Automated detection of data fabrication using statistical tools

Chris HJ Hartgerink, Jan G Voelkel, Jelte M Wicherts, Marcel ALM van Assen

25 July, 2017

Contents

Abstract **1**
Keywords 1

Introduction **1**

Theoretical framework **3**
Detecting data fabrication in summary statistics 4
Detecting data fabrication in raw data 5

Study 1 - detecting fabricated anchoring effects **7**
Methods 9
Results 11

Study 2 - detecting fabricated Stroop data **14**
Methods 14
Results 17
Discussion 19

Session info **19**

References **21**

Abstract

PLACEHOLDER

Keywords

PLACEHOLDER

Introduction

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 misconduct 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), which is likely to be an underestimate due to socially desirable responses. The detection rate of data fabrication is likely to be even lower; for example, only around a dozen cases become public in the United States and the Netherlands, despite covering several

hundreds of thousands of researchers. At best, this suggests 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. This idea is based on deterrence theory (Hobbes, 1651), which stipulates that with increased risk of detection, the utility of scientific misconduct will decrease and therefore fewer people will engage in such behaviors. Implementation of deterrence has not occurred equally across fabrication, falsification, and plagiarism. For plagiarism, it has been implemented with plagiarism scanners, a development that already started a long time ago (e.g., A. Parker & Hamblen, 1989). For data fabrication, detecting image manipulation has increasingly become possible and implemented. The Journal of Cell Biology and the EMBO journal scan 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, Wormer, & Ickstadt, 2016). These developments in detecting image manipulation have increased detection risk during the pre-publication and post-publication phase and increase 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, Casadevall, & Fang, 2016).

Statistical methods provide one way to improve detection of data fabrication in empirical research. Humans are notoriously bad at understanding and estimating randomness (**???**; Amos Tversky & Kahneman, 1971; A. Tversky & Kahneman, 1974), which could manifest itself in the fundamentally probabilistic data they try to fabricate. When data are fabricated, principles of statistics and randomness are easily violated at the univariate level, bivariate level, trivatiate level, or beyond (Haldane, 1948). Based on this idea, statistical methods that investigate whether the reported data are feasible under the theoretically probabilistic processes can be used to detect potential data fabrication.

Statistical methods to detect data fabrication have been applied in several cases of scientific misconduct in recent years and has potential for future application beyond these specific cases. For example, excessive consistency in papers by Fuji were highlighted with statistical methods (Carlisle, 2012; Carlisle, Dexter, Pandit, Shafer, & Yentis, 2015), resulting in 183 retractions (Oransky, 2015). Carlisle's method works as follows: in true randomized clinical trials (RCTs) baseline measurements should be identically distributed across groups. As such, the p-values for group comparisons would be expected to be uniformly distributed because the null hypothesis of identical distributions across groups is true by definition of the randomized design. In the Fuji papers, group comparisons showed excessive consistency, resulting primarily in high p-values (e.g., .99, .95) and a high mean p-value across the comparisons, where a mean p-value of .5 is expected. As an illustration, see Table 1, which depicts 10 hypothetical studies containing 100 participants per condition, for true randomized designs (Set 1) or for fabricated designs (Set 2). The mean p-value for the true randomized design Set 1 is 0.575, whereas the fabricated Set 2 has mean p-value 0.956. Other statistical methods to detect data fabrication are addressed in the theoretical sections. .

Table 1: Examples of means and standard deviations for a continuous outcome in genuine- and fabricated randomized clinical trials. Set 1 (S1) is randomly generated data under the null hypothesis of random assignment (assumed to be the genuine process), whereas Set 2 (S2) is generated under excessive consistency with equal groups. Each trial condition contains 100 participants. The p-values are the result of independent t-tests comparing the experimental and control conditions within each respective set.

Study	M_E (SD_E) [S1]	M_C (SD_C) [S1]	P-value [S1]	M_E (SD_E) [S2]	M_C (SD_C) [S2]	P-value [S2]
Study 1	48.432 (10.044)	49.158 (9.138)	0.594	52.274 (10.475)	63.872 (10.684)	0.918
Study 2	50.412 (10.322)	49.925 (9.777)	0.732	62.446 (10.454)	60.899 (10.398)	0.989
Study 3	51.546 (9.602)	51.336 (9.479)	0.877	62.185 (10.239)	55.655 (10.457)	0.951
Study 4	49.919 (10.503)	50.857 (9.513)	0.509	62.468 (10.06)	68.469 (10.761)	0.956

Study	M_E (SD_E) [S1]	M_C (SD_C) [S1]	P-value [S1]	M_E (SD_E) [S2]	M_C (SD_C) [S2]	P-value [S2]
Study 5	49.782 (11.167)	50.308 (8.989)	0.714	67.218 (10.328)	55.846 (10.272)	0.915
Study 6	48.631 (9.289)	49.29 (10.003)	0.630	62.806 (11.216)	66.746 (11.14)	0.975
Study 7	49.121 (9.191)	47.756 (10.095)	0.318	50.19 (10.789)	55.724 (10.302)	0.960
Study 8	49.992 (9.849)	51.651 (10.425)	0.249	54.651 (11.372)	55.336 (10.388)	0.995
Study 9	50.181 (9.236)	51.292 (10.756)	0.434	63.322 (11.247)	53.734 (11.488)	0.941
Study 10	49.323 (10.414)	49.879 (9.577)	0.695	60.285 (10.069)	54.645 (11.211)	0.960

Statistical methods to detect data fabrication, although developed to quantify suspicions in a specific paper, could be applied to screen multiple papers. 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 & George, 2015).

Nonetheless, prior to applying statistical methods to flag potentially problematic results, investigating whether such methods perform well enough in detecting data fabrication is required for responsible application. We hardly know how researchers might go about fabricating data. Cases such as Stapel, Fuji, Smeesters, and Sanna provide some insights, but are highly pre-selected (i.e., those who got caught/confessed) and as such, may be systematically biased. Relatively extensive descriptions in rare and partial autobiogrophical 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 detect data fabrication. Given that we do not know how researchers might fabricate data, the diagnosticity of these methods cannot realistically be assessed.

Throughout this paper, we inspect statistical methods to detect potential data fabrication that can be applied to summary statistics (Study 1) or raw data (Study 2). Even though structure and contents of data can look different depending on the structure of a study and the measures, there are certain common characteristics of empirical results and the underlying raw data that can be inspected. For example, summary statistics 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). Such common characteristics allow for the development of generic statistical methods that can be applied across a varied set of results to screen for problematic data. For each study, we review the theory of the specific methods we apply. However, the reviewed methods are not exhaustive of all methods available to test for potential data fabrication in empirical data (Anaya, 2016; Brown & Heathers, 2016; see also, Buyse et al., 1999; James Heathers, 2017).

Theoretical framework

In the current paper, we differentiate between statistical methods to detect potential data fabrication based on reported summary statistics or raw data. For summary statistics, we review p-value analysis, variance analysis, and effect size analysis as potential ways to detect data fabrication. P-value analysis can be applied whenever sufficient nonsignificant p-values are reported; variance analysis can be applied whenever a set of variances are reported for independent groups alongside the sample sizes per group; effect size analysis can be used whenever the effect size is reported or can be computed [e.g., an APA reported t- or F-statistic;@10.1525/collabra.71]. For raw data, we review digit analyses (i.e., the Newcomb-Benford law and terminal digit analysis) and multivariate associations as potential ways to detect data fabrication. Both types of digit analyses can be applied when ratio scale measures are present in the raw data and only terminal digit analysis can be applied when there are continuous measures. Multivariate associations can be investigated whenever there are two or more variables and data on that same relation is available from (assumably) genuine data sources.

Detecting data fabrication in summary statistics

P-value analysis

The distribution of a p-value is uniform if the null hypothesis is true and right-skewed if the alternative hypothesis is true (Fisher, 1925). The distribution of one p-value is the result of the population effect size, the precision of the estimate, and the observed effect size, whose properties carry over to a set of independent p-values if those p-values are independent. As such, the p-value distribution of a set of p-values is uniform when the null hypothesis is true, or right-skewed when the alternative hypothesis is true.

When a p-value distribution of independent p-values is not uniform or right-skewed, as would be theoretically expected, it can indicate potential data fabrication. For example, in the Fuji case, supposedly identically distributed groups were fabricated. However, this resulted in excessively large p-values and ultimately the identification of potential data fabrication by Carlisle (2012). Our illustration in Table 1 also shows the difference between expected data under the null distribution (Set 1) and what might result from excessively consistent and fabricated data (Set 2). More specifically, the expected value of a uniform p-value distribution is .5, but the fabricated data from our illustration result in a mean p-value of 0.956.

In order to test whether the distribution of nonsignificant and independent p-values might be fabricated, we proposed an adaptation of Fisher's method (S. P. O'Brien et al., 2016). This adaptation is a reversal of the original Fisher method (Fisher, 1925), which was introduced as a meta-analytic test for the presence of an effect. This test is computed as

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

where it tests for more smaller p-values than larger p-values across the k number of p-values. Reversing the original Fisher method, which tests for right-skew, into a Fisher method that tests for left-skew results in

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

where it now tests for more larger p-values than smaller p-values across the k number of p-values that fall above the threshold t. When the threshold is set to zero, it selects all p-values, but when this $t > 0$ the remaining p-values are rescaled to fit the original 0-1 range (i.e., by dividing by $1 - t$). Upon writing this paper, it became clear to us that this reversed Fisher method is similar to the operating principle of Carlisle's method testing for excessive homogeneity across baseline measurements in RCTs (Carlisle, 2012, 2017; Carlisle et al., 2015), which we had not realized before.

For example, this method can be applied to both the genuine- and fabricated results from our illustration in Table 1. Using the threshold $t = 0.05$ to only select the nonsignificant results, we retain $k = 10$ genuine p-values and $k = 10$ fabricated p-values. Calculating the χ^2 value for sets of k nonsigificant p-values results in a $\chi^2 = 0$ for the p-values resulting from the genuine data and $\chi^2 = 0$ for the p-values resulting from the fabricated data, with p-values 1 and 1 respectively, using $2k$ as the degrees of freedom of the χ^2-test. Another example, directly from the Fuji case (Carlisle, 2012), anecdotally illustrates that fabricated data can result in significant findings with the reversed Fisher method. p-values from five independent comparisons for one dependent measure (fentanyl dose, as extracted from Table 3 in Carlisle, 2012) show excessively high p-values, $\chi^2_{10} = 0, p = 1$.

Despite this test being useful for detecting data fabrication in nonsignificant p-values, one exception should be taken into account: wrongly specified one-tailed tests. For properly specified one-tailed tests, the p-value distribution is right-skewed. When wrongly specified, this distribution is reversed and becomes left-skew. As such, any data fabrication detected with this method would need to be inspected for misspecified one-tailed hypotheses to preclude false conclusions.

Variance analysis

Variance- or standard deviation estimates are typically reported to indicate dispersion, but just like the mean there should be sampling error in this estimate proportional to the sample size [i.e., $\sigma/\sqrt{2n}$ under the assumption of normality, p. 351;@yule1922]. A variance estimate follows a χ^2-distribution, which is dependent on the sample size (p. 445; Hogg & Tanis, 2001); that is

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

where N_j is the sample size of the jth group and MS_w is the normalizing constant resulting in a standardized variance z_j^2. The normalizing constant MS_w is computed as

$$MS_w = \frac{\sum_{j=1}^{k} (N_j - 1) s_j^2}{\sum_{j=1}^{k} (N_j - 1)}$$

where s_j^2 is the variance in the jth group.

The observed dispersion of the variances can be compared to various measures of expected dispersion in the variances. Dispersion can be operationalized in various ways, such as the standard deviation of the variances (denoted in this paper as SD_z; Simonsohn, 2013) or as the range of the variances (denoted as $max - min_z$). Too consistent results would indicate potential fabrication in the reported data. For example, in the Smeesters case three independent conditions from the same study ($n_k = 15$) were reported to have standard deviations 25.09, 24.58, and 25.65. The standard deviation of the standard deviations here is 0.54 (i.e., SD_z). Such consistency (or more consistency) would only be observed in 1.23% of 100,000 simulated replications (Simonsohn, 2013).

Effect size analysis

From our own experience, and anecdotal evidence elsewhere (Bailey, 1991), large effects have previously raised initial suspicions. Taking the observed effect size and transforming it into a correlation, allows for an easy way to assess how extreme the presented result is. One minus the observed correlation can be used as a measure for extreme effects (i.e., $1 - r$); as a heuristic, it can be regarded as a p-value. That is, this measure too ranges from zero to one and the more extreme the effect size, the smaller the value. This method specifically looks at situations where fabricators would want to fabricate the existence of an effect (not the absence of one).

Detecting data fabrication in raw data

Digit analysis

Raw data with ratio- or interval measures can be subjected to digit analysis under specific conditions. More specifically, the properties of the leading (first) digit (e.g., the 1 in 123.45) or the terminal (last) digit (e.g., the 5 in 123.45) can be examined. By analyzing these leading- and terminal digits for deviations from specific digit distributions, it might be possible to screen for problematic data. In this article we focus on leading digit analysis (i.e., Newcomb-Benford Law) and terminal digit analysis to detect potentially problematic data.

Newcomb-Benford law

The Newcomb-Benford law (Benford, 1938; NBL; Newcomb, 1881) states that leading digits do not have an equal probability of occuring under certain conditions. A leading digit is the left-most digit of a numeric

value, where a digit is any of the nine natural numbers $(1, 2, 3, ..., 9)$. The distribution of the leading digit, according to the NBL is

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

where d is the natural number of the leading digit and $P(d)$ is the probability of d occurring. Table 2 indicates the expected leading digit distribution based on the NBL. This expected distribution is typically compared to the observed distribution with a χ^2-test ($df = 9 - 1$), which requires a minimum of 45 observations based on the rule of thumb outlined by Agresti (2003) ($n = I \times J \times 5$, with I rows and J columns). The NBL has been applied to detect financial fraud (e.g., Cho & Gaines, 2007), voting fraud (e.g., Durtschi, Hillison, & Pacini, 2004), and also to detect problems in scientific data (e.g., **???**).

Table 2: The expected first digit distribution, based on the Newcomb-Benford Law.

Digit	Proportion
1	0.301
2	0.176
3	0.125
4	0.097
5	0.079
6	0.067
7	0.058
8	0.051
9	0.046

However, given that the NBL only applies under specific conditions that are rarely fulfilled in the social sciences, its applicability for detecting data fabrication in science can be questioned. First, the NBL only applies for true ratio scale measures (**???**; Hill, 1995). Second, sufficient range on the measure is required for the NBL to apply (i.e., range from $1 - 1000000$; **???**). Third, these measures should not be subject to digit preferences, for example due to psychological preferences for rounded numbers. Fourth, any form of truncation undermines the NBL (Nigrini, 2015). Moreover, some research has even indicated humans might be sensitive to fabricating data that are in line with the NBL (**???**; Burns, 2009), immediately undermining the applicability of the NBL. Nonetheless, considering the four conditions for the NBL to apply, we preregistered that this method would not prove fruitful (**???**).

Terminal digit analysis

Terminal digit analysis is based on the principle that the rightmost digit is the most random digit of a number, hence, is expected to be uniformly distributed under specific conditions (**???**, **???**). Terminal digit analysis is conducted with a χ^2-test ($df = 10 - 1$) on the digit occurrence counts (including zero), where the observed frequencies are compared with the expected uniform frequencies. Tthe rule of thumb outlined by Agresti (2003) indicates at least 50 observations are required to provide a meaningful test of the terminal digit distribution ($n = I \times J \times 5$, with I rows and J columns). Terminal digit analysis was developed during the Imanishi-Kari case by (**???**; for a history of this decade long case, see Kevles, 2000).

As an example, Figure 1 depicts the digit counts for the first- through fifth digit of a random, normally distributed variable. The first- and second digit distributions are clearly non-uniform, whereas the third-, fourth-, and fifth digit distributions are uniformly distributed.

As such, the rightmost digit can be expected to be uniformly distributed if sufficient precision is provided (**???**). For our purposes, sufficient precision is determined as the terminal digit being at least the third leading digit [i.e., minimally 1.23 or 12.3 or 123].

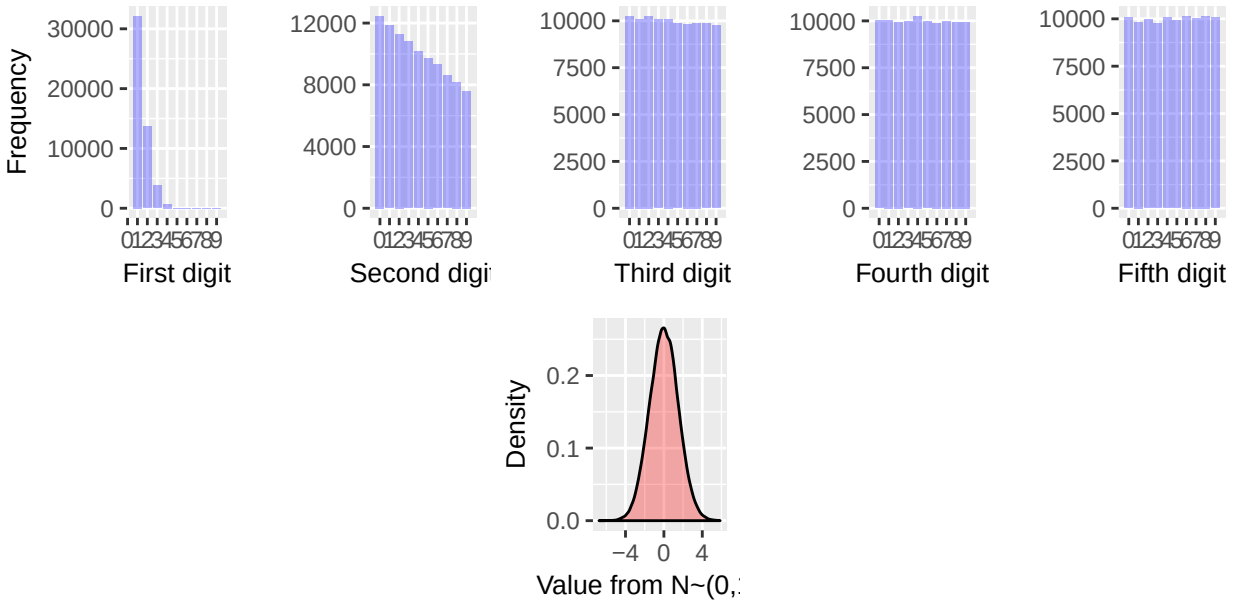


Figure 1: Illustration of how digit distributions evolve from first- through later digits. We sampled 100,000 values from a normal distribution, N (0, 1.5), to create these digit distributions.

Multivariate associations

True data occur within a web of relations, which can be observed in genuine data and easily forgotten while fabricating data. The multivariate relations between different variables arise from stochastic processes and are not readily known, hence difficult to take into account when someone wants to fabricate data. As such, using these multivariate associations to detect fabrication from genuine data might prove valuable.

The multivariate associations between different variables can be estimated from control data that are (assumably) genuine. For example, if the multivariate association between means (Ms) and standard deviations (SDs) is of interest, control data for that same measure can be collected from the literature, assuming the measure has been used in other studies. With these control data, a meta-analysis provides an overall estimate of the multivariate relation.

The multivariate relation from the genuine data is subsequently used to estimate how extreme the observed multivariate relation is. Consider the following fictitious example, regarding the multivariate association between Ms and SDs for a response latency task. Figure 2 depicts a simulated distribution of the association between Ms and SDs from the literature. The observed relation between Ms and SDs from two papers we want to (fictitiously) screen are 0.5 and 0.2. As such, we immediately see in Figure 2 that the former is flagged as being potentially anomalous (i.e., the red dot; two-tailed p-value 1.0274422×10^{-4} of the time), whereas the latter (blue dot) is not flagged (p-value: 1.0274422×10^{-4}).

Study 1 - detecting fabricated anchoring effects

We tested the performance of statistical methods to detect data fabrication in summary statistics with genuine- and fabricated summary statistics from four anchoring studies (Jacowitz & Kahneman, 1995; A. Tversky & Kahneman, 1974). 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 (A. Tversky & Kahneman, 1974), despite essentially posing the same factual question. A considerable amount of genuine

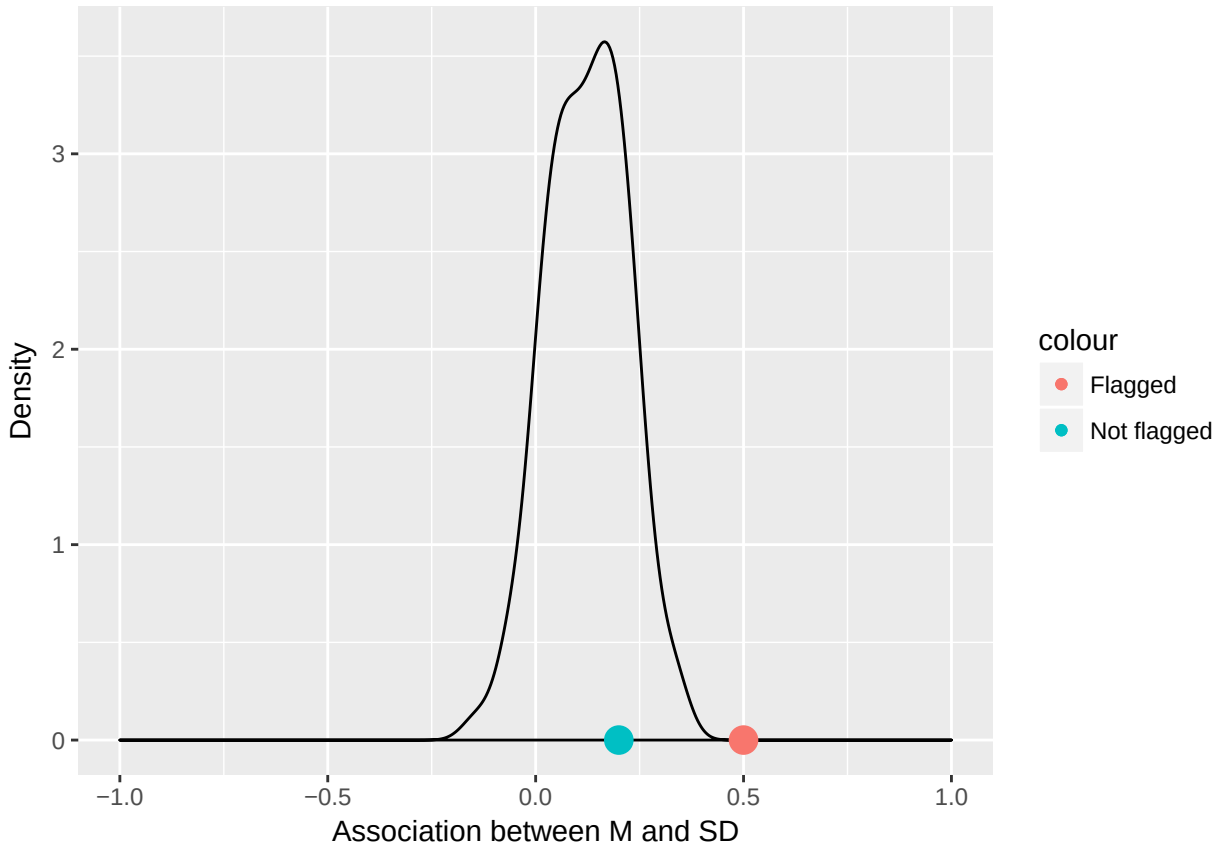


Figure 2: A fictitious distribution of observed association between Ms and SDs across 100 studies. The blue dot indicates the observed relation that is subject to screening for data fabrication.

datasets on this heuristic are freely available and we collected fabricated datasets within this study. This study was approved by the Tilburg Ethical Review Board (EC-2015.50).

Methods

We collected summary statistics for four anchoring studies: (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 (Jacowitz & Kahneman, 1995). Each of the four studies provided us with summary statistics 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 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 fabricating data, we assumed these data to be genuine. For each of the thirty-six locations we computed sample sizes, means, and standard deviations for each of the four conditions in the four anchoring studies (i.e., $3 \times 4 \times 4$) for each of the thirty-six locations. 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 statistics for the same four anchoring studies. Quotum sampling was used 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 Web of Science (WoS) with the filter set to the U.S. We sampled psychology researchers to improve familiarity with the anchoring effect (Jacowitz & Kahneman, 1995; A. Tversky & Kahneman, 1974), for which summary statistics 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 the WoS 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).

We invited a random sample of 1,000 researchers via e-mail 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 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 e-mail 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 $\times$ 2 conditions $\times$ 2 sexes $\times$ 2 statistics [mean and sd]) that fulfilled three hypotheses. We instructed respondents to fabricate results for the following hypotheses: there is (i) a main effect of condition, (ii) no effect of sex, and (iii) no interaction effect between condition and sex. We fixed the sample sizes to 25 per cell; respondents did not need to fabricate sample sizes. The fabricated summary statistics and their accompanying test results for these three hypotheses serve as the data to examine the properties of statistical tools to detect data fabrication.

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 3: 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/).

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 2 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 2. 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

We analyzed the genuine- and fabricated datasets for the four anchoring studies in four ways. First, we applied variance analyses to the reported variances of each of the four groups per study separately. Second, we applied the reversed Fisher method to the results of the gender and interaction hypotheses (i.e., nonsignificant results) across the four studies. Third, we combined the results from the variance analyses and the reversed Fisher method, using the original Fisher method (Fisher, 1925). Fourth, and not preregistered, we used effect size analysis (i.e., $1 - r$) that is a proxy of how extreme an effect is.

Specifically for the variance analyses, we deviated from the preregistration. Initially, we simultaneously analyzed the reported variances per study across the anchoring conditions. However, upon analyzing these values, we realized that the variance analyses assume that the reported variances are from the same population distribution, which is not necessarily the case for the anchoring conditions. Hence, we included two variance analyses per anchoring study (i.e., one for the high anchoring condition and one for the low anchoring condition). In the results we differentiate between these by using 'homogeneous' (across conditions) and 'heterogeneous' (separated for low- and high anchoring conditions).

For each of these statistical tests to detect data fabrication we carried out sensitivity and specificity analyses using Area Under Receiving Operator Characteristic (AUROC) curves. AUROC-analyses indicate the sensitivity (i.e., True Positive Rate [TPR]) and specificity (i.e., True Negative Rate [TNR]) for various decision criteria (e.g., $\alpha = 0, .01, .02, ..., .99, 1$). AUROC values indicate the probability that a randomly drawn fabricated- and genuine dataset can be correctly classified as fabricated and genuine (Hanley & McNeil,

1982). In other words, if $AUROC = .5$, correctly classifying a randomly drawn dataset in this sample is equal to a coin flip. For this setting, we follow the guidelines of (**???**) and regard any AUROC value $< .7$ as poor for detecting data fabrication, $.7 \leq$ AUROC $< .8$ as fair, $.8 \leq$ AUROC $< .9$ as good, and AUROC $\geq .9$ as excellent.

PROC package 10.1186/1471-2105-12-77

Results

```
##   variance_sd_p_overall_homo         variance_sd_p_study1
##                    0.2553419                    0.3725071
##         variance_sd_p_study2         variance_sd_p_study3
##                    0.3945869                    0.4971510
##         variance_sd_p_study4        maxmin_p_overall_homo
##                    0.4052707                    0.2496439
##              maxmin_p_study1              maxmin_p_study2
##                    0.3746439                    0.4138177
##              maxmin_p_study3              maxmin_p_study4
##                    0.5028490                    0.4298433
## variance_sd_p_overall_hetero     variance_sd_p_study1_low
##                    0.7592593                    0.6428063
##    variance_sd_p_study1_high     variance_sd_p_study2_low
##                    0.4380342                    0.7467949
##    variance_sd_p_study2_high     variance_sd_p_study3_low
##                    0.6153846                    0.6666667
##    variance_sd_p_study3_high     variance_sd_p_study4_low
##                    0.6502849                    0.7998575
##    variance_sd_p_study4_high      maxmin_p_overall_hetero
##                    0.5562678                    0.7475071
##          maxmin_p_study1_low         maxmin_p_study1_high
##                    0.6428063                    0.4380342
##          maxmin_p_study2_low         maxmin_p_study2_high
##                    0.7467949                    0.6153846
##          maxmin_p_study3_low         maxmin_p_study3_high
##                    0.6666667                    0.6502849
##          maxmin_p_study4_low         maxmin_p_study4_high
##                    0.7998575                    0.5562678
##                gender_fish_p           interaction_fish_p
##                    0.5180180                    0.5330330
##        fish_combine_3_homo_p      fish_combine_3_hetero_p
##                    0.6013514                    0.7334835
##        fish_combine_6_homo_p     fish_combine_10_hetero_p
##                    0.6456456                    0.7710210
##                        es_p
##                    0.7428775
```

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). These group-level comparisons provide an overview of the differences between the genuine- and fabricated data (see also Akhtar-Danesh & 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). However, there seem to be large group differences when we required subjects to fabricate significant data (i.e., condition hypothesis). Considering this, we also investigated how well effect sizes

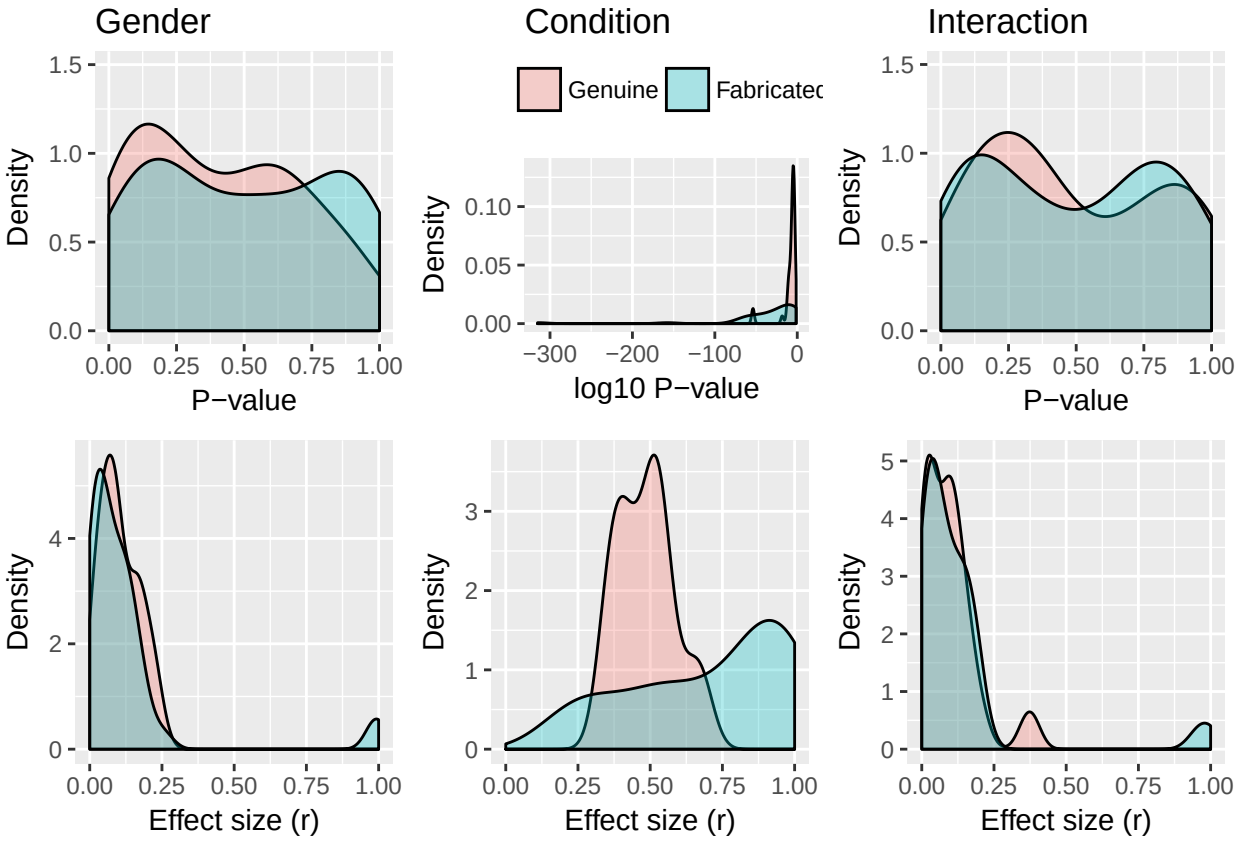


Figure 4: Overlay of density distributions for both genuine and fabricated data, per effect and type of result. We instructed respondents to fabricate nonsignificant data for the gender and interaction effects, and a significant effect for the condition effect.

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 variance analysis to detect data fabrication

Table 3 indicates that both operationalizations (i.e., SD_z and $max - min_z$) show similar performance based on the AUROC. All in all, their performance ranges from 0.303 through 0.796. As such, there is considerable variation for the various applications of the variance analyses.

Table 3: *Table XX.* Diagnosticity of using variance analyses to detect data fabrication, depicted with the AUROC-value.

Method	AUROC SD_z	AUROC $max - min_z$
Homogeneous, all studies combined	0.423	0.303
Homogeneous, study 1	0.367	0.374
Homogeneous, study 2	0.421	0.446
Homogeneous, study 3	0.510	0.520
Homogeneous, study 4	0.540	0.542
Heterogeneous, all studies combined	0.770	0.758
Heterogeneous study 1, low anchor condition	0.644	0.644
Heterogeneous study 1, high anchor condition	0.438	0.438

Method	AUROC SD_z	AUROC $max - min_z$
Heterogeneous study 2, low anchor condition	0.750	0.750
Heterogeneous study 2, high anchor condition	0.614	0.614
Heterogeneous study 3, low anchor condition	0.667	0.667
Heterogeneous study 3, high anchor condition	0.650	0.650
Heterogeneous study 4, low anchor condition	0.796	0.796
Heterogeneous study 4, high anchor condition	0.556	0.556

Combining the variance analyses across the different studies improves performance. This is as expected, considering that the sample size increases for the analyses (i.e., more reported variances are included) and that causes an increase in the statistical power to detect data fabrication.

More notably, combining the studies and taking the heterogeneous approach (i.e., separating anchoring conditions) greatly increases the performance to detect data fabrication considerably. Where the AUROC under homogeneous variances for $SD_z = 0.423$ ($max - min_z = 0.303$), under heterogeneous variances it increases to $SD_z = 0.77$ ($max - min_z = 0.758$). Further inspecting the heterogeneous variance of variances analysis indicates that no false positives occur until $\alpha = 0.13$, making this the optimal alpha level based on this sample (but note the small sample).

Performance of p-values analysis to detect data fabrication

Table 4 indicates that methods using nonsignificant p-values to detect data fabrication are hardly better than chance level in the current sample. We asked researchers to fabricate data for nonsignificant effect sizes, thinking they might be unable to produce uniformly distributed p-values. However, these results (and the density plot in Figure XX) indicate that widespread detection based on this is not promising.

Table 4: *Table XX.* Diagnosticity of using p-value analyses to detect data fabrication, depicted with the AUROC-value.

Method	AUROC
Reversed Fisher method gender hypothesis	0.521
Reversed Fisher method interaction hypothesis	0.535

Performance of combining variance- and p-value analyis to detect data fabrication

Table 5 indicates that combining the variance- and p-value methods provides little beyond the methods separately. The overall performance of this combination is driven by the variance analyses, given that the p-value analysis yields little more than chance classification. When combining the results from variance analyses per anchoring condition (i.e., 10 results, SD_z heterogeneous) and the p-value analyses, a minor improvement occurs over the heterogeneous variance analysis (all studies combined, $AUROC = 0.77$). However, this difference is negligible and potentially due to sampling error.

Table 5: *Table XX.* Diagnosticity of combining variance- and p-value analyses to detect data fabrication, depicted with the AUROC-value.

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 (10 results, SD_z heterogeneous)	0.771

Performance of extreme effects to detect data fabrication

Table 6 indicates that using effect sizes (i.e., $1 - r$) is a simple but effective way to detect data fabrication ($AUROC = 0.744$). Compared to the variance analyses several sections ago, its performance in this sample is a bit worse (i.e., 0.744 compared to 0.77). However, it makes up for this by computational parsimony. Whereas the variance analyses require a considerable amount of effort to implement, computing the correlation and taking the inverse is a relatively simple task. Further inspecting the effect size approach to detecting data fabrication indicates that no false positives occur until $\alpha = 0.31$ (i.e., $r > 0.69$), making this the optimal alpha level based on this sample (but note the small sample).

Table 6: *Table XX.* Diagnosticity of using effect sizes to detect data fabrication, depicted with the AUROC-value.

Method	AUROC
Effect sizes $(1 - r)$	0.744

Study 2 - detecting fabricated Stroop data

We investigated automated detection of data fabrication in raw data using statistics as an extension of Study 1. 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 (**???**) and during data collection (https://osf.io/fc35g).

To test the validity of statistical methods to detect data fabrication in raw data, we investigated raw data of Stroop experiments (**???**). In a Stroop experiment, 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 a Stroop experiment is the response latency (in this study we used milliseconds). Participants in actual studies are typically presented with a set of these Stroop tasks, 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

Data collection

We collected twenty-one genuine datasets on the Stroop task from the Many Labs 3 project (https://osf.io/n8xa7/; **???**). 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; **???**). Similar to Study 1, we assumed these data to be genuine due to the minimal individual gains for fabricating data and the transparency of the project. 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 4). The Stroop effect was calculated as a t-test of the difference between the congruent and incongruent conditions ($H_0 : \mu = 0$).

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

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 5: 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/.

Dutch and Flemish researchers to allow for face-to-face interviews 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 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 (original: https://osf.io/7qhy8/) and the template spreadsheet via Qualtrics (see Figure X; https://osf.io/2qrbs/). 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 euros for participation; to incentivize them to fabricate (supposedly) hard to detect data they could win an additional 100 euros if they belonged to one out of three top fabricators. The contents of the interview were transcribed for further research on qualitatively assessing how researchers might fabricate experimental data.

Data analysis

To automatically detect data fabrication using statistics, we performed sixteen analyses. These sixteen analyses consisted of four NBL digit analyses, four terminal digit analyses, two variance analyses, four multivariate association analyses, a combination test, and effect sizes (using effect sizes was not preregistered).

For the digit analyses, we separated the Ms and SDs per condition and conducted χ^2-tests for each per data set. As such, for one data set, we conducted NBL digit analyses on the first digits of (i) the mean response latencies in the congruent condition, (ii) the mean response latencies in the incongruent condition, (iii) the standard deviation of the response latencies in the congruent condition, and (iv) the standard deviation of the response latencies in the incongruent condition. For the terminal digit analyses, we took the same four sets, but tested on the final digits.

For the variance analyses, we analyzed the standard deviations of the response latencies separated per condition per data set. That is, we analyzed the standard deviations of the response latencies in the congruent condition for excessive consistency separate from the standard deviations of the incongruent condition.

For the multivariate association analyses, we estimated how extreme the observed correlations between the means and standard deviations within and across conditions were. More specifically, we did this for (i) the correlation between the means across conditions, (ii) the standard deviations across conditions, (iii) the means and standard deviations within the congruent condition, and (iv) the means and standard deviations within the incongruent condition. Based on a meta-analysis of the observed correlations from the Many Labs 3 data, we estimated the parametric distribution of the multivariate relations for each of those four under investigation. Using the estimated parametric distribution, we computed the two-tailed p-value under that distribution.

We also conducted a combination of the terminal digit analyses, the variance analyses, and the analyses based on multivariate associations. To this end, we took the p-values of the 10 statistical tests (i.e., four terminal digit analyses, two variance analyses, four analyses of the multivariate associations) and combined them using the Fisher method (Fisher, 1925).

Study 1 showed that effect sizes are a potentially valuable tool to detect data fabrication, which we replicate in Study 2. Based on the data sets, we computed effect sizes for the Stroop effect based on the Many Labs 3 scripts (osf.io/XXXX). Using a difference t-test ($H_0 : \mu = 0$) we computed the t-value and its constituent

effect size as a correlation using

$$r = \sqrt{\frac{\frac{t^2}{df_2}}{\frac{t^2}{df_2} + 1}}$$

Similar to Study 1, we computed the AUROC for each of these statistical methods to detect data fabrication. To recapitulate, if $AUROC = .5$, correctly classifying a randomly drawn dataset in this sample is equal to a coin flip. We follow regard any AUROC value $< .7$ as poor for detecting data fabrication, $.7 \leq$ AUROC $< .8$ as fair, $.8 \leq$ AUROC $< .9$ as good, and AUROC $\geq .9$ as excellent (**???**).

Results

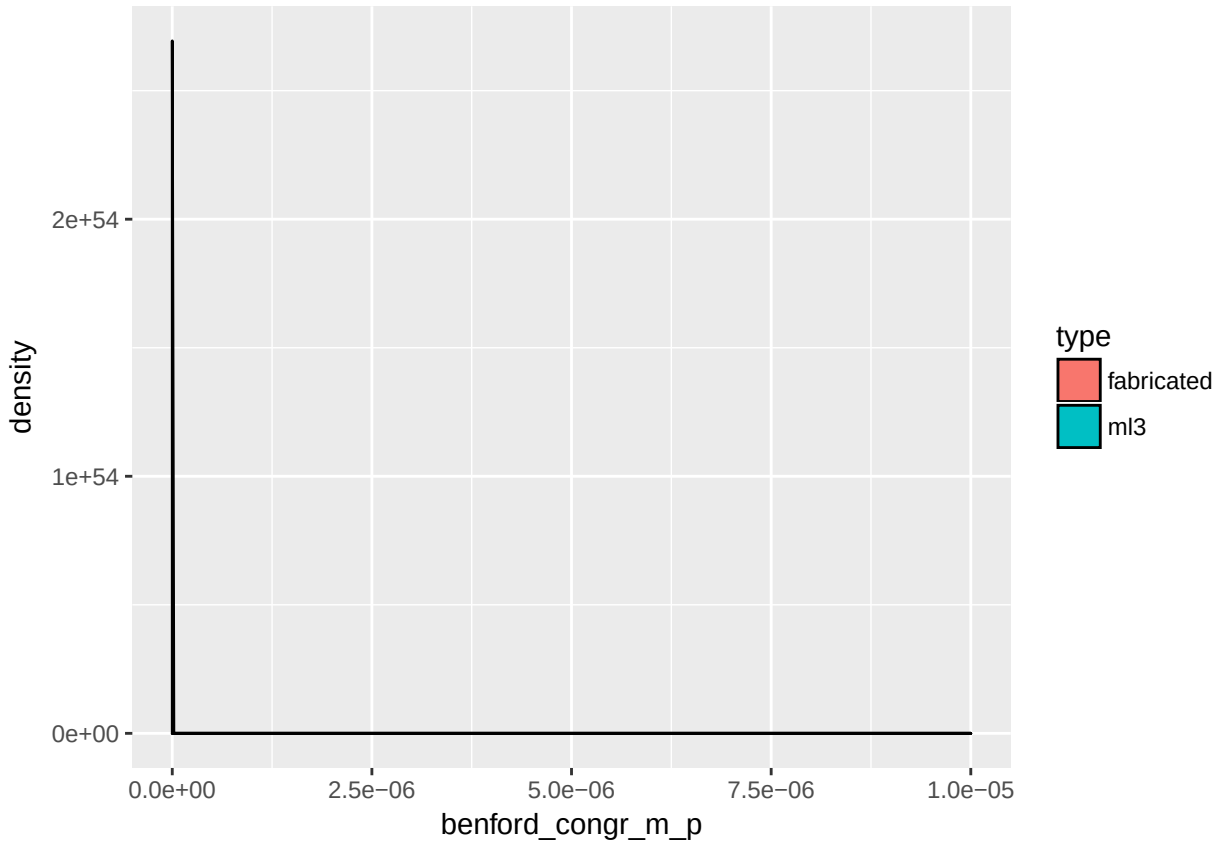
```
##     benford_congr_m_p    benford_congr_sd_p   benford_incongr_m_p
##           0.039115646           0.562925170           0.023809524
##  benford_incongr_sd_p    terminal_congr_m_p   terminal_congr_sd_p
##           0.161564626           0.003401361           0.035714286
##  terminal_incongr_m_p terminal_incongr_sd_p           std_congr_p
##           0.000000000           0.013605442           0.500000000
##         std_incongr_p   p_f_mult_m_sd_congr p_f_mult_m_sd_incongr
##           0.500000000           0.884353741           0.926870748
##   p_f_mult_m_m_across p_f_mult_sd_sd_across          one_min_es_r
##           0.765306122           0.838435374           0.981292517
```

Performance of NBL to detect data fabrication

Table 7: *Table XX.* Diagnosticity of using the Newcomb-Benford law (NBL) to detect data fabrication, depicted with the AUROC-value.

Method	AUROC
NBL, congruent means	0.500
NBL, congruent SDs	0.531
NBL, incongruent means	0.500
NBL, incongruent SDs	0.321

```
ggplot(dat, aes(x = benford_congr_m_p)) + geom_density(aes(fill = type)) + xlim(0, .00001)
```



Performance of terminal digit analysis to detect data fabrication

Table 8: *Table XX.* Diagnosticity of using terminal digit analysis to detect data fabrication, depicted with the AUROC-value.

Method	AUROC
Terminal digit analysis, congruent means	0.002
Terminal digit analysis, congruent SDs	0.043
Terminal digit analysis, incongruent means	0.018
Terminal digit analysis, incongruent SDs	0.014

Performance of variance analysis to detect data fabrication

Table 9: *Table XX.* Diagnosticity of using variance analysis to detect data fabrication, depicted with the AUROC-value.

Method	AUROC
Variance analysis, congruent condition	0
Variance analysis, incongruent condition	0

Performance of multivariate associations to detect data fabrication

Table 10: *Table XX.* Diagnosticity of using multivariate associations to detect data fabrication, depicted with the AUROC-value.

Method	AUROC
Multivariate association means and SDs, congruent condition	0.749
Multivariate association means and SDs, incongruent condition	0.837
Multivariate association means, across conditions	0.594
Multivariate association SDs, across conditions	0.719

Performance of combining

Table 11: *Table XX.* Diagnosticity of using variance analysis to detect data fabrication, depicted with the AUROC-value.

Method	AUROC
Fisher combination of terminal, variance, and multivariate	0.554

Performance of effect sizes

Table 12: *Table XX.* Diagnosticity of using variance analysis to detect data fabrication, depicted with the AUROC-value.

Method	AUROC
Effect size $(1 - r)$	0.984

Discussion

Session info

```
sessionInfo()
```

```
## R version 3.4.0 (2017-04-21)
## Platform: x86_64-redhat-linux-gnu (64-bit)
## Running under: Fedora 26 (Workstation Edition)
##
## Matrix products: default
## BLAS/LAPACK: /usr/lib64/R/lib/libRblas.so
##
## 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
```

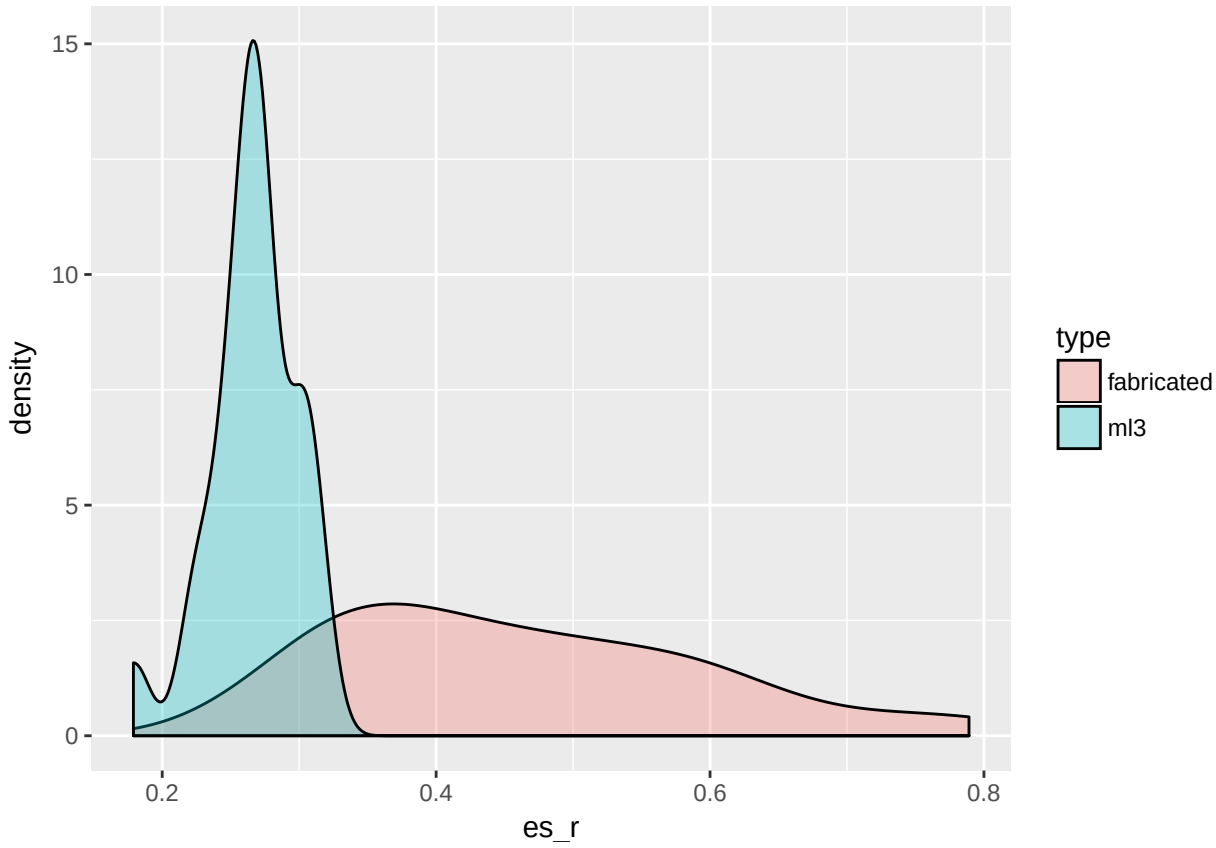


Figure 6: The effect size distributions from Many Labs 3 and those fabricated by the participants

```
##
## other attached packages:
##  [1] stringr_1.2.0     plyr_1.8.4        metafor_2.0-0
##  [4] Matrix_1.2-9      reshape2_1.4.2    dplyr_0.7.1
##  [7] data.table_1.10.4 lsr_0.5           car_2.0-19
## [10] httr_1.2.1        xtable_1.7-1      gridExtra_2.2.1
## [13] ggplot2_2.2.1     latex2exp_0.4.0   foreign_0.8-67
## [16] knitr_1.16        pROC_1.10.0
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.11     highr_0.6        compiler_3.4.0   bindr_0.1
##  [5] tools_3.4.0      digest_0.6.12    nlme_3.1-131     lattice_0.20-35
##  [9] evaluate_0.10.1  tibble_1.3.3     gtable_0.2.0     pkgconfig_2.0.1
## [13] rlang_0.1.1      yaml_2.1.14      bindrcpp_0.2     rprojroot_1.2
## [17] grid_3.4.0       nnet_7.3-12      glue_1.1.1       R6_2.2.1
## [21] rmarkdown_1.6    magrittr_1.5     backports_1.1.0  scales_0.4.1
## [25] htmltools_0.3.6  MASS_7.3-47      assertthat_0.2.0 colorspace_1.3-2
## [29] labeling_0.3     stringi_1.1.5    lazyeval_0.2.0   munsell_0.4.3
```

References

Agresti, A. (2003). *Categorical data analysis* (Vol. 482). London, United Kingdom: John Wiley & Sons. Retrieved from https://mathdept.iut.ac.ir/sites/mathdept.iut.ac.ir/files/AGRESTI.PDF

Akhtar-Danesh, N., & Dehghan-Kooshkghazi, M. (2003). How does correlation structure differ between real and fabricated data-sets? *BMC Medical Research Methodology*, *3*(1). http://doi.org/10.1186/1471-2288-3-18

Anaya, J. (2016). The grimmer test: A method for testing the validity of reported measures of variability. *PeerJ Preprints*, *4*, e2400v1. http://doi.org/10.7287/peerj.preprints.2400v1

Bailey, K. R. (1991). Detecting fabrication of data in a multicenter collaborative animal study. *Controlled Clinical Trials*, *12*(6), 741–752. http://doi.org/10.1016/0197-2456(91)90037-m

Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society*, *78*(4), 551–572. Retrieved from http://www.jstor.org/stable/984802

Bik, E. M., Casadevall, A., & Fang, F. C. (2016). The prevalence of inappropriate image duplication in biomedical research publications. *mBio*, *7*(3), e00809–16. http://doi.org/10.1128/mbio.00809-16

Brown, N. J. L., & Heathers, J. A. J. (2016). The grim test: A simple technique detects numerous anomalies in the reporting of results in psychology. *PeerJ Preprints*, *4*, e2064v1. http://doi.org/10.7287/peerj.preprints.2064v1

Burns, B. D. (2009). Sensitivity to statistical regularities : People (largely) follow Benford's law. In *Proceedings of the thirty first annual conference of the cognitive science society*. Austin, TX: Cognitive Science Society. Retrieved from http://wayback.archive.org/web/20170619175106/http://csjarchive.cogsci.rpi.edu/Proceedings/2009/papers/637/paper637.pdf

Buyse, M., George, S. L., Evans, S., Geller, N. L., Ranstam, J., Scherrer, B., ... Verma, B. L. (1999). The role of biostatistics in the prevention, detection and treatment of fraud in clinical trials. *Statistics in Medicine*, *18*(24), 3435–3451. http://doi.org/10.1002/(SICI)1097-0258(19991230)18:24<3435::AID-SIM365>3.0.CO;2-O

Carlisle, J. B. (2012). The analysis of 168 randomised controlled trials to test data integrity. *Anaesthesia*, *67*(5), 521–537. http://doi.org/10.1111/j.1365-2044.2012.07128.x

Carlisle, J. B. (2017). Data fabrication and other reasons for non-random sampling in 5087 randomised,

controlled trials in anaesthetic and general medical journals. *Anaesthesia*. http://doi.org/10.1111/anae.13938

Carlisle, J. B., Dexter, F., Pandit, J. J., Shafer, S. L., & 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. http://doi.org/10.1111/anae.13126

Cho, W. K. T., & Gaines, B. J. (2007). Breaking the (benford) law: Statistical fraud detection in campaign finance. *The American Statistician*, *61*(3), 218–223. Retrieved from http://www.jstor.org/stable/27643897

Durtschi, C., Hillison, W., & Pacini, C. (2004). The effective use of benford's law to assist in detecting fraud in accounting data. *Journal of Forensic Accounting*, *5*(1), 17–34.

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS ONE*, *4*(5), e5738. http://doi.org/10.1371/journal.pone.0005738

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

Haldane, J. B. S. (1948). The faking of genetical results. *Eureka*, *6*, 21–28. Retrieved from http://wayback.archive.org/web/20170206144438/http://www.archim.org.uk/eureka/27/faking.html

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29–36. http://doi.org/10.1148/radiology.143.1.7063747

Hartgerink, C., & George, S. (2015). Problematic trial detection in ClinicalTrials.gov. *Research Ideas and Outcomes*, *1*, e7462. http://doi.org/10.3897/rio.1.e7462

Hill, T. P. (1995). A statistical derivation of the significant-digit law. *Statistical Science*, *10*(4), 354–363. Retrieved from http://www.jstor.org/stable/2246134

Hobbes, T. (1651). *Leviathan*. Oxford University Presss.

Hogg, R. V., & Tanis, E. A. (2001). *Probability and statistical inference*. New Jersey, NJ: Prentice-Hall.

Jacowitz, K. E., & Kahneman, D. (1995). Measures of anchoring in estimation tasks. *Personality & Social Psychology Bulletin*, *21*, 1161–1166. http://doi.org/10.1037/e722982011-058

James Heathers. (2017). Introducing SPRITE (and the Case of the Carthorse Child). Retrieved from http://wayback.archive.org/web/20170515092023/https://hackernoon.com/introducing-sprite-and-the-case-of-the-carthorse-child-586gi=66761f959132

Kevles, D. J. (2000). *The baltimore case: A trial of politics, science, and character*. WW Norton & Company.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, 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

Koppers, L., Wormer, H., & 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*. http://doi.org/10.1007/s11948-016-9841-7

Newcomb, S. (1881). Note on the frequency of use of the different digits in natural numbers. *American Journal of Mathematics*, *4*(1/4), 39. http://doi.org/10.2307/2369148

Nigrini, M. (2015). Chapter eight. detecting fraud and errors using benford's law. In S. J. Miller (Ed.), *Benfords law*. Princeton University Press. http://doi.org/10.1515/9781400866595-011

Oransky, I. (2015). The Retraction Watch Leaderboard. Retrieved from http://wayback.archive.org/web/20170206163805/http://retractionwatch.com/the-retraction-watch-leaderboard/

O'Brien, S. P., Danny Chan, Leung, F., Ko, E. J., Kwak, J. S., Gwon, T., ... Bouter, L. (2016). Proceedings of the 4th world conference on research integrity. *Research Integrity and Peer Review*, *1*(S1). http://doi.org/

10.1186/s41073-016-0012-9

Parker, A., & Hamblen, J. (1989). Computer algorithms for plagiarism detection. *IEEE Transactions on Education*, *32*(2), 94–99. http://doi.org/10.1109/13.28038

Simonsohn, U. (2013). Just post it. *Psychological Science*, *24*(10), 1875–1888. http://doi.org/10.1177/0956797613480366

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

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677–680. http://doi.org/10.1126/science.103.2684.677

The Journal of Cell Biology. (2015). About the Journal. Retrieved from https://web.archive.org/web/20150911132421/http://jcb.rupress.org/site/misc/about.xhtml

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*(2), 105–110. http://doi.org/10.1037/h0031322

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