conducted on means and standard deviations across cells per study, in Study 2 we conducted the test on means and standard deviations across respondents per condition. The technical aspects of this test remain equivalent to the application in Study 1 (see equations 1 and 2). As such, the only difference is that N_j is no longer the sample size, but the number of Stroop trials presented to the participant.

We analyzed the multivariate associations between the means and standard deviations of the participant-level data in four ways. We computed the observed multivariate associations between the means across conditions (i.e., $r(M_{congruent},M_{incongruent})$), the standard deviations across conditions (i.e., $r(SD_{congruent},SD_{incongruent})$), and across means and standard deviations within conditions (i.e., $r(M_{congruent},SD_{congruent})$ and $r(M_{incongruent},SD_{incongruent})$). We subsequently compared the observed multivariate correlations from a fabricated dataset with their respective observed multivariate correlations from the genuine data. We computed the probability that the genuine data yielded similar, or more extreme, multivariate correlations (i.e., a p-value). Given 21 genuine datasets are available, only 21 p-values are possible (i.e., lowest p-values possible are 0, 0.048, and 0.095).

We combined the results from the four terminal digit analyses, the two variance analyses, and the four multivariate analyses with the Fisher method. This result was used to determine the top three fabricators, by rank-ordering the p-values for each fabricator (higher p-value equating to less evidence for fabrication). This omnibus method aggregates evidence from the various methods, making it potentially more sensitive for various fabrication behaviors. We evaluated each of the methods to detect data fabrication using AUROC-curves, using the same procedures as in Study 1.

Results

Discussion

General discussion

References

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

Baker, M. (2016). Stat-checking software stirs up psychology. *Nature*, 540(7631):151–152.

Benford, F. (1938). The Law of Anomalous Numbers. *Proceedings of the American Philosophical Society*, 78(4):551–572.

Bik, E. M., Casadevall, A., and Fang, F. C. (2016). The prevalence of inappropriate image duplication in biomedical research publications. *mBio*, 7(3).

Carlisle, J. B. (2012). The analysis of 168 randomised controlled trials to test data integrity. *Anaesthesia*, 67(5):521–537.

Carlisle, J. B., Dexter, F., Pandit, J. J., Shafer, S. L., and Yentis, S. M. (2015). Calculating the probability of random sampling for continuous variables in submitted or published randomised controlled trials. *Anaesthesia*, 70(7):848–858.

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., Baranski, E., Bernstein, M. J., Bonfiglio, D. B., Boucher, L., Brown, E. R., Budiman, N. I., Cairo, A. H., Capaldi, C. A., Chartier, C. R., Chung, J. M., Cicero, D. C., Coleman, J. A., Conway, J. G., Davis, W. E., Devos, T., Fletcher, M. M., German, K., Grahe, J. E., Hermann, A. D., Hicks, J. A., Honeycutt, N., Humphrey, B., Janus, M., Johnson, D. J., Joy-Gaba, J. A., Juzeler, H., Keres, A., Kinney, D., Kirshenbaum, J., Klein, R. A., Lucas, R. E., Lustgraaf, C. J., Martin, D., Menon, M., Metzger, M., Moloney, J. M., Morse, P. J., Prislin, R., Razza, T., Re, D. E., Rule, N. O., Sacco, D. F., Sauerberger, K., Shrider, E., Shultz, M., Siemsen, C., Sobocko, K., Sternglanz, R. W., Summerville, A., Tskhay, K. O., van Allen, Z., Vaughn, L. A., Walker, R. J., Weinberg, A., Wilson, J. P., Wirth, J. H., Wortman, J., and Nosek, B. A. (2016). Many labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, 67:68 – 82. Special Issue: Confirmatory.

Fanelli, D. (2009). How many scientists fabricate and falsify research? a systematic review and meta-analysis of survey data. *PloS one*, 4(5):e5738.

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.

Goodman, S. N. (1999). Toward evidence-based medical statistics. 1: The p value fallacy. *Annals of Internal Medicine*, 130(12):995.

Haldane, J. B. S. (1948). The faking of genetical results. *Eureka*, 6:21–28.

Hartgerink, C. and George, S. (2015). Problematic trial detection in ClinicalTrials.gov. *Research Ideas and Outcomes*, 1:e7462.

Hartgerink, C., Wicherts, J., and van Assen, M. (2016). The value of statistical tools to detect data fabrication. *Research Ideas and Outcomes*, 2:e8860.

Hobbes, T. (1909/1651). *Leviathan.* Oxford University Presss. http://www.gutenberg.org/ebooks/3207.

Hoekstra, R., Finch, S., Kiers, H. A. L., and Johnson, A. (2006). Probability as certainty: Dichotomous thinking and the misuse of p values. *Psychonomic Bulletin & Review*, 13(6):1033–1037.

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.

Koppers, L., Wormer, H., and Ickstadt, K. (2016). Towards a systematic screening tool for quality assurance and semiautomatic fraud detection for images in the life sciences. *Science and Engineering Ethics*, pages 1–16.

Mosimann, J., Dahlberg, J., Davidian, N., and Krueger, J. (2002). Terminal digits and the examination of questioned data. *Accountability in research*, 9(2):75–92.

Mosimann, J. E., Wiseman, C. V., and Edelman, R. E. (1995). Data fabrication: Can people generate random digits? *Accountability in research*, 4(1):31–55.

Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., and Wicherts, J. M. (2015). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, 48(4):1205–1226.

Oransky, I. (2015). The Retraction Watch Leaderboard.

Parker, A. and et al. (1989). Computer algorithms for plagiarism detection.

Sijtsma, K. (2015). Playing with Data-Or How to Discourage Questionable Research Practices and Stimulate Researchers to Do Things Right. *Psychometrika*.

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

Stapel, D. (2014). *Ontsporing [Derailment]*.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103(2684):677–680.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18(6):643–662.

The Journal of Cell Biology (2015). About the Journal. `https://web.archive.org/web/20150911132421/http://jcb.rupress.org/site/misc/about.xhtml`. Accessed: 2015-9-11.

Tversky, A. and Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, 76:105–110.

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

SessionInfo

```
> sessionInfo()

R version 3.3.2 (2016-10-31)
Platform: x86_64-redhat-linux-gnu (64-bit)
Running under: Fedora 25 (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_1.1.0     plyr_1.8.4        reshape2_1.4.2    dplyr_0.5.0
 [5] data.table_1.10.0 lsr_0.5           effects_3.1-2     car_2.0-19
 [9] httr_1.2.1        xtable_1.7-1      gridExtra_2.2.1   ggplot2_2.2.1
[13] latex2exp_0.4.0   foreign_0.8-67    pROC_1.8

loaded via a namespace (and not attached):
 [1] Rcpp_0.12.9      magrittr_1.5     splines_3.3.2    MASS_7.3-45
 [5] munsell_0.4.3    colorspace_1.3-2 lattice_0.20-34  R6_2.2.0
 [9] minqa_1.2.4      tools_3.3.2      nnet_7.3-12      grid_3.3.2
[13] nlme_3.1-128     gtable_0.2.0     DBI_0.5-1        lme4_1.1-12
[17] lazyeval_0.2.0   assertthat_0.1   tibble_1.2       Matrix_1.2-7.1
[21] nloptr_1.0.4     stringi_1.1.2    scales_0.4.1

>
```