

# Detection of data fabrication using statistical tools

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

*29 July, 2018*

## Loading required package: viridisLite

## Abstract

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 faced Stapel; medical sciences faced Poldermans and Macchiarini; life sciences faced Voignet; physical sciences faced Schön — these are just a few examples of research misconduct cases in the last decade. Overall, an estimated 2% of all scholars admit to falsifying or fabricating research results at least once (Fanelli, 2009), which due to its self-report nature is likely to be an underestimate. 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 that there are several hundreds of thousands of researchers in these countries. At best, this suggests a detection rate 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 behavior. This idea is based on deterrence theory (Hobbes, 1651), which stipulates that increased risk of detection decreases the expected utility of scientific misconduct, hence, fewer people will engage in it. Detection techniques have developed differently for fabrication, falsification, and plagiarism. Plagiarism scanners have been around the longest (e.g., A. Parker & Hamblen, 1989) and are widely implemented not only at journals but also in the evaluation of student theses (e.g., with commercial services such as Turnitin). For data fabrication, developments around detecting image manipulation are more recent, with some tools even being implemented at journals. For example, the Journal of Cell Biology and the EMBO journal scan each submitted image for potential manipulation (The Journal of Cell Biology, 2015; 2017), which supposedly increases the risk of detecting (blatant) image manipulation. More recently, algorithms are being developed to automate the scanning of images for such manipulations (Koppers, Wormer, & Ickstadt, 2016). The application of such tools can also help researchers systematically evaluate research articles in order to estimate the extent to which image manipulation occurs (4% of all papers are estimated to contain manipulated images, Bik, Casadevall, & Fang, 2016) or what factors are predictive of image manipulation (Fanelli, Costas, Fang, Casadevall, & Bik, 2018).

Detection methods for data fabrication in empirical research are often based on a mix of psychology theory and statistics theory. Because humans are notoriously bad at understanding and estimating randomness (???, Haldane, 1948; Nickerson, 2000; Amos Tversky & Kahneman, 1971; A. Tversky & Kahneman, 1974), this could manifest itself in the fundamentally probabilistic data they try to fabricate. Whether the data and outcomes of analyses based on these data are in line with the (at least partly probabilistic) processes that are assumed to underlie them, may indicate deviations from the reported protocol, potentially even data fabrication.

Statistical methods have proven to be of importance in initiating data fabrication investigations or in assessing scope of potential data fabrication. For example, Kranke, Apfel, and Roewer skeptically perceived Fuji's data (Peter Kranke, Apfel, & Roewer, 2000) and used statistical methods to contextualize their skepticism. At the time, a reviewer perceived them to be on a “crusade against Fujii and his colleagues” (P. Kranke, 2012) and further investigation remained absent. Only when Carlisle extended the systematic investigation to 168 of Fuji's papers (Carlisle, 2012; Carlisle & Loadman, 2016; Carlisle, Dexter, Pandit, Shafer, & Yentis, 2015) did events cumulate into an investigation- and ultimately retraction of 183 of

Fuji's peer-reviewed papers ("Joint editors-in-chief request for determination regarding papers published by dr. yoshitaka fujii," 2013; Oransky, 2015). In another example, the Stapel case, statistical evaluation of his oeuvre occurred after he had already confessed to fabricating data, which resulted in 58 retractions of papers (co-)authored by Stapel (Levelt, 2012; Oransky, 2015).

In order to determine whether the application of statistical methods to detect data fabrication is responsible, their diagnostic value requires further investigation to inform decisions about the utility of these methods. Specifically, many of the developed statistical methods to detect data fabrication are quantifications of case specific suspicions by researchers. Hence, these could be considered mere proposals and their diagnostic value (i.e., sensitivity and specificity) unknown until they are thoroughly validated outside of those specific cases. Side-by-side comparisons of these proposed statistical methods has also been difficult through the in-casu origin of these methods. Moreover, the efficacy of these methods based on known cases is likely to be biased, considering that an unknown amount of undetected cases are not included. With respect to the utility of these statistical methods, questions about whether the sensitivity and specificity are permissible in light of the severe professional- and personal consequences of potential research misconduct need to be asked (regretfully, the STAP case brings this to the fore very clearly; Cyranoski, 2015). These methods might have utility in misconduct investigations, but not in large-scale applications to screen the literature, depending on the diagnostic values resulting from a controlled investigation.

In this article, we investigate the diagnostic performance of statistical methods to detect data fabrication. These statistical methods (detailed next) have not previously been validated using both genuine- and fabricated data. We present two studies where we try to distinguish assumably genuine- from known fabricated data based on statistical methods. These studies investigate methods to detect data fabrication in summary statistics (Study 1) or in raw data (Study 2). In Study 1, we invited researchers to fabricate summary statistics for a set of four anchoring studies, for which we also had genuine data from the Many Labs 1 initiative (<https://osf.io/pqf9r>; Klein et al., 2014). In Study 2, we invited researchers to fabricate raw data for a Stroop experiment, for which we also had genuine data from the Many Labs 3 initiative (<https://osf.io/n8xa7/>; Ebersole et al., 2016). Before presenting these studies, we expand on the theoretical framework of the investigated statistical methods to detect data fabrication.

## Theoretical framework

In the current paper, we differentiate between statistical methods to detect potential data fabrication based on reported summary statistics or raw data. Below, we expand on the theoretical underpinnings of these methods. For summary statistics, we review  $p$ -value analysis, variance analysis, and effect size analysis as potential ways to detect data fabrication.  $P$ -value analyses can be applied whenever a set of nonsignificant  $p$ -values are reported; variance analysis can be applied whenever a set of variances and accompanying sample sizes are reported for independent, randomly assigned groups; 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; Chris H. J. Hartgerink, Wicherts, & Van Assen, 2017). For raw data, we review digit analyses (i.e., the Newcomb-Benford law and terminal digit analysis) and multivariate associations between variables as potential ways to detect data fabrication. The Newcomb-Benford law can be applied on ratio- or count scale measures that have sufficient digits and that are not truncated (Hill & Schürger, 2005); terminal digit analysis can be applied whenever measures have sufficient digits (see also Mosimann, Wiseman, & Edelman, 1995). Multivariate associations can be investigated whenever there are two or more numerical variables available 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 single or a set of independent  $p$ -values is uniform if the null hypothesis is true; it is right-skewed if the alternative hypothesis is true (Fisher, 1925). If the model assumptions of the underlying process hold, 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  $p$ -values if those  $p$ -values are independent.

When assumptions underlying the model used to compute a  $p$ -value are violated,  $p$ -value distributions can take on a variety of shapes. For example, when optional stopping (i.e., adding batches of participants until you have a statistically significant result) occurs and the null hypothesis is true,  $p$ -values just below .05 become more frequent (C. H. Hartgerink, Aert, Nuijten, Wicherts, & Assen, 2016; Lakens, 2015). However, when optional stopping occurs under the alternative hypothesis or when other researcher degrees of freedom are used, a right-skewed distribution for significant  $p$ -values can still occur (C. H. Hartgerink et al., 2016; Ulrich & Miller, 2015).

When independent  $p$ -values are not right-skewed or uniformly distributed (as would be theoretically expected), it can indicate potential data fabrication. For example, in the Fuji case, data of supposedly randomly assigned groups were fabricated. In truly randomly assigned groups, the measurements of different groups (prior to an intervention) can be assumed to be generated by the same probabilistic process, resulting in uniformly distributed  $p$ -values when comparing these groups using statistical tests. However, in the Fuji case Carlisle observed many large  $p$ -values, which ultimately led to the identification of potential data fabrication (Carlisle, 2012). The cause of these large  $p$ -values is that Fuji, when fabricating the data, underappreciated the effect of randomness, thereby creating groups of data that were too similar conditional on the null hypothesis of no differences between the groups. In Table 1 we illustrate the difference between expected data under the null distribution (Set 1) and excessively consistent and potentially fabricated data (Set 2). More specifically, the expected value of a uniform  $p$ -value distribution is .5, but the fabricated data from our illustration have a mean  $p$ -value of 0.956.

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 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

In order to test whether a distribution of independent  $p$ -values might be fabricated, we previously proposed using the Fisher method (Fisher, 1925; S. P. O’Brien et al., 2016). The Fisher method originally was intended as a meta-analytic tool, which tests whether there is sufficient evidence for an effect (i.e., right-skewed  $p$ -value distribution). This test is computed as

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

where it tests for more smaller  $p$ -values than larger  $p$ -values across the  $k$  number of  $p$ -values. Reversing this results in

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

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  $t$  (i.e., the Fisher method now tests for left-skew). When  $t = 0$ , all  $p$ -values are selected. When  $t > 0$  the remaining  $p$ -values are rescaled to fit the original 0-1 range (i.e., dividing by

$1 - t$ ). This test is similar (but not equivalent) to Carlisle’s method testing for excessive homogeneity across baseline measurements in RCTs (Carlisle, 2012, 2017; Carlisle et al., 2015).

As an example, we apply the reversed Fisher method to both the genuine- and fabricated results from Table 1. Using the threshold  $t = 0.05$  to select only the nonsignificant results from Table 1, we retain  $k = 10$  genuine  $p$ -values and  $k = 10$  fabricated  $p$ -values. This results in  $\chi^2_{2 \times 10} = 18.362, p = 0.564$  for the genuine data (Set 1), and  $\chi^2_{2 \times 10} = 66.848, p = 6 \times 10^{-7}$  for the fabricated data (Set 2). Another more practical example directly from the Fuji case (Carlisle, 2012), illustrates that actual fabricated data can result in significant findings with the reversed Fisher method. For example,  $p$ -values extracted from the original Table 3 (fentanyl dose; Carlisle, 2012) for five independent comparisons also show excessively high  $p$ -values,  $\chi^2_{2 \times 5} = 19.335, p = 0.036$ . However, based on this anecdotal evidence little can be said about the sensitivity, specificity, and utility of this method.

We note that misspecified one-tailed tests can also result in excessive amounts of large  $p$ -values. For correctly specified one-tailed tests, the  $p$ -value distribution is right-skewed if the alternative hypothesis is true. When the alternative hypothesis is true, but the effect is in the opposite direction of the hypothesized effect (e.g., a negative effect when a one-tailed test for a positive effect is conducted), this results in a left-skewed  $p$ -value distribution. As such, any potential data fabrication detected with this method would need to be inspected for misspecified one-tailed hypotheses to preclude false conclusions. In the studies we present in this manuscript, misspecification of one-tailed hypothesis testing is not an issue because we prespecified the effect and its direction to the participants.

## Variance analysis

Sample variance or standard deviation estimates are typically reported to indicate dispersion in the data. As is common in many empirical research papers, the mean is reported alongside its SD because it is a valuable tool in determining how diverse participants are on a specific measure. For example, if a sample has a reported age of  $M(SD) = 21.05(2.11)$  we know this group is both younger and more homogeneous than another group with reported  $M(SD) = 42.78(17.83)$ . For this section, we will talk about standard deviations and variances in the data across various groups as is common in an experimental design.

Similar to the estimate of the mean in the data, there is sampling error in the estimated variance in the data (i.e., dispersion of the variance). The sampling error of the estimated variance is inversely related to the sample size. For example, under the assumption of normality the sampling error of a given standard deviation can be estimated as  $\sigma/\sqrt{2n}$  (p. 351, Yule, 1922), where  $n$  is the sample size of the group. Additionally, if an observed random variable  $x$  is normally distributed, the standardized variance of  $x$  is  $\chi^2$ -distributed (p. 445; Hogg & Tanis, 2001); that is

$$var(x) \sim \frac{\chi^2_{N_j-1}}{N_j - 1}$$

where  $N$  is the sample size of the  $j$ th group. We can compute the unstandardized variance by computing the Mean Squares within ( $MS_w$ ) 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 reported variance and  $N_j$  the reported sample size in group  $j$ . When calculating  $MS_w$ , equality of variances across  $j$  groups is assumed. As such, under normality and equality of variances, we can simulate the expected distribution of variances in the data by multiplying the results of the  $\chi^2$  distribution with  $MS_w$ . Conversely, the variances reported in a paper can be standardized by dividing the observed variances by  $MS_w$ ; we denote standardized variances with  $z^2$  hereafter.

In order to compute the dispersion of the standard deviation, we simulate the distribution of expected variances under the null model. The null model is that the data and its subsequent standard deviations arise from a true probabilistic process, when assuming normality and equality of variances. In each iteration  $i$  of the simulation, we generate standardized variances for each group  $j$ . Combining the

distribution of  $var(x)$  and  $MS_w$ , the distribution of the standardized variances in group  $j$  follows a  $\chi^2$ -distribution in the form of

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

Upon simulating standardized variances for all  $j$  groups, we compute two dispersion measures of those variances.

For each iteration, we calculate the standard deviation and range of the estimated variances in the data. Repeating this process across  $i$  iterations provides an estimated density function for the expected dispersion of the variances (either in its range or SD). By comparing the observed dispersion of the variances with the expected dispersion of variances, we can estimate how extreme our observations are. More specifically, we can compute how many iterations show equally- or more extreme consistency in the data to compute a bootstrapped  $p$ -value. For our purposes, too little dispersion in the variances may indicate potential fabrication in the reported data (Simonsohn, 2013). This could be the result of the fabricator underestimating sampling fluctuations due to intuitively misunderstanding probabilistic processes, resulting in generating too little randomness (i.e., error) in data (Mosimann et al., 1995). Observed dispersion of the standardized variances can be operationalized 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  $max_z - min_z$ ).

As an example, we apply the variance analysis to the illustration from Table 1 and the Smeesters case. For the reported standard deviations in Table 1, we apply the variance analysis across both the experimental and control conditions, separating the genuine- and fabricated sets. For the genuine data (Set 1), we find that the reported mean standard deviation is 9.868 with a standard deviation of 0.595; for the fabricated data (Set 2), we find that the reported mean standard deviation is 10.667 with a standard deviation of 0.456. These summary statistics of the standard deviations already indicate there is a difference between the genuine- and fabricated data. Variance analysis, as explained previously, helps us quantify how extreme this difference is: Set 1 has no excessive consistency in the dispersion of the standard deviations ( $p=0.214$ ), whereas Set 2 does show excessive consistency in the dispersion of the standard deviations ( $p = 0.006$ ). In words, out of 100,000 theoretically expected samples under the null model of independent groups with equal variances on a normally distributed measure,  $2.142 \times 10^4$  showed less dispersion in standard deviations for Set 1, whereas only 572 showed less dispersion in standard deviations for Set 2. As a non-fictional example, three independent conditions from the one study in the Smeesters case ( $n_k = 15$ ) were reported to have standard deviations 25.09, 24.58, and 25.65. The standard deviation of these standard deviations is 0.54 (i.e.,  $SD_z$ ). Such consistency in standard deviations (or even more) would only be observed in 1.21% of 100,000 simulated replications (Simonsohn, 2013).

## Effect sizes

Large effects have previously been opted to arise from dubious origins (Lieberman, Berkman, & Wager, 2009), but there is sufficient evidence that large effects can arise from data fabrication. For example, in the misconduct investigations in the Stapel case, effect sizes were one particular indicator of data fabrication in certain papers (Levelt, 2012). Some papers showed extreme explained variances of up to 95%. Moreover, Akhtar-Danesh & Dehghan-Kooshkghazi (2003) asked faculty members from three universities to fabricate data sets and found that the fabricated data showed much larger effect sizes than the genuine data. From our own anecdotal experience, we have found that large effect sizes raised initial suspicions of data fabrication (e.g.,  $d > 20$ ). In clinical trials, extreme effect sizes are also used to identify potentially fabricated data in multi-site trials while the study is still being conducted (Bailey, 1991).

Effect sizes can be reported in research reports in various ways. For example, the most common ways effect sizes are reported in papers are as a standardized mean difference (e.g.,  $d$ ), as an explained variance (e.g.,  $R^2$ ), or as a test statistic. A test statistic is also a measure of effect size, albeit in a not directly interpretable form. A test result such as  $t(59) = 3.55$  corresponds to  $d=0.924$  and  $r=0.176$  (Chris H. J. Hartgerink et al., 2017). These effect sizes can readily be recomputed based on data extracted with `statcheck` across thousands of results (Hartgerink, 2016; Nuijten, Hartgerink, Assen, Epskamp, & Wicherts, 2015).

Observed effect sizes can subsequently be compared with the effect distribution of other studies investigating the same effect. For example, if a study on the ‘foot-in-the-door’ technique yields an effect size of

$r = .8$ , we can collect other studies that investigate the ‘foot-in-the-door’ effect and compare how extreme that  $r = .8$  is in comparison to the other studies. If the largest observed effect size in the control data is  $r = .2$  and a reasonable number of studies on the ‘foot-in-the-door’ effect have been done, this can be considered extreme and a flag for potential data fabrication. 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

The properties of leading (first) digits (e.g., the 1 in 123.45) or terminal (last) digits (e.g., the 5 in 123.45) may be examined in raw data. By analyzing these leading- and terminal digits for deviations from specific and theoretically expected digit distributions, it might be possible to screen for fabricated data. Here we focus on testing leading digit based on the Newcomb-Benford Law (NBL) and testing terminal digits based on measurement error in order to detect potentially fabricated data.

For leading digits, the Newcomb-Benford Law or NBL (Benford, 1938; Newcomb, 1881) states that these digits do not have an equal probability of occurring under certain conditions but a monotonically decreasing probability. 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 is, according to the NBL:

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

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  $\chi^2$ -test ( $df = 9 - 1$ ). In order to make such a comparison feasible, it 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 (Bauer & Gross, 2011; Hüllemann, Schüpfer, & Mauch, 2017).

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, the NBL only applies under specific conditions that are rarely fulfilled in the social sciences. Hence, its applicability for detecting data fabrication in science can be questioned. First, the NBL only applies for true ratio scale measures (Berger & Hill, 2011; Hill, 1995). Second, sufficient range on the measure is required for the NBL to apply (i.e., range from at least  $1 - 1000000$  or  $1 - 10^6$ ; Fewster, 2009). 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; Diekmann, 2007), immediately undermining the applicability of the NBL.

For terminal digits, 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 (Mosimann & Ratnaparkhi, 1996; Mosimann et al., 1995). Terminal digit analysis is conducted with a  $\chi^2$ -test

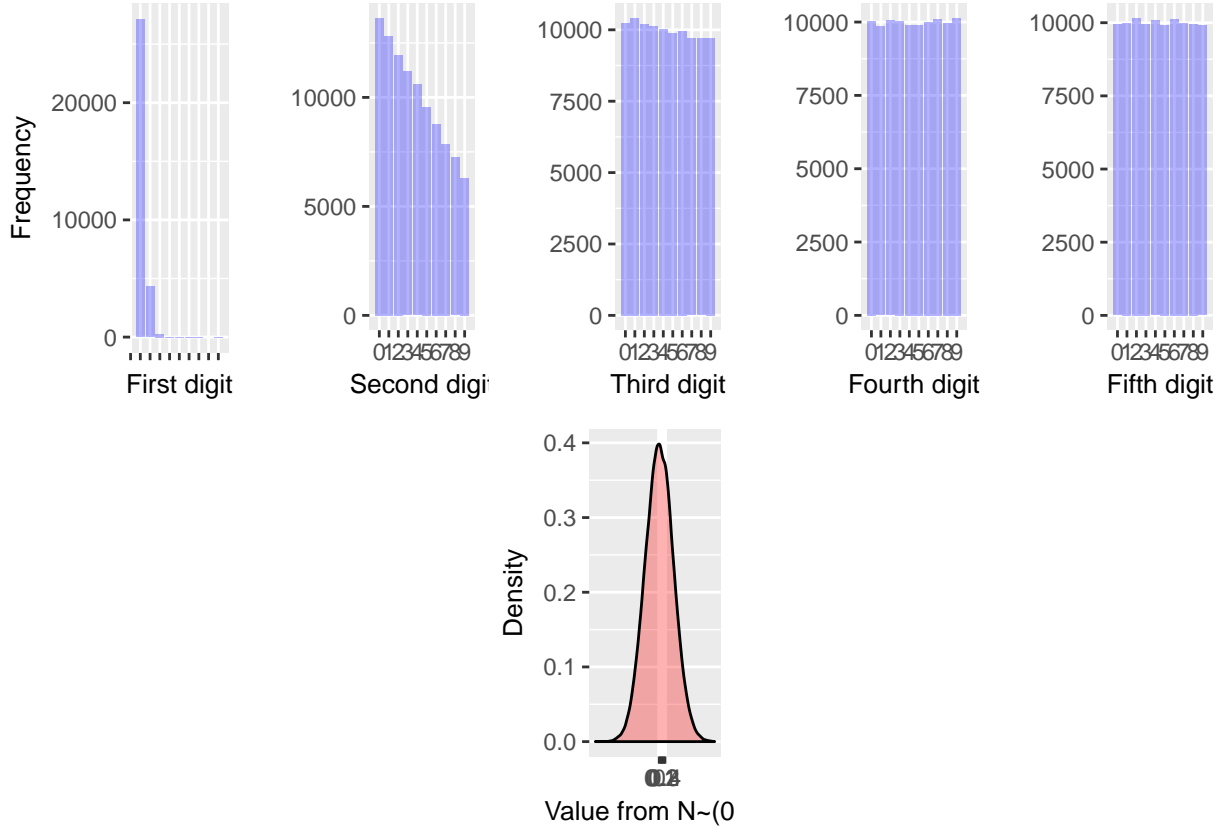


Figure 1: Illustration of how digit distributions evolve from first- through later digits. We sampled 100,000 values from a normal distribution to create these digit distributions.

( $df = 10 - 1$ ) on the digit occurrence counts (including zero), where the observed frequencies are compared with the expected uniform frequencies. The 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 Mosimann & Ratnaparkhi (1996; for a history of this decade long case, see Kevles, 2000).

Figure 1 depicts simulated digit counts for the first- through fifth digit of a random, normally distributed variable (i.e.,  $N \sim (0, 1)$ ). The first- and second digit distributions are clearly non-uniform, whereas the third digit distribution seems slightly non-uniform, and the 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 (Mosimann et al., 1995). What sufficient precision is, we investigated by running a small simulation study, drawing 500 random values from a normal distribution ( $N \sim (0, 1)$ ) thousand times and conducting a terminal digit test for each of the first five digits. For the third-, fourth-, and fifth- digits, tests operated on nominal  $\alpha$  levels (i.e., under  $\alpha = .05$ , false positives were 0.945, 0.955, 0.959, respectively). Hence, sufficient precision for our purposes is determined as the terminal digit being conducted on at least the third leading digit (i.e., minimally 1.23 or 12.3 or 123).

## Multivariate associations

Variables or measurements included in one study can have multivariate relations that might be non-obvious to researchers. Hence, such relations between variables or measurements might be forgotten by people who fabricate data. Fabricators might also simply be practically unable to fabricate data that also show these multivariate associations, even if they are knowledgeable of these. For example, in response time latencies, there is a negative relation between mean response time and the variance of the response time, where lower mean response times are accompanied by a lower variance due to truncation. Given that the genuine multivariate relations between different variables arise from stochastic processes and are not readily known in either their form or size, these might be difficult to take into account when someone

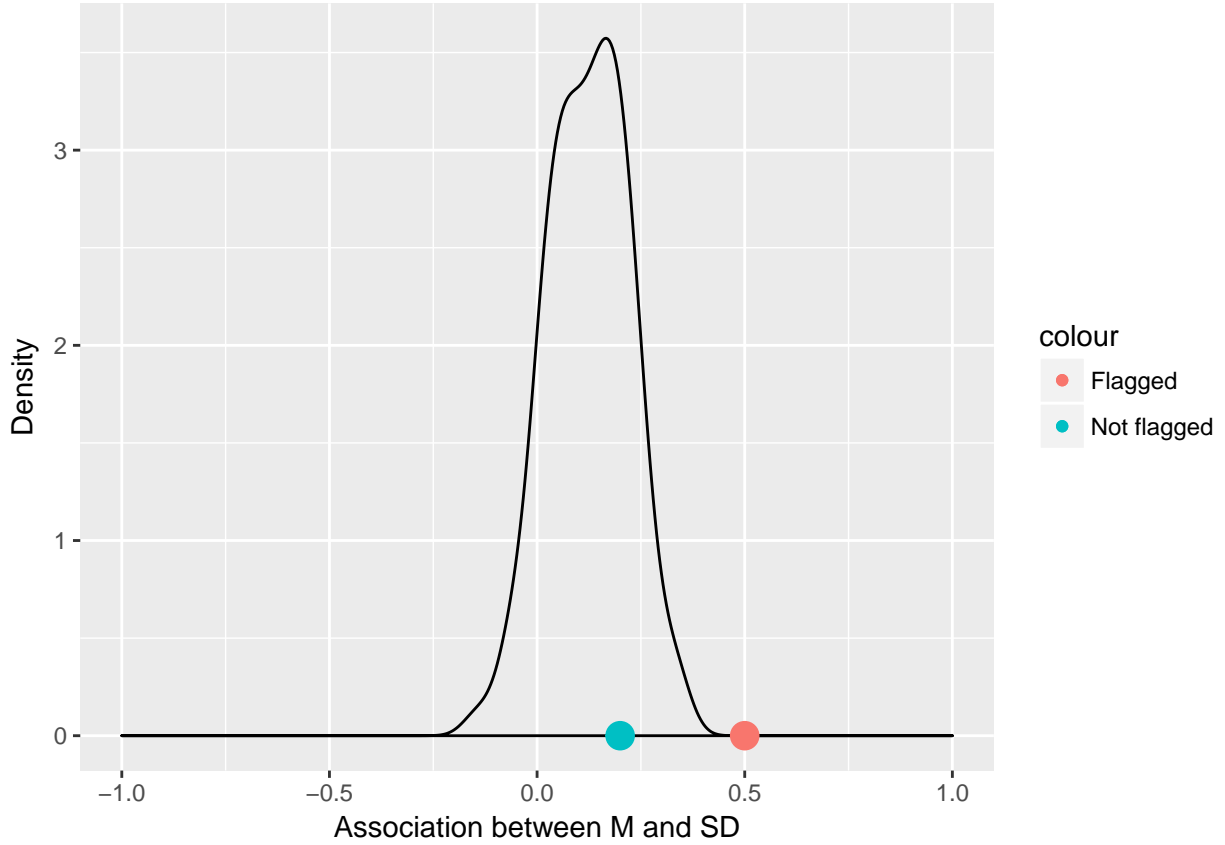


Figure 2: A fictitious distribution of 100 simulated observed associations between Ms and SDs arising from  $N(.123, .1)$ . The red dot indicates the observed relation that is flagged for further screening for potential data fabrication.

wants to fabricate data. As such, using multivariate associations to discern fabricated data from genuine data might prove worthwhile.

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 that can subsequently be used in a parametric  $z$ -test (assuming normality).

The multivariate associations from the genuine data are subsequently used to estimate how extreme the observed and investigated multivariate relation is. Consider the following fictitious example, regarding the multivariate association between Ms and SDs for a response latency task mentioned earlier. Figure 2 depicts a (simulated) population distribution of the association between Ms and SDs from the literature ( $N \sim (.123, .1)$ ). 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 fabricated (i.e., the red dot; two-tailed  $p$ -value  $1.027 \times 10^{-4}$ ), whereas the latter (blue dot) is not flagged ( $p$ -value: 0.431).

## Study 1 - detecting fabricated summary statistics

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:

Do you think the percentage of African countries in the UN is above or below [10% or 65%]?  
What do you think is the percentage of African countries in the UN?

These differently anchored questions yield mean responses of 25% and 45%, respectively (A. Tversky & Kahneman, 1974), despite essentially posing the same factual question. A considerable amount of (assumably) genuine data sets on the anchoring heuristic are freely available (<https://osf.io/pqf9r>; Klein et al., 2014). In our study we asked researchers to fabricate summary statistics on anchoring experiments on the same studies. This study was approved by the Tilburg Ethical Review Board (EC-2015.50; <https://osf.io/7tg8g/>).

## Methods

We collected genuine- and fabricated summary statistics for four anchoring studies: (i) distance from San Francisco to New York, (ii) human 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 our 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 participant. The test results available are the main effect of the anchoring condition, the main effect of gender, and the interaction effect between the anchoring conditions and gender conditions. For current purposes, a participant 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/tshx8/>. Throughout this report, we will indicate which facets were not preregistered or deviate from the preregistration (for example by denoting "(not preregistered)" or "(deviation from preregistration)").

## Data collection

We downloaded thirty-six genuine data sets 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 felt confident to assume these data are genuine. For each of the thirty-six locations we computed three summary statistics (i.e., sample sizes, means, and standard deviations) for each of the four conditions in the four anchoring studies (i.e.,  $3 \times 4 \times 4$ ; data: <https://osf.io/5xgcp/>). We computed these summary statistics from the raw ML data, which were cleaned using the original analysis scripts from the ML project.

The sampling frame for the participants asked to fabricate data 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). We filtered for U.S. researchers to ensure familiarity with the imperial measurement system, which is the scale of some of the anchoring studies and in order to reduce heterogeneity across fabricators. We searched WoS on October 13, 2015. In total, 2,038 unique corresponding e-mails were extracted from 2,014 papers (due to multiple corresponding authors).

From these 2,038 psychology researchers, we e-mailed a random sample of 1,000 researchers to participate in this study (April 25, 2016; [osf.io/s4w8r](https://osf.io/s4w8r)). We used Qualtrics and removed identifying information not essential to the study (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 participants that they could stop at any time without providing a reason. If they wanted, participants received a \$30 Amazon gift card as compensation for their participation 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 (this procedure is explained in the Data Analysis section). 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](https://osf.io/w984b).

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 out template spreadsheet used in the fabrication process of Study 1. Respondents fabricated data in the yellow cells, which were used to automatically 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>).

Each participant was instructed to fabricate 32 summary statistics ( $4 \text{ studies} \times 2 \text{ anchoring conditions} \times 2 \text{ sexes} \times 2 \text{ statistics [mean and sd]}$ ) that corresponded to three hypotheses. We instructed participants to fabricate results for the following hypotheses: there is (i) a positive main effect of the anchoring condition, (ii) no effect of sex, and (iii) no interaction effect between condition and sex. We fixed the sample sizes to 25 per cell; participants did not need to fabricate sample sizes. These 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.

We provided participants 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 3 depicts an example of this spreadsheet (original: <https://osf.io/w6v4u>). We requested participants to fill out the yellow cells with fabricated data, which includes means and standard deviations for the four conditions. Using these values, statistical tests are automatically computed and shown in the “Current result” column instantaneously. If these results supported the (fabrication) hypotheses, a checkmark appeared as depicted in Figure 3. We required participants to copy-paste the yellow cells into Qualtrics. This provided a standardized response format that could be automatically processed in the analyses. Technically, participants could provide a response that did not correspond to the instructions but nobody did.

Upon completion of the data fabrication, we debriefed respondents within Qualtrics ([osf.io/rg3qc/](https://osf.io/rg3qc/)). Respondents answered several questions about their statistical knowledge and approach to data fabrication. Lastly we reminded them that data fabrication is widely condemned by professional organizations, institutions, and funding agencies alike. This reminder was intended to minimize potential carry-over effects of the unethical behavior into actual research practice (Mazar, Amir, & Ariely (2008); although a recent multilab replication contests this finding, [osf.io/cwavm/](https://osf.io/cwavm/)). We rewarded participation with a \$30 Amazon gift card and the fabricated results that we had most difficulty to detect received a bonus \$50 Amazon gift card. Using quota sampling, we collected as many responses as possible for the available 36 rewards, resulting in 39 fabricated data sets (<https://osf.io/e6zys/>; 3 participants did not participate for a bonus).

## Data analysis

We analyzed the genuine- and fabricated data sets for each of the anchoring studies in four ways. First, we applied the reversed Fisher method to the results of the gender and interaction hypotheses separately (i.e., statistically nonsignificant results) across the four studies. Second, we applied variance analyses to the reported variances of the four studies. Third, and not preregistered, we used the effect sizes of the anchoring effect to detect fabricated data based on the premise that fabricated statistically significant effects would be (much) larger than genuine statistically significant effects. Fourth, we combined the results from the reversed Fisher method and variance analyses using the original Fisher method (a meta-analysis method; Fisher, 1925).

Specifically for the variance analyses, we substantially deviated from the preregistration (<https://osf.io/>

tshx8/) and added multiple analyses. Initially, we simultaneously analyzed the reported variances across studies and across the anchoring conditions, combining it into one overall variance analysis. However, only upon analyzing these values, we realized that the variance analyses assume that the included variances are from the same (standardized) population distribution. Homogeneous population of variances is not necessarily the case for the different anchoring conditions. Hence, we included subgrouped variance analyses, where we analyzed each anchoring study separately and also added deeper subgroup analyses where we split each study into two variance analyses. The latter split the studies into two (more) homogeneous subsets (i.e., the low/high anchoring condition collapsed across gender). As such, the only preregistered result is the overall variance analysis [homogeneity] under both the  $SD_z$  and  $max_z - min_z$  operationalizations. We added separate analyses per study (assuming homogeneous variances across anchoring conditions) and analyses assuming heterogeneous variances across anchoring conditions, for both operationalizations.

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 summarize the sensitivity (i.e., True Positive Rate [TPR]) and specificity (i.e., True Negative Rate [TNR]) for various decision criteria (e.g.,  $\alpha = 0, .01, .02, \dots, .99, 1$ ). For our purposes, AUROC values indicate the probability that a randomly drawn fabricated- and genuine dataset can be correctly classified as fabricated or genuine based on the result of the analysis (Hanley & McNeil, 1982). In other words, if  $AUROC = .5$ , correctly classifying a randomly drawn dataset as fabricated (or genuine) in this sample is equal to 50%. For this setting, we follow the guidelines of Youngstrom (2013) 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. We conducted all analyses using the pROC package (Robin et al., 2011).

## Results

Figure 4 shows a group-level comparison of the genuine- ( $k = 36$ ) and fabricated ( $k = 39$ )  $p$ -values and relevant effect sizes ( $r$ ). These group-level comparisons provide a general overview of the differences between the genuine- and fabricated data. Figure 4 indicates that there are few group differences between fabricated and genuine summary statistics from the anchoring studies when statistically nonsignificant effects are inspected (i.e., gender and interaction hypotheses). However, there seem to be larger group differences when we required participants to fabricate statistically significant summary statistics (i.e., anchoring hypothesis). We zoom in on more specific results next; Figure 4 already indicates that statistically nonsignificant effects are likely to be less discerning between fabricated- and genuine data in this sample than statistically significant effects.

### P-value analysis

When we apply the reversed Fisher method to the statistically nonsignificant effects, results indicate its performance is approximately equal to chance classification. We asked researchers to fabricate data for statistically nonsignificant effect sizes across four anchoring studies, thinking they might be unable to produce uniformly distributed  $p$ -values due to widespread misunderstanding of what a  $p$ -value indicates (Goodman, 2008; Sijtsma, Veldkamp, & Wicherts, 2015). Our results indicate that using the statistically nonsignificant effects and analyzing them with the reversed Fisher method is not effective in detecting data fabrication when genuine data is available to compare to. More specifically, when using statistically nonsignificant gender effects in these anchoring studies we find  $AUROC = 0.501$ , 95% CI [0.468-0.535]; for statistically nonsignificant interaction effects in these anchoring studies we find  $AUROC = 0.516$ , 95% CI [0.483-0.549]. In other words, results from this sample indicate that detection of fabricated data using the distribution of statistically nonsignificant  $p$ -values to detect excessive amounts of high  $p$ -values does not seem promising.

### Variance analysis

We computed the AUROC values for the variance analyses with the directional hypothesis that genuine data would show more variation than fabricated data. In other words, fabricated data was expected to show less dispersion in reported variances when compared to genuine data. AUROC results of all 14

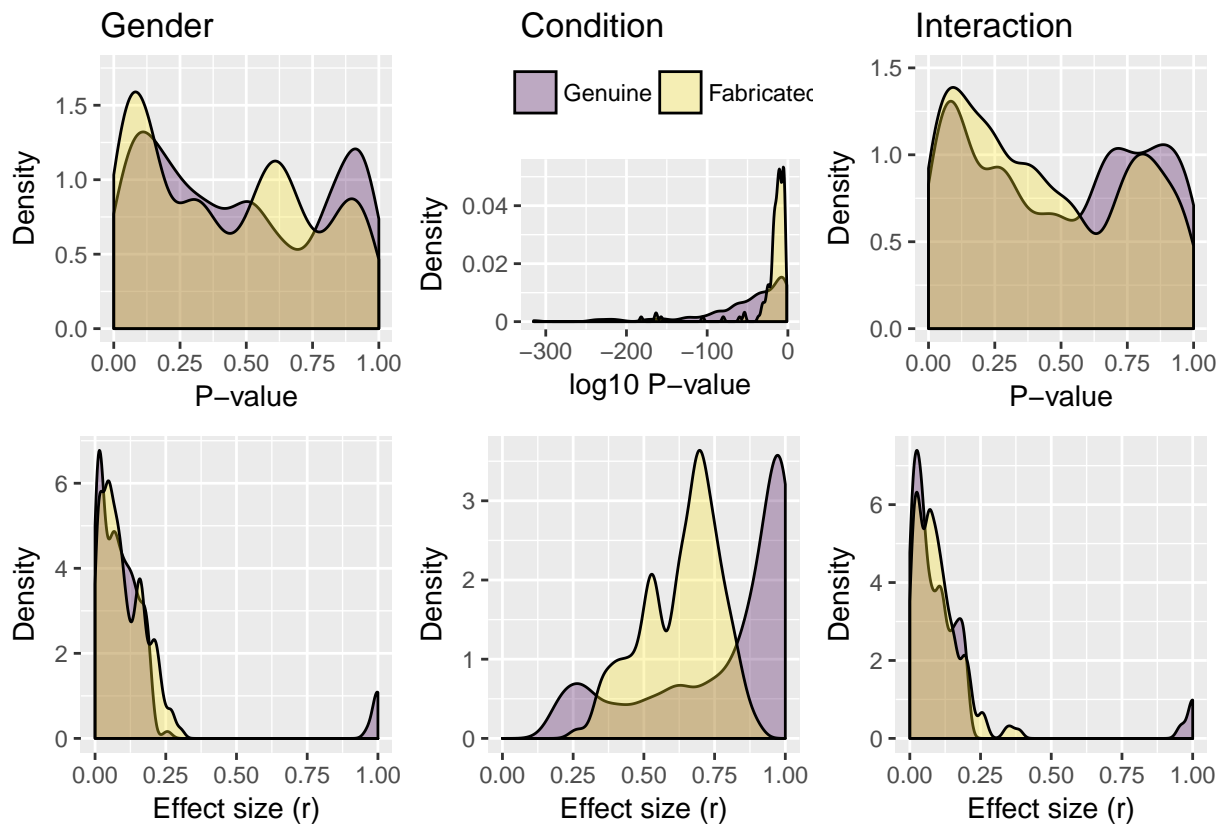


Figure 4: Overlay of (smoothed) density distributions for both genuine and fabricated data across four anchoring studies, per effect and type of result. We instructed respondents to fabricate nonsignificant summary statistics for the gender and interaction effects, and a significant effect for the condition effect.

analyses are presented in Table 3. Of these 14, we preregistered only the variance analyses inspecting the standardized variances across all studies under both the  $SD_z$  and  $max_z - min_z$  operationalizations, assuming homogeneous population variances (<https://osf.io/tshx8/>). All other analyses have not been preregistered and should therefore be considered exploratory.

Table 3: Area Under Receiving Operator Characteristic (AUROC) values for each variance analysis and operationalization, including its 95% Confidence Interval. Heterogeneity assumes population variances differ for the low- and high anchoring conditions, whereas homogeneity assumes equal population variances across anchoring conditions.

Population variance assumption	Study	$SD_z$	$max_z - min_z$
Heterogeneity	Overall	0.761 [0.733-0.788]	0.827 [0.8-0.853]
Homogeneity	Overall	0.264 [0.235-0.293]	0.544 [0.507-0.58]
Homogeneity	Study 1	0.373 [0.339-0.406]	0.488 [0.474-0.502]
Homogeneity	Study 2	0.395 [0.36-0.429]	0.634 [0.608-0.66]
Homogeneity	Study 3	0.498 [0.463-0.533]	0.563 [0.539-0.588]
Homogeneity	Study 4	0.401 [0.367-0.435]	0.561 [0.527-0.594]
Heterogeneity	Study 1, low anchoring	0.438 [0.406-0.47]	0.487 [0.481-0.493]
Heterogeneity	Study 1, high anchoring	0.615 [0.582-0.647]	0.501 [0.492-0.51]
Heterogeneity	Study 2, low anchoring	0.652 [0.621-0.683]	0.625 [0.607-0.643]
Heterogeneity	Study 2, high anchoring	0.556 [0.523-0.589]	0.528 [0.515-0.541]
Heterogeneity	Study 3, low anchoring	0.643 [0.612-0.674]	0.542 [0.53-0.553]
Heterogeneity	Study 3, high anchoring	0.747 [0.719-0.775]	0.691 [0.669-0.712]
Heterogeneity	Study 4, low anchoring	0.667 [0.636-0.697]	0.595 [0.577-0.614]
Heterogeneity	Study 4, high anchoring	0.798 [0.773-0.823]	0.756 [0.733-0.779]

Results indicate that (1) heterogeneity of population variances directly affects the efficacy of detecting data fabrication with variance analyses, (2)  $max_z - min_z$  is relatively more robust to violations of homogeneity than  $SD_z$ , and (3) that there is considerable fluctuation across subgroup results. When comparing the combined variance analyses, we see that the homogeneity assumption drastically decreases the classification performance if violated ( $SD_z$ ; homogeneity:  $AUROC = 0.264$ , 95% CI [0.235-0.293], heterogeneity:  $AUROC = 0.761$ , 95% CI [0.733-0.788]), but less so for the  $max_z - min_z$  operationalization ( $max_z - min_z$ ; homogeneity:  $AUROC = 0.544$ , 95% CI [0.507-0.58], heterogeneity:  $AUROC = 0.827$ , 95% CI [0.8-0.853]). Lastly, we see that variance analyses separated per study or anchoring condition within a study are quite variable (ranging from 0.373-0.798), which suggests that a combined analysis of variances across homogeneous subsets of standard deviations is preferred. After all, a priori selection of one specific subset seems infeasible in practice.

Overall, variance analyses work fairly well if the homogeneity assumption is fulfilled for subgroups and all variances available are analyzed jointly. More specifically, we see that both the  $SD_z$  and  $max_z - min_z$  operationalizations perform approximately the same ( $AUROC = 0.761$ , 95% CI [0.733-0.788] and  $AUROC = 0.827$ , 95% CI [0.8-0.853], respectively). Given that  $max_z - min_z$  seems to be relatively more robust to violations of the assumption of equal variances, we recommend using that over the previously proposed  $SD_z$  operationalization (Simonsohn, 2013).

### Combining p-value- and variance analyses

We combined the results from the different variance analyses with the  $p$ -value analyses of the nonsignificant gender- and interaction effects. Because we conducted combined and subsetted variance analyses under two different assumptions (homogeneous- or heterogeneous variances across anchoring conditions), we also conducted a set of combinations. Table 4 depicts the results for each of these combined analyses. In the first two rows, we depict the combination of the  $p$ -value analyses with the reversed Fisher method of the nonsignificant gender effect and the nonsignificant interaction effect, and the overall variance analyses under heterogeneity of variances (first and second row) or homogeneity of variances (third and

fourth row). The variance analyses are either combined over all studies (first and third row) or split per study/subgroup (second and fourth row, respectively).

Table 4: Area Under Receiving Operator Characteristic (AUROC) values for the various combined  $p$ -value- and variance analyses, with corresponding 95% Confidence Intervals. Heterogeneity assumes population variances differ for the low- and high anchoring conditions, whereas homogeneity assumes equal population variances across anchoring conditions. Overall indicates that the variance analysis was conducted across all studies simultaneously. Split indicates the variance analyses are separated per study or per anchoring condition, for homogeneous and heterogeneous approaches, respectively.

comb	study	AUROC
Gender, interaction, variance $SD_z$ (heterogeneity, overall, $k = 1$ )	Overall	0.647 [0.616-0.677]
Gender, interaction, variance $SD_z$ (heterogeneity, split, $k = 8$ )	Overall	0.684 [0.655-0.714]
Gender, interaction, variance $SD_z$ (homogeneity, overall, $k = 1$ )	Overall	0.58 [0.548-0.611]
Gender, interaction, variance $SD_z$ (homogeneity, split, $k = 4$ )	Overall	0.605 [0.573-0.636]

Results as presented in Table 4 indicate the combination of the  $p$ -value analyses and variance analyses performs poorly. This combination method performs only as well as its constituent methods and would be expected to decrease performance if it introduces more noise than signal. Given that the  $p$ -value analyses of the gender- and interaction effects already performed at chance level, and the variance analyses performed reasonably poor for all but the combined method with subgroups, it would be expected that this combination would not be more promising in detecting data fabrication.

## Effect sizes

Using the statistically significant effect sizes from the anchoring studies, we are able to differentiate between the fabricated- and genuine results fairly well. Figure 4 (middle column, second row) indicates that the fabricated statistically significant effects are considerably different. If we inspect the effect size distributions ( $r$ ), we see that the median fabricated effect size is 0.891 whereas the median genuine effect size is 0.661 (median difference= 0.23). In contrast to the fabricated nonsignificant effects, which resembled the genuine data more, the statistically significant effects seem to have been harder to fabricate for the participants. We asked participants to fabricate statistically significant main effects for each of the four anchoring studies; our results indicate that effect sizes across the four studies show consistent results in differentiating between fabricated- and genuine results. More specifically, we see that the *AUROC* for the studies approximate .75 each (0.743, 95% CI [0.712-0.774]; 0.734, 95% CI [0.702-0.767]; 0.737, 95% CI [0.706-0.768]; 0.755, 95% CI [0.724-0.786]; respectively). In other words, given a randomly drawn genuine- and fabricated anchoring effect size, there is approximately a 75% chance that the larger effect size is the fabricated one in this sample. Based on these results, it seems that using effect sizes to detect data fabrication is a parsimonious and fairly effective method.

## Fabricating effects with Random Number Generators (RNGs)

Fabricated effects might seem more genuine when participants used Random Number Generators (RNGs). RNGs are typically used in computer-based simulation procedures where data is generated that are supposed to arise from probabilistic processes. Given that our framework of detecting data fabrication rests on the lack of intuitive understanding of humans at drawing values from probability distributions, those participants who used an RNG might come closer to fabricating seemingly genuine data. Hence, those data might be harder to detect.

We split analyses for those participants who did and those who did not use random number generators. In our sample of 39 participants, 11 used RNGs. Figure 5 shows the same density distributions as in Figure

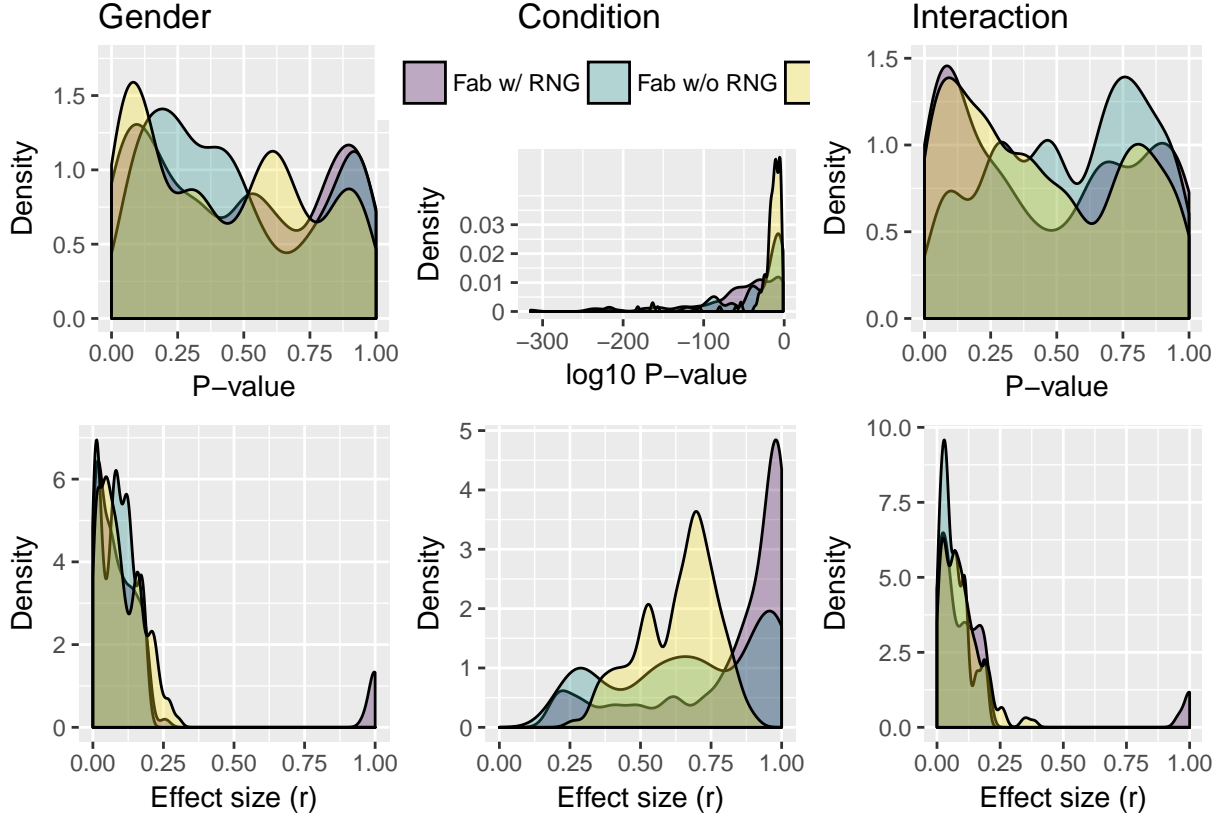


Figure 5: Overlay of (smoothed) density distributions for fabricated results using random number generators (RNGs), fabricated results without using RNGs, and genuine effects. These are split per effect and type of result. Respondents self-selected to use (or not use) RNGs in their fabrication process.

4, except that this time the density distributions of the fabricated data are split into those fabricated with RNGs and those without RNGs (according to self-report by the participants).

Based on Figure 5 we conclude that using RNGs creates less exaggerated summary statistics, but still larger than genuine ones. Based on Figure 5 it seems that the use of RNGs produces somewhat more uniformly distributed statistically nonsignificant  $p$ -values than those without RNGs, but that difference is not confirmed by the AUROC values (gender, with RNG  $AUROC = 0.455$  95% CI [0.405-0.504], without RNG  $AUROC = 0.52$  95% CI [0.482-0.557]; interaction, with RNG  $AUROC = 0.601$  95% CI [0.558-0.644], without RNG  $AUROC = 0.482$  95% CI [0.444-0.52]). For the best performing variance analysis (i.e., heterogeneity over all four anchoring studies with  $max_z - min_z$  operationalization) classification performance is barely different between those data fabricated with ( $AUROC = 0.78$  95% CI [0.728-0.833]) or without RNGs ( $AUROC = 0.845$  95% CI [0.817-0.874]). For effect sizes, Table 5 specifies the differences in sample estimates of the AUROC between the groups of fabricated results with and without RNGs (as compared to the genuine data). These results indicate that the participants who used RNGs are relatively more difficult to detect as fabricated (mean probability of 0.604 that the larger effect is fabricated if presented with one genuine and fabricated effect size), when compared to the participants who did not use a RNG (mean probability of 0.797 that the larger effect is fabricated if presented with one genuine and fabricated effect size). Based on these results, it seems that only effect sizes become substantially less effective at detecting fabricated data.

Study	AUROC RNG, k=11	AUROC no RNG, k=28
-------	-----------------	--------------------

Table 5: AUROC values for detecting data fabrication based on effect sizes for those participants who used Random Number Generators (RNGs) and those participants who did not use RNGs, including 95% Confidence Interval. Split based on self-report data on whether RNGs were used by the participant.

Study	AUROC RNG, k=11	AUROC no RNG, k=28
Study 1	0.553 [0.489-0.617]	0.817 [0.785-0.85]
Study 2	0.641 [0.578-0.705]	0.771 [0.734-0.807]
Study 3	0.578 [0.512-0.645]	0.8 [0.767-0.832]
Study 4	0.641 [0.581-0.702]	0.8 [0.764-0.835]

## Discussion

We presented the first controlled study on detecting data fabrication at an individual data set level. As far as we could find, previous efforts only looked at group-level comparisons of genuine- and fabricated data (Akhtar-Danesh & Dehghan-Kooshkghazi, 2003), inspected properties of individually fabricated sets of data without comparing them to genuine data, or did not contextualize these data in a realistic study with specific hypotheses (Mosimann et al., 1995). We explicitly asked researchers to fabricate results for an actual effect within their research domain, which was contextualized in realistic hypotheses, and was compared to genuine data on the same effect.

We applied various statistical methods to classify genuine- from fabricated data and found that those related to statistically significant summary statistics performed fairly well. The results of the reversed Fisher method on the statistically nonsignificant effects performed at chance level. Variance analyses and the statistically significant effects, on the other hand, performed fairly well at classifying fabricated from genuine data. Variance analyses performed marginally better than using significant effect sizes in this sample.

Using a Random Number Generator (RNG) to fabricate summary statistics could decrease the probability of detecting a fabricated dataset, depending on the type of analysis. Using RNGs substantially decreased the performance of using effect sizes to classify fabricated- from genuine data. This indicates that data fabricated by humans without RNGs might be excessively bold. However, it also showcases that methods to detect data fabrication with effect sizes is potentially likely to fail when RNGs are involved. On the other hand, using RNGs did not substantially decrease the performance of the variance analysis that analyzed the subsetted anchoring conditions. We will investigate in Study 2 whether this is similar for raw data and revisit this issue in the general discussion.

For the reversed Fisher method, results indicated that our prediction was wrong and that participants did not fabricate excessive amounts of high  $p$ -values when told to fabricate statistically nonsignificant effects. More specifically, the analysis of nonsignificant  $p$ -values appeared to perform at chance level, disputing our prediction that the misinterpretation of  $p$ -values as “the probability that there is an effect” would lead to more high  $p$ -values. Further research might investigate how  $p$ -value interpretations affect the expected distribution of the  $p$ -values to investigate what effect the interpretation of  $p$ -values has on the expected distribution. Additionally, given that the  $p$ -value analyses we proposed conceptually resembles analyses proposed by Carlisle (2012), our result suggests caution in analyzing for too many high  $p$ -values outside of randomisation tests such as Carlisle performed.

Specific to the variance analyses, we noted that the assumption of homogeneous population variances had not previously been explicated nor tested for robustness to violations. In Simonsohn (2013) it remains implicit that the variances grouped together in an analysis should arise from a homogeneous population distribution. Our results indicate that the classification performance of variance analyses is considerably dependent on fulfilling this assumption. The alternative operationalization we included inspects the range of standard deviations ( $max_z - min_z$ ) instead of the variance of standard deviations ( $SD_z$ ), which seemed to be more robust to violations of this assumption. Hence, we recommend to use variance analyses



with subgrouping of variances into those that are likely to be from the same population distribution (e.g., based on anchoring condition here) and use the range of standard deviations ( $max_z - min_z$ ).

We note that the presented results might be particular to the anchoring effect and not replicable with other effects. After all, mental fabrication strategies may be dependent on the type of effect or measurement that is being fabricated. In the anchoring studies, data needed to be fabricated for numbers that are ranged in the hundreds or thousands. Such relatively large values might feel more unintuitive to think about than smaller numbers in the singles or tens. Hence, our results might be better at detecting data fabrication because of this increased lack of intuitiveness. Other kinds of studies that are easier for fabricators to think about in terms of fabricating realistic data might prove more difficult to classify. For example, we might question how results based on Likert scale items might show different kinds of results from these anchoring studies.

With respect to our study design, we discovered that we included several non-U.S. researcher against our initial aim. We filtered Web of Science on U.S. origin, but found out that this meant that one of the authors on the paper was U.S. based. As such, corresponding authors might still be non-U.S. Based on a search through the open ended comments of the participant's responses, there was no mention of issues in fabricating the data related to the metric or imperial system. Hence, there is no explicit reason to assume participants had issues with fabricating imperial measurements if they use the metric system in their daily life. We also discovered that some of the original Many Labs trials converted metric estimates to imperial estimates.

Despite testing various statistical methods to detect data fabrication, we did not test all available statistical methods to detect data fabrication in summary statistics. SPRITE (J. A. Heathers, Anaya, Zee, & Brown, 2018), GRIM (N. J. L. Brown & Heathers, 2016), and GRIMMER (Anaya, 2016) are some examples of other statistical methods that test for faulty or fabricated summary statistics (see also Buyse et al., 1999). However, these methods were not applicable in the studies we presented, because they require ordinal scale measures. It seems that, combined with the question of whether current results of detecting fabricated data replicate in Likert scale studies, validating these other methods would be a fruitful avenue for further research.

## Study 2 - detecting fabricated raw data

In Study 2 we tested the performance of statistical methods to detect data fabrication in raw data. Our procedure is comparable to Study 1: We asked actual researchers to fabricate data that they thought would go undetected. However, instead of summary statistics, in Study 2 we asked participants to fabricate lower level data (i.e., raw data) and included a face-to-face interview (Chris H J Hartgerink, Voelkel, Wicherts, & Assen, 2017). A preregistration of this study occurred during the seeking of funding (Hartgerink, Wicherts, & Assen, 2016) and during data collection (<https://osf.io/fc35g>). Just like Study 1, this study was approved by the Tilburg Ethical Review Board (EC-2015.50; <https://osf.io/7tg8g/>).

To test the validity of statistical methods to detect data fabrication in raw data, we investigated raw data of Stroop experiments (Stroop, 1935). In a Stroop experiment, participants are asked to determine the color a word is presented in (i.e., word colors) and where 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 (e.g., 'red' presented in green). The dependent variable in a Stroop experiment is the response latency, typically in milliseconds. Participants in actual Stroop studies are usually presented with a set of these Stroop tasks, where the mean and standard deviation per condition serve as the raw data for analyses (see also Ebersole et al., 2016). The Stroop effect is often computed as the difference in mean response latencies between the congruent and incongruent conditions.

## Methods

### Data collection

We collected twenty-one genuine data sets on the Stroop task 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 data sets, accidentally overlooking the

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 6: Example of a filled out 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>.

online sample; Hartgerink et al., 2016). 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 each participant their response latencies in both within-subjects conditions of congruent trials and incongruent trials (i.e., two M-SD combinations for each participant). This format was also the basis for the template spreadsheet that we requested participants to use to supply the fabricated data (see also Figure 6 or <https://osf.io/2qrbs/>). We calculated the Stroop effect as a *t*-test of the difference between the congruent and incongruent conditions ( $H_0 : \mu_{\bar{X}_1 - \bar{X}_2} = 0$ ).

We collected twenty-eight fabricated data sets on the Stroop task 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 face-to-face interviews on how the data were fabricated. We chose the period 2005-2015 to prevent a decrease in the probability that the corresponding author would still be reachable via the given corresponding e-mail address. 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. Subsequently, we implemented a second, unplanned, and not ethically reviewed 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' quota 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 or 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 6; <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. JGV transcribed the contents of the interview, CHJH blind-reviewed these transcripts to remove any potentially personally identifiable information (these transcripts are freely available for anyone to use at <https://doi.org/10.5281/zenodo.832490>)(<https://doi.org/10.5281/zenodo.832490>)).

## Data analysis

To detect data fabrication in raw data using statistical tools, we performed a total of sixteen analyses (preregistration: <https://osf.io/ecxvn/>). These sixteen analyses consisted of four NBL digit analyses, four terminal digit analyses, two variance analyses, four multivariate association analyses (deviated from preregistration; used parametric instead of non-parametric approach), a combination test of these methods, and effect sizes at the summary statistics level (the latter replicating Study 1 and was not preregistered).

For the digit analyses, we separated the *Ms* and *SDs* per within-subjects condition and conducted  $\chi^2$ -tests for each per data set. As such, for one data set, we conducted digit analyses on the 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 NBL, we used the first (or leading) digit, whereas for the terminal digit analyses we tested the same sets but on the final digit.

For the variance analyses, we analyzed the standard deviations of the response latencies separated for the within-subjects conditions. That is, we analyzed the standard deviations of the response latencies in the congruent condition for excessive consistency separately from the standard deviations of the incongruent condition. We conducted this analysis for each genuine- or fabricated dataset, using the  $max_z - min_z$  operationalization (not preregistered; based on results from Study 1 indicating it is more robust to violations of the assumption of equal variances).

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 the (i) correlation between the means across conditions, (ii) standard deviations across conditions, (iii) means and standard deviations within the congruent condition, and (iv) means and standard deviations within the incongruent condition. We did this by computing a random-effects estimate of the observed (Fisher transformed) correlations from the Many Labs 3 data. The estimated effect distribution served as the parametric model for each of those four relations under investigation ( $N \sim (\mu, \tau)$ ). Using the estimated parametric distribution, we computed two-tailed *p*-values for each fabricated- and genuine dataset.

We also combined the terminal digit analyses, the variance analyses, and the analyses based on multivariate associations using the Fisher method. More specifically, we included the *p*-values of 10 statistical tests: Four terminal digit analyses, two variance analyses, and four analyses of the multivariate associations. We excluded the NBL digit analyses because we a priori expected that psychological measures (e.g., response times) are rarely true ratio scales with sufficient range to show the NBL properties in the first digit, hence that this type of analysis would not be productive in detecting data fabrication in these types of data (preregistration: [doi.org/10.3897/rio.2.e8860](https://doi.org/10.3897/rio.2.e8860)).

Study 1 showed that effect sizes are a potentially valuable tool to detect data fabrication, which we exploratively replicate in Study 2. This was not preregistered because we had not yet determined results of

Study 1 before designing Study 2. Based on the genuine- and fabricated data sets, we computed effect sizes for the Stroop effect based on the effect computation from the Many Labs 3 scripts (<https://osf.io/qs8tp/>). Using a  $t$ -test of the difference between the congruent and incongruent conditions ( $H_0 : \mu = 0$ ) we computed the  $t$ -value and its constituent effect size as a correlation using (Chris H. J. Hartgerink et al., 2017)

$$r = \sqrt{\frac{\frac{F \times df_1}{df_2}}{\frac{F \times df_1}{df_2} + 1}}$$

where  $df_1 = 1$ ,  $F = t^2$ , and  $df_2$  is the degrees of freedom of the  $t$ -test. We can simplify the effect size calculation to

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

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 regard AUROC values  $< .7$  as poor for detecting data fabrication,  $.7 \leq AUROC < .8$  as fair,  $.8 \leq AUROC < .9$  as good, and  $AUROC \geq .9$  as excellent (Youngstrom, 2013). We also explore whether using Random Number Generators (RNGs) affects the detection of fabricated data in our sample. We conducted all analyses using the `pROC` package (Robin et al., 2011).

## Results

### Digit analyses

Figure 7 shows the aggregated first digit distributions of the genuine- and fabricated data side-by-side with the expected first digit distributions according to the NBL. In the first row the first digit distributions of the means are presented, for both the congruent condition (left column) and incongruent condition (right column). The first row indicates that the first digit distributions of the genuine- and fabricated mean response times do not adhere to the NBL. The first digit distributions of the standard deviations (second row) adhere to the NBL more than the means at first glance, but still deviate substantially from what would be expected according to the NBL. These aggregate results already suggest that using the NBL to test for data fabrication is definitely not appropriate (means) or might be appropriate (standard deviations).

The AUROC results indicate that using the Newcomb-Benford Law is at best on par with chance level classification of genuine- and fabricated data. More specifically, for the congruent standard deviations, using the results of the NBL test are on par with chance classification ( $AUROC = 0.553$ , 95% CI [0.389-0.717]). Values from other measures showcase that the fabricated data are actually more in line with the NBL than the genuine data. After all, the NBL test is directional, where fabricated data would be expected to be less in line with the NBL and therefore yield smaller  $p$ -values, hence we also classified that way. However, for the (in)congruent means and the incongruent standard deviations the AUROC values indicate that we are classifying the wrong way with this test (congruent means:  $AUROC = 0.039$ , 95% CI [0-0.087]; incongruent means:  $AUROC = 0.024$ , 95% CI [0-0.059]; incongruent standard deviations:  $AUROC = 0.156$ , 95% CI [0.045-0.268]).

Figure 8 shows the aggregated terminal digit distributions of the genuine- and fabricated data side-by-side with the expected terminal digit distributions. In the first row the terminal digit distributions of the means are presented, for both the congruent (left column) and incongruent (right column) conditions. The first row shows that the terminal digit distributions of the genuine- and fabricated mean response times are approximately uniform with only minor differences between the genuine- and fabricated data. The terminal digit distributions of the standard deviations (second row) show slightly more deviation from uniformly distributed digits, but still approximate the expected distribution of terminal digits reasonably well. Based on these aggregate digit distributions, it seems like the classification based on the terminal digit analyses will not be able to differentiate between genuine- and fabricated data particularly well.

The AUROC results indicate that using terminal digit analyses is also at best on par with chance level classification of genuine- and fabricated data. More specifically, for the incongruent standard deviations,

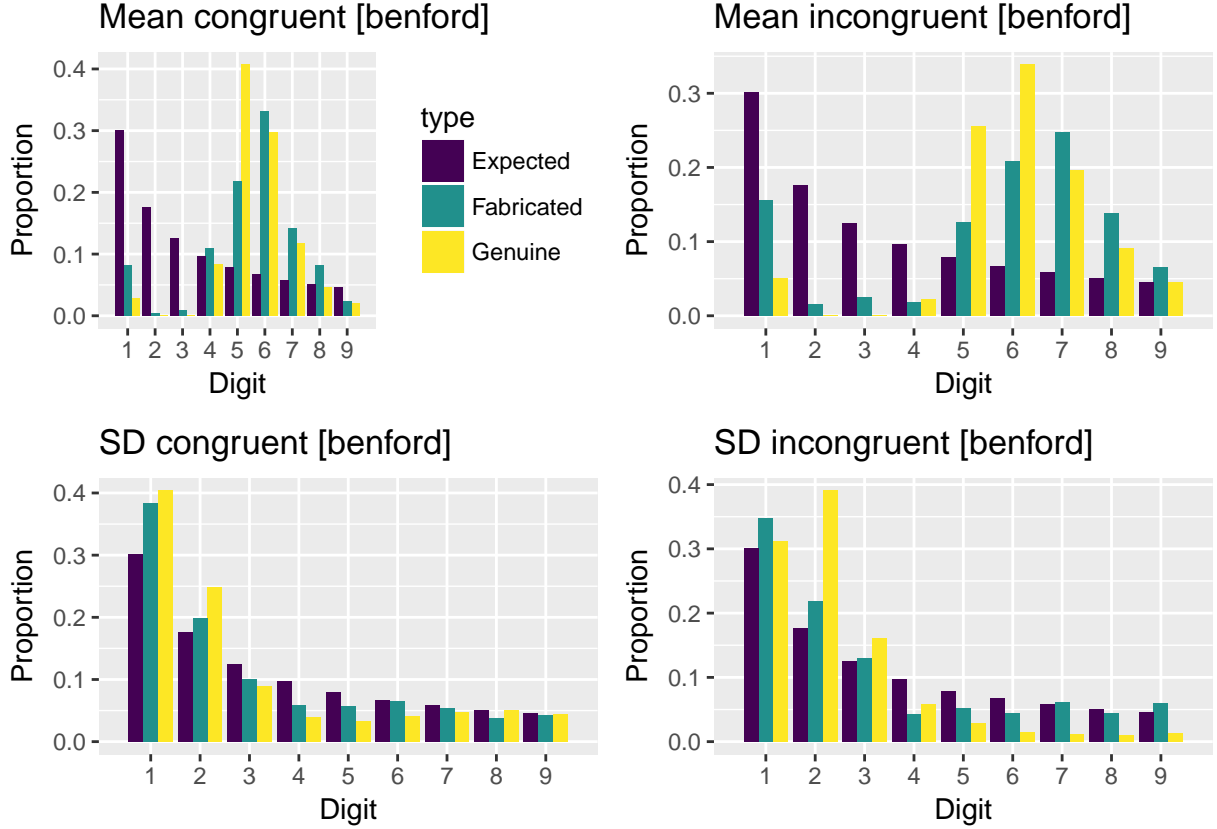


Figure 7: First digit distributions for the (in)congruent means and standard deviations, aggregated across all Many Labs 3 datasets or across the datasets fabricated by the participants.

using the results of the terminal digit analysis are on par with chance classification ( $AUROC = 0.511$ , 95% CI [0.343-0.679]). Other conditions showcase that the fabricated data are slightly more in line with a uniform digit distribution than the genuine data. The terminal digit test is directional (similar to the NBL), where fabricated data would be expected to be less uniform and therefore yield smaller  $p$ -values. However, for the (in)congruent means and the incongruent standard deviations the AUROC values indicate that we are classifying the wrong way with this test, albeit less so than with the NBL results presented before (congruent means:  $AUROC = 0.383$ , 95% CI [0.222-0.543]; incongruent means:  $AUROC = 0.387$ , 95% CI [0.226-0.548]; congruent standard deviations:  $AUROC = 0.401$ , 95% CI [0.241-0.562]).

### Variance analysis

Results indicate that the fabricated- and genuine data can be perfectly separated based on the variance analyses. More specifically, the AUROC of both the variance analyses for the congruent standard deviations and the incongruent standard deviations is  $AUROC = 1$  (confidence intervals cannot be reliably computed in this case). We note that these results are likely to be sample specific and do not mean to imply that this method will always be able to separate the genuine- from fabricated data perfectly. However, they also indicate that given the number of standard deviations participants had to fabricate ( $k = 25$ ), it was difficult for participants to make them look similar to those found in the genuine data. Nonetheless, upon closer inspection, results of the variance analyses showed a maximum  $p$ -value of 0.006 across both the genuine- and the fabricated data, which indicates that the variance analysis is oversensitive in the absolute sense, despite classifying so well in a relative sense. After all, if we would use the traditional notion of hypothesis testing and say  $p < \alpha$  leads to the conclusion that the data are fabricated,  $\alpha = .01$  would cause all included genuine- and fabricated datasets to be regarded as fabricated.

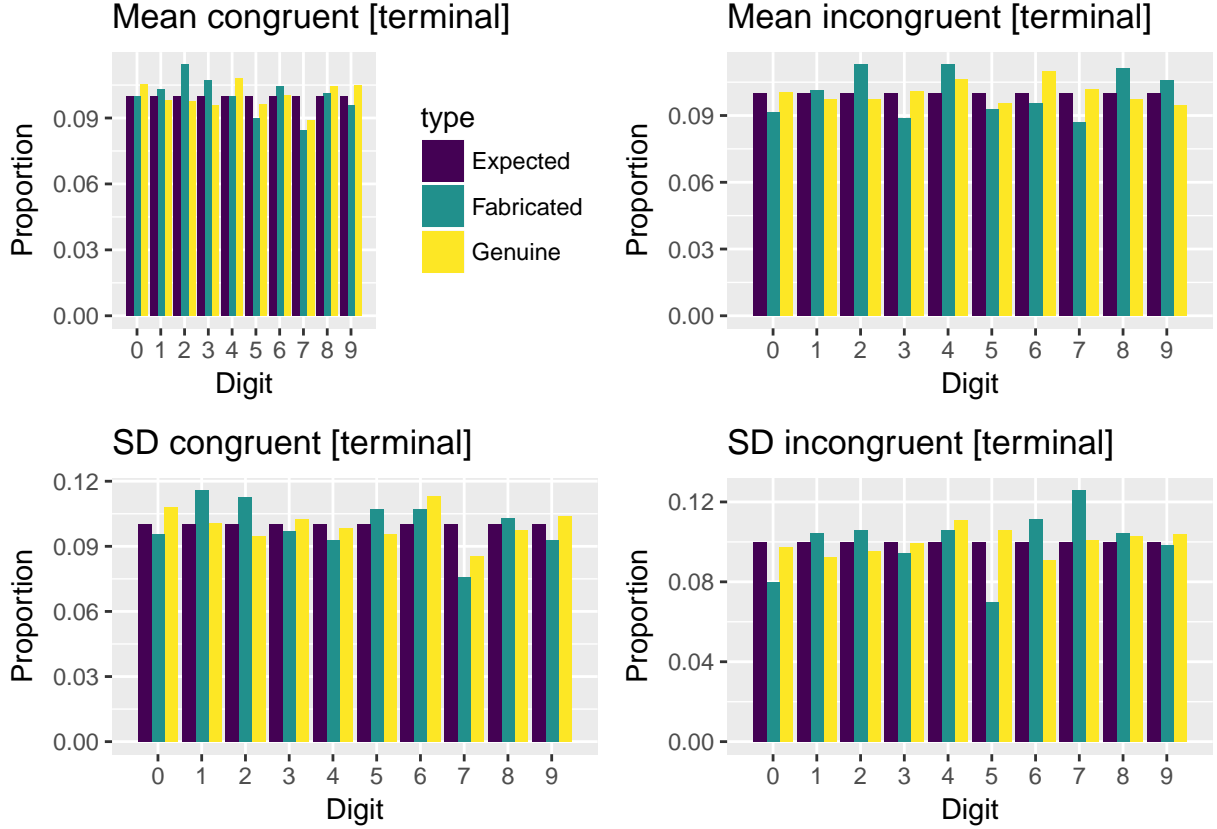


Figure 8: Terminal digit distributions for the (in)congruent means and standard deviations, aggregated across all Many Labs 3 datasets or across the datasets fabricated by the participants.

### Multivariate associations

Using the parametric test of multivariate associations, results indicate classification is fair to good in the current sample. Figure 9 shows the density distributions of the various multivariate associations (rows 1-2), which already indicates the genuine data are less spread out and more normally distributed when compared to the fabricated multivariate associations. Using the parametric estimates of the associations to test the various sets of multivariate relations between the (in)congruent means and standard deviations, AUROC values range from 0.714 through 0.842. More specifically, the AUROC for the various sets of relations (going clockwise with the first four figures in Figure 9) are  $AUROC = 0.818$ , 95% CI [0.689-0.947] for M-SD in the congruent condition,  $AUROC = 0.833$ , 95% CI [0.705-0.962] for M-SD in the incongruent condition,  $AUROC = 0.714$ , 95% CI [0.568-0.861] for M-M across conditions,  $AUROC = 0.842$ , 95% CI [0.732-0.952] for SD-SD across conditions.

### Combining variance, terminal digit, and associational analyses

We combined both variance analyses, the terminal digit analyses, and the tests of the multivariate associations with the Fisher method. More specifically, we combined the results of the congruent- and incongruent variance analyses (i.e., two per dataset); the terminal digit analysis of the congruent and incongruent means and standard deviations (i.e., four per dataset); the parametric tests of the relation between the (in)congruent means and standard deviations, the relation between the means across conditions, and the relation between the standard deviations across conditions (i.e., four per dataset).

Results of the combined analysis perform excellent at classifying fabricated- and genuine data in this sample. More specifically, the results for the combination method indicate  $AUROC = 0.959$  (95% CI [0.912-1]). This combination method is affected by the effectiveness of the individual methods involved; given that the performance of the multivariate associations and variance analyses ranged from sufficient to excellent, it makes sense that this combination method also performs quite well. However, also here we

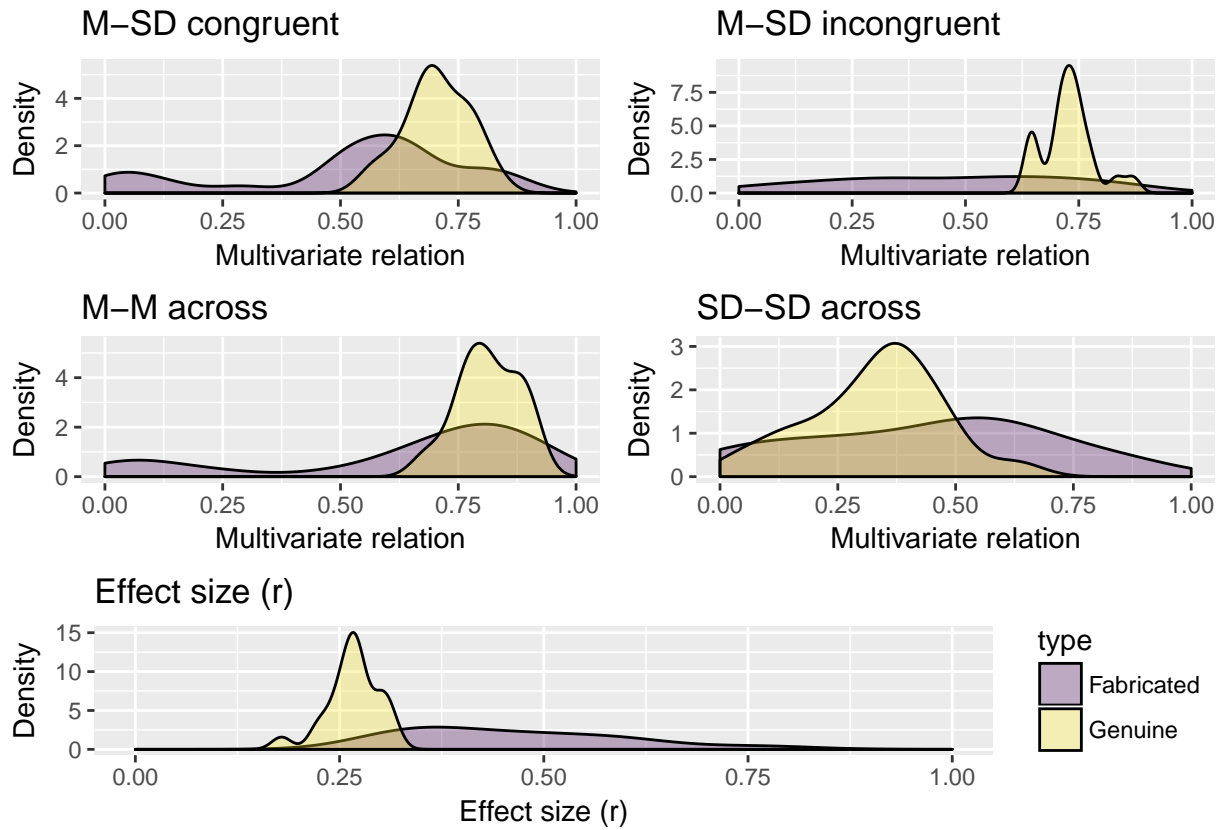


Figure 9: Density distributions of the multivariate relations (first two rows) and the effect sizes (final row), split for the genuine and fabricated data.

note that the maximum  $p$ -value of the combination of these tests is 0.003, indicating that in a traditional hypothesis testing sense it is oversensitive at regular alpha levels (i.e., .05 or .01).

### Effect sizes

Figure 9 (final row) shows the density distributions of the fabricated- and genuine Stroop effect sizes, which is an excellent classifier of fabricated/genuine data in this sample. More specifically, the classification performance for detecting fabricated data in this sample is  $AUROC = 0.981$ , 95% CI [0.954-1] (the 95% CI is truncated at 1). Upon closer inspection of the effect sizes, we note that only three (of 28) fabricated effect sizes fall within the range of genuine effect sizes. As such, this is a particularly good result within this sample.

### Fabricating effects with Random Number Generators (RNGs)

Table 6: AUROC values with 95% confidence intervals for each test, when split for those with Random Number Generators (RNGs) and those without.

Test	With RNG (k=19)	Without RNG (k=9)
Benford, congruent means	0.035 [0-0.087]	0.048 [0-0.144]
Benford, congruent sds	0.506 [0.315-0.698]	0.651 [0.431-0.87]
Benford, incongruent means	0.023 [0-0.064]	0.026 [0-0.082]
Benford, incongruent sds	0.115 [0.008-0.223]	0.243 [0.015-0.472]
Combination w Fisher method	0.957 [0.9-1]	0.963 [0.895-1]
Effect size (r)	0.985 [0.957-1]	0.974 [0.918-1]
Multivariate association, M-M across	0.662 [0.481-0.842]	0.825 [0.603-1]

Test	With RNG (k=19)	Without RNG (k=9)
Multivariate association, M-SD congruent	0.85 [0.707-0.992]	0.751 [0.488-1]
Multivariate association, M-SD incongruent	0.802 [0.637-0.967]	0.899 [0.702-1]
Multivariate association, SD-SD across	0.484 [0.272-0.695]	0.688 [0.421-0.955]
Parametric test of Multivariate association, M-M across	0.662 [0.481-0.842]	0.825 [0.603-1]
Parametric test of Multivariate association, M-SD congruent	0.85 [0.707-0.992]	0.751 [0.488-1]
Parametric test of Multivariate association, M-SD incongruent	0.802 [0.637-0.967]	0.899 [0.702-1]
Parametric test of Multivariate association, SD-SD across	0.847 [0.717-0.977]	0.831 [0.671-0.991]
Terminal digits, congruent means	0.388 [0.206-0.57]	0.37 [0.132-0.609]
Terminal digits, congruent sds	0.439 [0.253-0.624]	0.323 [0.087-0.559]
Terminal digits, incongruent means	0.36 [0.186-0.534]	0.444 [0.181-0.708]
Terminal digits, incongruent sds	0.573 [0.383-0.763]	0.381 [0.162-0.6]
Variance analysis, congruent sds (maxmin)	1 [1-1]	1 [1-1]
Variance analysis, incongruent sds (maxmin)	1 [1-1]	1 [1-1]

Using Random Number Generators (RNGs) in the raw data fabrication procedure slightly changes results but not substantively so. AUROC values for all of these tests are presented side by side in Table 6, split on those participants who said they used ( $k = 19$ ) or did not use RNGs ( $k = 9$ ) to fabricate data (based on manual coding of the interview transcripts).. For leading digits, Figure 10 indicates the digit distributions split for those fabricated with and without RNGs and results indicate relatively minor differences; Figure 11 shows the digit distributions of terminal digits, where results again indicate minor differences. Figure 12 shows the same results as Figure 9, where the raw data looks slightly more genuine for multivariate associations fabricated with RNGs when compared to those raw data fabricated without RNG. Going clockwise through Figure 10, we see a slight normalization of results for M-SD congruent associations, virtually no changes for M-SD incongruent associations, a reasonable amount of normalization for M-M associations across conditions, an almost linear increase in SD-SD associations across conditions, and virtually no changes for the fabricated effect sizes. The effect size distribution is approximately similar for both data fabricated with and without RNGs. Given these minor and inconsistent changes to the density distributions, we do not regard RNGs as having substantial effects on the effectiveness of statistical methods to detect data fabrication in this sample.

## Discussion

Our second study investigated how well statistical methods can discern between individual genuine- and fabricated raw data. To that respect, results indicate that digit analyses of raw data perform at chance level, variance analyses of raw data perform excellent, and multivariate relations between variables in the raw data perform fairly to excellent. Moreover, the summary statistic effect size strikes a surprisingly good balance between efficacy and parsimony for classifying fabricated- from genuine raw data (only superseded in performance by the more complex variance analyses). It is somewhat ironic that the summary statistic of the effect performs so well in classifying the genuine- from fabricated data. This replicates the finding from Study 1 that effect sizes are a valuable piece of information to discern genuine- from fabricated data. Random Number Generators (RNGs) had no substantial affect on the classification of fabricated- from genuine raw data.

Our results confirmed our prediction that leading digit analyses (i.e., NBL) are not fruitful in detecting fabricated response times. The Newcomb-Benford Law is frequently observed in various natural phenomena (e.g., population numbers) but Figure 7 (clearly) indicates this is not the case for summary statistics of response times. Response times are untruncated ratio measures in theory that technically should satisfy the NBL's requirements, but in practice response time measures are truncated severely (e.g., nobody can respond within <50 milliseconds and few take longer than 2000 milliseconds). In other words, if there is no evidence that already calibrated the data against the NBL, we recommend against using it as a generic method to detect data fabrication in experimental studies.

Going against our predictions, participants fabricated raw data that was almost indistinguishable from the genuine raw data when looking at terminal digit analyses. Given the theoretical framework we use, where humans are bad at fabricating stochastic processes that underlie data collection procedures, we expected that participants would be unable to fabricate uniformly distributed terminal digits. Our sample



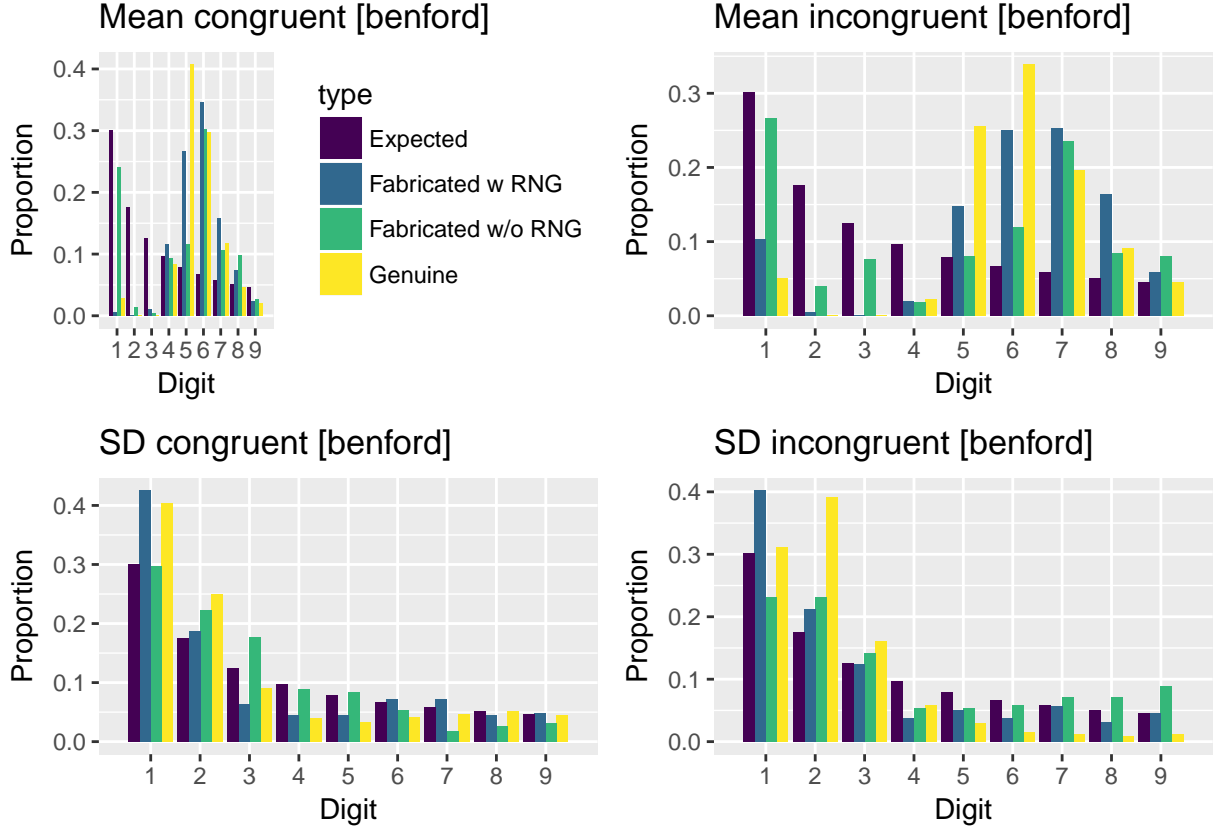


Figure 10: First digit distributions for the (in)congruent means and standard deviations, aggregated across all Many Labs 3 datasets or across the datasets fabricated by the participants.

indicates this is not the case. Moreover, given that these stochastic processes are expected to be better included when data is fabricated with RNGs, it was a surprise that this did not affect classification performance. This raises questions with respect to whether the framework of human’s lack of intuitive understanding of probabilities manifests itself in fabricated raw data.

Study 2 replicated the effectiveness of variance analyses and effect sizes to detect data fabrication, but failed to replicate the effect of RNGs on detection. With these mixed results with respect to the effect of RNGs, we note the same limitation as for the terminal digit analysis, which is that our theoretical framework of intuitions for probabilities might not manifest itself in fabricated data. Hence, further research might look into correlating the (lack of) expertise on probabilities and the kind of data that is fabricated. With respect to variance analyses and effect sizes, our results suggest that these are the most promising methods.

We originally planned to extend Study 2 with a qualitative exploration of the fabrication process. We transcribed all 28 interviews, but due to time constraints did not get around to doing so. We note that all transcripts are available online () and that the initial work can be found online as well. We invite anyone with an interest to look at these documents and build on our work further.

## General discussion

We presented the first two empirical studies on detecting individual sets of fabricated data, where the fabricated data pertained to existing experiments and detection occurred purely by using statistical methods. By comparing results from genuine- and fabricated data across summary statistics and raw data, it seems like classification based on statistically significant effect sizes strikes the best balance between parsimony, effectiveness, and usability. Variance analyses on the other hand are a well performing option that is somewhat more complex in its application. We bundled our functions for the variance- and digit analyses and the (reversed) Fisher method in the `ddf` (short for detecting data fabrication) package for

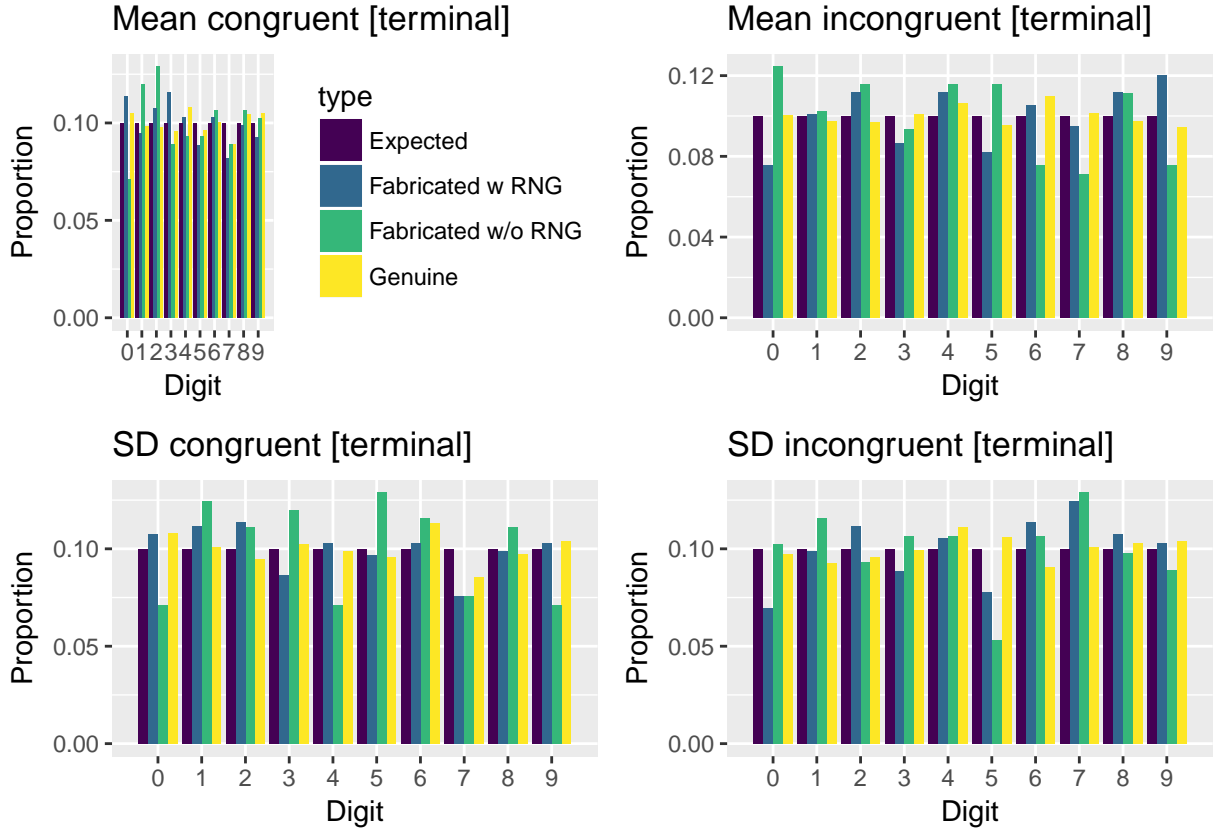


Figure 11: Terminal digit distributions for the (in)congruent means and standard deviations, aggregated across all Many Labs 3 datasets or across the datasets fabricated by the participants.

R, which is available through GitHub (<https://github.com/chartgerink/ddfab>) for application in further research and development.

We designed the current study to have sufficient information to detect data fabrication within a given set of data, but not necessarily to infer our results to a larger population. As such, the sample sizes of the presented studies provide little certainty when used to make inferences. This was not our primary goal, but could be the goal of further research. Our studies have highlighted that variance- and effect size analysis are methods that look promising to detect problematic data. Our descriptive results with confidence intervals may be regarded as an initial step in understanding the effectiveness of these methods to detect data fabrication. Next, we highlight some of the difficulties that remain.

All presented results throughout the two studies pertain to relative comparisons between genuine- and fabricated data. Hence, all statements about the performance of classification is dependent on the availability of unbiased genuine data to compare to and cannot readily be done by using generic decision criteria such as *alpha*-levels. As we saw for example in the variance analyses for Study 2, there was excellent relative classification, but absolute classification as many researchers are used to by comparing  $p < \alpha$  remained problematic. More specifically, we would have classified all datasets as fabricated if we had used the traditional hypothesis testing approach. It is for exactly this reason we refrain from formulating general decision rules for the methods presented in this paper. One potential exception might be for really excessively large effect sizes (e.g.,  $r > .95$ ) as an initial screening of a set of effects.

Because we included the Many Labs data [@@@] we had (assumably) unbiased estimates of the effects under investigation, which is key for relative comparisons. If we had used the peer-reviewed literature on the anchoring effect (Study 1) or the Stroop effect (Study 2), we would likely find inflated effect size estimates of the anchoring- or Stroop effects due to publication bias. These inflated effect size estimates would likely result in worsened classification of genuine- and fabricated data because publication bias results in inflated effect sizes (M. B. Nuijten, Assen, Veldkamp, & Wicherts, 2015) and our studies indicate fabricating data has a similar effect. That they have the same effect in turn conflates the detection of fabricated data. Collecting an unbiased genuine effect distribution thus requires careful attention; when

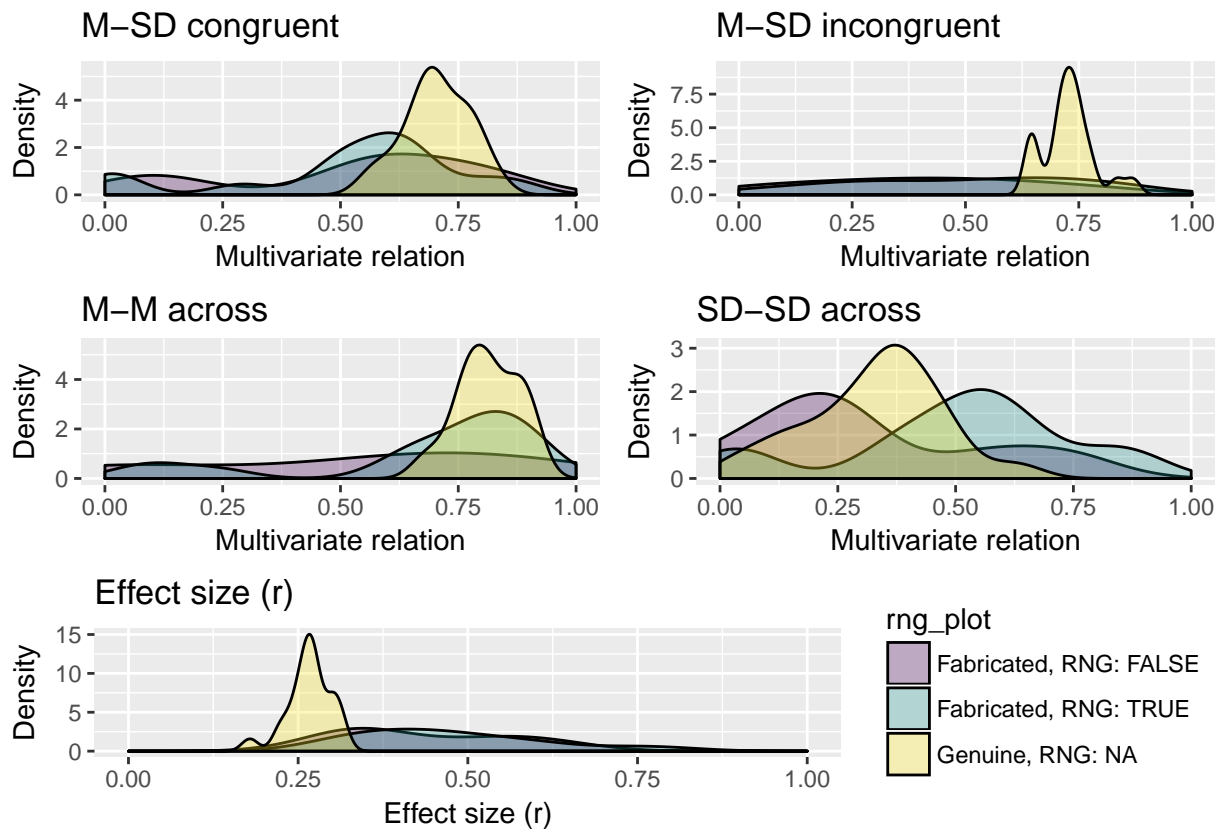


Figure 12: Density distributions of the multivariate relations (first two rows) and the effect sizes (final row), split for the genuine data, the fabricated data without using Random Number Generators RNGs), and fabricated data with using RNGs.

such care is not taken detection rates are likely to be conservative. We recommend retrieving unbiased effect size distributions for an effect from large-scale replication projects, such as Registered Replication Reports (e.g., Cheung et al., 2016) and building systemic efforts to reduce publication bias (see also ???).

Our results depend on the (majority of the) Many Labs data being genuine. If it turns out at any point that this is not the case, our results are invalidated. We remain confident that the Many Labs data are genuine for a variety of reasons of which we present a few next. First, the sheer amount of people involved results in a distribution of responsibility that also limits the effect if one person were to fabricate data. Second, the number of people involved also minimizes the individual reward it would have to fabricate data given that any utility would have to be shared across all researchers involved. Third, the projects actively made all individual research files available and participating researchers in the ML were made aware of this from the very start. Fourth, the analyses of the Many Labs are not conducted by the same individuals who collected the data. We cannot exclude the possibility of malicious actors in the ML studies, but also have no evidence that suggests there would be.

Highly relevant to the application of these kinds of methods beyond these studies is that our study design implies a high prevalence of data fabrication, which directly affects the positive predictive value of these statistical methods. The positive predictive value is the chance of getting a true positive when a positive result is found. More specifically, Study 1 by design has a prevalence of 52% of data fabrication and Study 2 has a prevalence of 57%. This greatly affects the positive predictive value (PPV) of these methods if they would be applied in a more general setting. After all, even if we could classify all fabricated data correctly and falsely regard genuine data as fabricated in 5% of the cases, then with a prevalence of 2% (Fanelli, 2009) the positive predictive value would only be 29%. This is a best case scenario that would cause approximately 1 out of 3 cases of ‘detected data fabrication’ to be false. Given that we saw that the variance analyses for example are oversensitive, it is likely that PPV will be even lower. The likely low prevalence of data fabrication thus makes widespread application of any statistical method to detect data fabrication problematic.

We recommend against widespread application of these statistical methods to detect potential data fabrication. As we explained just now, a low prevalence will drastically decrease its effectiveness. If prevalence appears to be much higher than expected, we might revisit this recommendation. These statistical methods used throughout our studies can still be used as screening tools in review processes and as tools in formal misconduct investigations where prevalence is supposedly higher than in the general population of research results. As we mentioned before, excessively large effect sizes might be used as a screening approach for further manual investigation, but we note that results from these individual tests might provide confirmation bias in any subsequent inspection of the results. As such, if any of these statistical tools are used, we recommend to solely use them to screen for indications of potential data anomalies, which are subsequently further inspected by a blinded researcher to prevent confirmation bias.

We note that our studies have been regarded as unethical by some due to the nature of asking participants to fabricate data (see for example Naomi Ellemers, 2017). We understand and respect that asking researchers to show one of the most widely condemned scientific behaviors is risky. While designing these studies, we also asked ourselves whether this was an appropriate design and ultimately regarded it was for a set of reasons. First, there was little utility in simulating potential data fabrication strategies because there is little to no knowledge of how researchers actually fabricate data. Second, the cases of data fabrication known to us are severely self-selected (i.e., based on detection bias), which would limit the ecological validity of any tests we could do on those. These two reasons made it necessary for us to collect fabricated data. After we had come to that decision, we also regarded that we should minimize the negative effect it had on the researchers participating. We attempted to minimize this negative effect by using findings from psychology research to decrease potential carry-over of this controlled misbehavior. Despite that some of our participants indicated that they felt initial unease with fabricating data for the study, no participants reached out afterwards indicating feeling conflicted. Moreover, we actively attempt to maximize returns of the data collected by sharing all the information we gathered openly and without restrictions. We consider these reasons to balance the design and ask of our study from our participants.

Another ethical issue is the dual use of these kinds of statistical methods to detect data fabrication. Dual use is the ethical issue where the development of knowledge can be used for both good- and evil purposes, hence, whether we should want to morally conduct this research. A traditional example is the research into biological agents that might be used for chemical warfare. For our research, a data fabricator might use our research to test their fabricated data until it goes undetected based on these methods. There is no inherent way to control whether malicious actors do this and one might argue that this is sufficient

reason to not conduct this kind of research to begin with. However, borrowing from copyright law (???), we argue that the potential ethical uses of these methods (improved detection of fabricated data by a potential many) outweigh the potential unethical uses of these methods (undermining detection by a potential few). Secrecy in this respect would actually enhance the ability of malicious actors to remain undetected, because when they find a way to exploit the system fewer people can investigate suspicions they might have. Hence, we regard the ethical issue of dual use to ultimately weigh in favor of doing the research, although we recognize that this might start a competition in undermining detection of problematic data.

Some of our participants in Study 2 indicated using the Many Labs (or other open) data to fabricate their own dataset. During the interviews, some participants indicated that they thought this would make it more difficult to detect their data as fabricated. We did not investigate evidence for this claim specifically (this could be avenue for further research) but we note that our detection in Study 2 performed well despite some participants using genuine data. Moreover, we note that open data might actually facilitate the detection of fabricated data for two reasons. First, open data improves the unbiased estimation of effect sizes, where the peer-reviewed literature inflates estimated effect sizes due to publication bias. As we mentioned before, having these unbiased effect size estimates seem key to detecting issues. Second, if data are fabricated based on existing data, it is more likely to be detected if it is based on open data than when based on closed data. For example, in the LaCour case, data was fabricated based on open data [4]. Researchers detected that this data had been fabricated because it seemed to be a(n almost) linear transformation of an open dataset [4]. As such, we see no concrete evidence to support the claim that open data could lead to worsened detection of fabricated data, but we also recognize that this does not exclude it as an option. We see the effect of open data on detection of data fabrication as a fruitful avenue for further research.

All in all, we see a need for unbiased effect size estimates to provide meaningful comparisons of genuine- and potentially fabricated data, but even when those are available the (potentially) low positive predictive value of widespread detection of data fabrication is going extremely difficult. Hence, we recommend meta-research to focus on more effective systemic reforms to make progress on the root cause of data fabrication possible. The root cause is likely to be the incentive system that rewards bean-counts of outputs and does not put them in the context of a larger collective scientific effort where validity counts. In other words, rewarding scientists for behaviors that are not primarily in science’s interest. Our premise in these two research studies was after the fact detection of a problem, but we recognize that prior to the fact addressing of the underlying causes that give rise to data fabrication is much more sustainable and effective. Nonetheless, we also recognize that there will always be malintent involved for some researchers, and we recommend that meta-research engage in more penetration testing of how those with malintent can fool a system.

Table 7: Raw results of the variance analyses congruent, sorted on test result [for JMW and MvA, not final manuscript]

	id	type	result
324	0jg.txt	Fabricated	0.00000
369	1zm.txt	Fabricated	0.00000
399	2f5.txt	Fabricated	0.00000
429	3wn.txt	Fabricated	0.00000
459	8nb.txt	Fabricated	0.00000
519	h65.txt	Fabricated	0.00000
534	hsu.txt	Fabricated	0.00000
549	jgg.txt	Fabricated	0.00000
594	o2f.txt	Fabricated	0.00000
609	ojh.txt	Fabricated	0.00000
624	pkl.txt	Fabricated	0.00000
639	sel.txt	Fabricated	0.00000
669	tjv.txt	Fabricated	0.00000
729	z26.txt	Fabricated	0.00000
654	t5g.txt	Fabricated	0.00010
339	19e.txt	Fabricated	0.00025
354	1se.txt	Fabricated	0.00025

	id	type	result
384	2a9.txt	Fabricated	0.00025
414	3pl.txt	Fabricated	0.00025
444	82z.txt	Fabricated	0.00025
474	ez8.txt	Fabricated	0.00025
489	g2f.txt	Fabricated	0.00025
504	h5w.txt	Fabricated	0.00025
564	jmqr.txt	Fabricated	0.00025
579	nbu.txt	Fabricated	0.00025
684	tyo.txt	Fabricated	0.00025
699	ulr.txt	Fabricated	0.00025
714	yty.txt	Fabricated	0.00025
39	cebersole.carleton	Genuine	0.00051
279	cebersole.toronto	Genuine	0.00082
9	cebersole.ashland	Genuine	0.00084
129	cebersole.mississippi	Genuine	0.00093
294	cebersole.vcu	Genuine	0.00097
99	cebersole.miami	Genuine	0.00101
204	cebersole.plu	Genuine	0.00101
219	cebersole.psuabington	Genuine	0.00117
24	cebersole.bradley	Genuine	0.00119
54	cebersole.davis	Genuine	0.00120
114	cebersole.michst	Genuine	0.00127
144	cebersole.montana	Genuine	0.00129
174	cebersole.nova	Genuine	0.00132
69	cebersole.florida	Genuine	0.00138
249	cebersole.sdsu	Genuine	0.00145
189	cebersole.osu	Genuine	0.00163
264	cebersole.texasam	Genuine	0.00181
309	cebersole.virginia	Genuine	0.00196
234	cebersole.riverside	Genuine	0.00252
84	cebersole.ithaca	Genuine	0.00288
159	cebersole.mturk	Genuine	0.00621

Table 8: Raw results of the variance analyses incongruent, sorted on test result [for JMW and MvA, not final manuscript]

	id	type	result
325	0jg.txt	Fabricated	0.00000
370	1zm.txt	Fabricated	0.00000
400	2f5.txt	Fabricated	0.00000
430	3wn.txt	Fabricated	0.00000
460	8nb.txt	Fabricated	0.00000
520	h65.txt	Fabricated	0.00000
550	jgg.txt	Fabricated	0.00000
610	ojh.txt	Fabricated	0.00000
625	pkl.txt	Fabricated	0.00000
670	tjv.txt	Fabricated	0.00000
685	tyo.txt	Fabricated	0.00000
730	z26.txt	Fabricated	0.00000
475	ez8.txt	Fabricated	0.00007
340	19e.txt	Fabricated	0.00008
715	yty.txt	Fabricated	0.00008
580	nbu.txt	Fabricated	0.00009
505	h5w.txt	Fabricated	0.00019
490	g2f.txt	Fabricated	0.00023

	id	type	result
355	1se.txt	Fabricated	0.00025
385	2a9.txt	Fabricated	0.00025
415	3pl.txt	Fabricated	0.00025
445	82z.txt	Fabricated	0.00025
535	hsu.txt	Fabricated	0.00025
565	jmq.txt	Fabricated	0.00025
595	o2f.txt	Fabricated	0.00025
640	sel.txt	Fabricated	0.00025
655	t5g.txt	Fabricated	0.00025
700	ulr.txt	Fabricated	0.00025
40	cebersole.carleton	Genuine	0.00051
280	cebersole.toronto	Genuine	0.00082
10	cebersole.ashland	Genuine	0.00084
130	cebersole.mississippi	Genuine	0.00093
295	cebersole.vcu	Genuine	0.00097
100	cebersole.miami	Genuine	0.00101
205	cebersole.plu	Genuine	0.00101
220	cebersole.psuabington	Genuine	0.00117
25	cebersole.bradley	Genuine	0.00119
55	cebersole.davis	Genuine	0.00120
115	cebersole.michst	Genuine	0.00127
145	cebersole.montana	Genuine	0.00129
175	cebersole.nova	Genuine	0.00132
70	cebersole.florida	Genuine	0.00138
250	cebersole.sdsu	Genuine	0.00145
190	cebersole.osu	Genuine	0.00163
265	cebersole.texasam	Genuine	0.00181
310	cebersole.virginia	Genuine	0.00196
235	cebersole.riverside	Genuine	0.00252
85	cebersole.ithaca	Genuine	0.00288
160	cebersole.mturk	Genuine	0.00621

## 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](http://doi.org/10.1016/0197-2456(91)90037-m)
- Bauer, J., & Gross, J. (2011). Difficulties detecting fraud? The use of benford’s law on regression tables. *Methodological Artefacts, Data Manipulation and Fraud in Economics and Social Science*. <http://doi.org/10.1515/9783110508420-010>
- 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>
- Berger, A., & Hill, T. P. (2011). A basic theory of benford’s law. *Probability Surveys*, 8(0), 1–126. <http://doi.org/10.1214/11-ps175>
- 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. *Social Psychological and Personality Science*, 8(4), 363–369. <http://doi.org/10.1177/1948550616673876>
- 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](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., & Loadman, J. A. (2016). Evidence for non-random sampling in randomised, controlled trials by yuhji saito. *Anaesthesia*, 72(1), 17–27. <http://doi.org/10.1111/anae.13650>
- 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>
- Cheung, I., Campbell, L., LeBel, E. P., Ackerman, R. A., Aykutoğlu, B., Bahník, Š., . . . al. (2016). Registered replication report. *Perspectives on Psychological Science*, 11(5), 750–764. <http://doi.org/10.1177/1745691616664694>
- 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>
- Cyranoski, D. (2015). Collateral damage: How one misconduct case brought a biology institute to its knees. *Nature*, 520(7549), 600–603. <http://doi.org/10.1038/520600a>
- Diekmann, A. (2007). Not the first digit! Using benford’s law to detect fraudulent scientific data. *Journal of Applied Statistics*, 34(3), 321–329. <http://doi.org/10.1080/02664760601004940>
- 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.
- Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., . . . al. (2016). Many labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, 67, 68–82. <http://doi.org/10.1016/j.jesp.2015.10.012>
- 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>
- Fanelli, D., Costas, R., Fang, F. C., Casadevall, A., & Bik, E. M. (2018). Testing hypotheses on risk factors for scientific misconduct via matched-control analysis of papers containing problematic image duplications. *Science and Engineering Ethics*. <http://doi.org/10.1007/s11948-018-0023-7>
- Fewster, R. M. (2009). A simple explanation of benford’s law. *The American Statistician*, 63(1), 26–32. <http://doi.org/10.1198/tast.2009.0005>
- Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburg, United Kingdom: Oliver Boyd.
- Goodman, S. (2008). A dirty dozen: Twelve p-value misconceptions. *Seminars in Hematology*, 45(3), 135–140. <http://doi.org/10.1053/j.seminhematol.2008.04.003>
- 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. (2016). 688,112 statistical results: Content mining psychology articles for statistical test results. *Data*, 1(3), 14. <http://doi.org/10.3390/data1030014>
- Hartgerink, C. H. J., Voelkel, J. V., Wicherts, J. M., & Assen, M. A. van. (2017, July). Transcripts of 28 interviews with researchers who fabricated data for an experiment. <http://doi.org/10.5281/zenodo.832490>
- Hartgerink, C. H. J., Wicherts, J. M., & Van Assen, M. A. L. M. (2017). Too good to be false: Nonsignificant results revisited. *Collabra: Psychology*, 3(1), 9. <http://doi.org/10.1525/collabra.71>
- Hartgerink, C. H., Aert, R. C. van, Nuijten, M. B., Wicherts, J. M., & Assen, M. A. van. (2016). Distributions of p-values smaller than .05 in psychology: What is going on? *PeerJ*, 4, e1935. <http://doi.org/10.7717/peerj.1935>
- Hartgerink, C., Wicherts, J., & Assen, M. van. (2016). The value of statistical tools to detect data fabrication. *Research Ideas and Outcomes*, 2, e8860. <http://doi.org/10.3897/rio.2.e8860>
- Heathers, J. A., Anaya, J., Zee, T. van der, & Brown, N. J. (2018). Recovering data from summary statistics: Sample parameter reconstruction via iterative TEchniques (SPRITE). <http://doi.org/10.7287/peerj.preprints.26968v1>
- 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>
- Hill, T. P., & Schürger, K. (2005). Regularity of digits and significant digits of random variables. *Stochastic Processes and Their Applications*, 115(10), 1723–1743. <http://doi.org/10.1016/j.spa.2005.05.003>
- Hobbes, T. (1651). *Leviathan*. Oxford University Press.
- Hogg, R. V., & Tanis, E. A. (2001). *Probability and statistical inference*. New Jersey, NJ: Prentice-Hall.
- Hüllemann, S., Schüpfer, G., & Mauch, J. (2017). Application of benford’s law: A valuable tool for detecting scientific papers with fabricated data? *Der Anaesthetist*, 66(10), 795–802. <http://doi.org/10.1007/s00101-017-0333-1>
- 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>
- Joint editors-in-chief request for determination regarding papers published by dr. yoshitaka fujii. (2013). *International Journal of Obstetric Anesthesia*, 22(1), e1–e21. <http://doi.org/10.1016/j.ijoa.2012.10.001>
- 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>
- Kranke, P. (2012). Putting the record straight: Granisetron’s efficacy as an antiemetic “post-fujii”. *Anaesthesia*, 67(10), 1063–1067. <http://doi.org/10.1111/j.1365-2044.2012.07318.x>
- Kranke, P., Apfel, C. C., & Roewer, N. (2000). Reported data on granisetron and postoperative nausea and vomiting by fujii et al. are incredibly nice! *Anesthesia & Analgesia*, 90(4), 1004. <http://doi.org/10.1213/00000539-200004000-00053>
- Lakens, D. (2015). Comment: What p-hacking really looks like: A comment on masicampo and lalande (2012). *Quarterly Journal of Experimental Psychology*, 68(4), 829–832. <http://doi.org/10.1080/17470218.2014.982664>
- Levelt. (2012). *Flawed science: The fraudulent research practices of social psychologist Diederik Stapel*. Retrieved from <https://www.commissielevelt.nl/>
- Lieberman, M. D., Berkman, E. T., & Wager, T. D. (2009). Correlations in social neuroscience aren’t voodoo: Commentary on vul et al. (2009). *Perspectives on Psychological Science*, 4(3), 299–307.

<http://doi.org/10.1111/j.1745-6924.2009.01128.x>

Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, 45(6), 633–644. <http://doi.org/10.1509/jmkr.45.6.633>

Mosimann, J. E., & Ratnaparkhi, M. V. (1996). Uniform occurrence of digits for folded and mixture distributions on finite intervals. *Communications in Statistics - Simulation and Computation*, 25(2), 481–506. <http://doi.org/10.1080/03610919608813325>

Mosimann, J. E., Wiseman, C. V., & Edelman, R. E. (1995). Data fabrication: Can people generate random digits? *Accountability in Research*, 4(1), 31–55. <http://doi.org/10.1080/08989629508573866>

Naomi Ellemers. (2017). Ethisch klimaat op het werk: Op zoek naar het nieuwe normaal [Ethical climate at work: Searching for the new normal]. Retrieved from [https://wayback.archive.org/web/20180726070256/https://www.scoop-program.org/images/Tekst\\_Oratie\\_Naomi\\_Ellemers\\_9\\_februari\\_2017.pdf](https://wayback.archive.org/web/20180726070256/https://www.scoop-program.org/images/Tekst_Oratie_Naomi_Ellemers_9_februari_2017.pdf)

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>

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, 5(2), 241–301. <http://doi.org/10.1037/1082-989x.5.2.241>

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>

Nuijten, M. B., Assen, M. A. L. M. van, Veldkamp, C. L. S., & Wicherts, J. M. (2015). The replication paradox: Combining studies can decrease accuracy of effect size estimates. *Review of General Psychology*, 19(2), 172–182. <http://doi.org/10.1037/gpr0000034>

Nuijten, M. B., Hartgerink, C. H. J., Assen, M. A. L. M. van, Epskamp, S., & Wicherts, J. M. (2015). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, 48(4), 1205–1226. <http://doi.org/10.3758/s13428-015-0664-2>

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>

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). PROC: An open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, 12(1), 77. <http://doi.org/10.1186/1471-2105-12-77>

Sijtsma, K., Veldkamp, C. L. S., & Wicherts, J. M. (2015). Improving the conduct and reporting of statistical analysis in psychology. *Psychometrika*, 81(1), 33–38. <http://doi.org/10.1007/s11336-015-9444-2>

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

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18(6), 643–662. <http://doi.org/10.1037/h0054651>

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>

Ulrich, R., & Miller, J. (2015). P-hacking by post hoc selection with multiple opportunities: Detectability by skewness test?: Comment on simonsohn, nelson, and simmons (2014). *Journal of Experimental*

*Psychology: General*, 144(6), 1137–1145. <http://doi.org/10.1037/xge0000086>

Youngstrom, E. A. (2013). A primer on receiver operating characteristic analysis and diagnostic efficiency statistics for pediatric psychology: We are ready to roc. *Journal of Pediatric Psychology*, 39(2), 204–221. <http://doi.org/10.1093/jpepsy/jst062>

Yule, G. U. (1922). An introduction to the theory of statistics. Retrieved from <https://ia800205.us.archive.org/13/items/cu31924013993187/cu31924013993187.pdf>

(2017). *Nature*, 546(7660), 575–575. <http://doi.org/10.1038/546575a>