Detection of data fabrication using statistical tools

Hartgerink, Chris; Voelkel, Jan G.; Wicherts, Jelte M.; van Assen, Marcel

DOI:
10.31234/osf.io/jkws4

Publication date:
2019

Document Version
Early version, also known as pre-print

Detection of data fabrication using statistical tools

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

19 August, 2019

Contents

Abstract

Scientific misconduct potentially invalidates findings in many scientific fields. Improved detection of unethical practices like data fabrication is considered to deter such practices. In two studies, we investigated the diagnostic performance of various statistical methods to detect fabricated quantitative data from psychological research. In Study 1, we tested the validity of statistical methods to detect fabricated data at the study level using summary statistics. Using (arguably) genuine data from the Many Labs 1 project on the anchoring effect ($k = 36$) and fabricated data for the same effect by our participants ($k = 39$), we tested the validity of our newly proposed 'reversed Fisher method', variance analyses, and extreme effect sizes, and a combination of these three indicators using the original Fisher method. Results indicate that the variance analyses perform fairly well when the homogeneity of population variances is accounted for and that extreme effect sizes perform similarly well in distinguishing genuine from fabricated data. The performance of the 'reversed Fisher method' was poor and depended on the types of tests included. In Study 2, we tested the validity of statistical methods to detect fabricated data using raw data. Using (arguably) genuine data from the Many Labs 3 project on the classic Stroop task ($k = 21$) and fabricated data for the same effect by our participants ($k = 28$), we investigated the performance of digit analyses, variance analyses, multivariate associations, and extreme effect sizes, and a combination of these four methods using the original Fisher method. Results indicate that variance analyses, extreme effect sizes, and multivariate associations perform fairly well to excellent in detecting fabricated data using raw data, while digit analyses perform at chance levels. The two studies provide mixed results on how the use of random number generators affects the detection of data fabrication. Ultimately, we consider the variance analyses, effect sizes,

and multivariate associations valuable tools to detect potential data anomalies in empirical (summary or raw) data. However, we argue against widespread (possible automatic) application of these tools, because some fabricated data may be irregular in one aspect but not in another. Considering how violations of the assumptions of fabrication detection methods may yield high false positive or false negative probabilities, we recommend comparing potentially fabricated data to genuine data on the same topic.

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 having falsified or fabricated research results at least once during their career (Fanelli, 2009), which due to its self-report nature is likely to be an underestimate of the true rate of misconduct. The detection rate of data fabrication is likely to be even lower; for example, among several hundreds of thousands of researchers working in the United States and the Netherlands, only around a dozen cases become public each year. At best, this suggests a detection rate below 1% among those 2% who admit to fabricating or falsifying data — the tip of a seemingly much larger iceberg.

The ability to detect fabricated data may help deter researchers from fabricating data in their work. Deterrence theory (e.g., Hobbes, 1651) states that improved detection of undesirable behaviors decreases the expected utility of said behaviors, ultimately leading to fewer people to 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 in practice, not only at journals but also in the evaluation of student theses (e.g., with commercial services such as Turnitin). Various tools have been developed to detect image manipulation and some of these tools have been implemented at biomedical journals to screen for fabricated or falsified images. For example, the Journal of Cell Biology and the EMBO journal scan each submitted image for potential image manipulation (The Journal of Cell Biology, 2015; 2017), which supposedly increases the risk of detecting (blatant) image manipulation. Recently developed algorithms even allow automated 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 in the literature (4% of all papers are estimated to contain manipulated images; Bik, Casadevall, & Fang, 2016) and to study factors that predict image manipulation (Fanelli, Costas, Fang, Casadevall, & Bik, 2018).

Methods to detect fabrication of quantitative data 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; Wagenaar, 1972), they might create fabricated data that fail to follow the fundamentally

probabilistic nature of genuine data. Data and outcomes of analyses based on these data that fail to align with the (at least partly probabilistic) processes that are assumed to underlie genuine data may indicate deviations from the reported data collecting protocol, potentially even data fabrication or falsification.

Statistical methods have proven to be of importance in initiating data fabrication investigations or in assessing the scope of potential data fabrication. For example, Kranke, Apfel, and Roewer skeptically perceived Fujii'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 Fujii's papers for misconduct (Carlisle, 2012; Carlisle & Loadsman, 2016; Carlisle, Dexter, Pandit, Shafer, & Yentis, 2015) did events cumulate into an investigation- and ultimately retraction of 183 of Fujii'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 ultimately 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 and valuable, we need to study their diagnostic value. Specifically, many of the developed statistical methods to detect data fabrication are quantifications of case specific suspicions by researchers, but these applications do not inform us on their diagnostic value (i.e., sensitivity and specificity) outside of those specific cases. Side-by-side comparisons of different statistical methods to detect data fabrication 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 is not included. Using different statistical methods to detect fabricated data using genuine versus fabricated data yields information on the sensitivity and specificity of the detection tools. This is important because of the severe professional- and personal consequences of accusations of potential research misconduct (as illustrated by the STAP case; Cyranoski, 2015). These methods might have utility in misconduct investigations where the prior chances of misconduct are high, but their diagnostic value in large-scale applications to screen the literature are unclear.

In this article, we investigate the diagnostic performance of various statistical methods to detect data fabrication. These statistical methods (detailed next) have not previously been validated systematically in research using both genuine and fabricated data. We present two studies where we try to distinguish (arguably) genuine data from known fabricated data based on these statistical methods. These studies investigate methods to detect data fabrication in summary statistics (Study 1) or in individual level (raw) data (Study 2) in psychology. 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 individual level data for a classic 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 discuss the theoretical framework of the investigated statistical methods to detect data fabrication.

Theoretical framework

Statistical methods to detect potential data fabrication can be based either on reported summary statistics that can often be retrieved from articles or on the raw (underlying) data if these are available. Below we detail p-value analysis, variance analysis, and effect size analysis as potential ways to detect data fabrication using summary statistics. 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 calculated (e.g., an APA reported t- or F-statistic; C. Hartgerink, Wicherts, & Van Assen, 2017). Among the methods that can be applied to uncover potential fabrication using raw data, we consider digit analyses (i.e., the Newcomb-Benford law and terminal digit analysis) and multivariate associations between variables. 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 also 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 (arguably) 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, while it is right-skewed if the alternative hypothesis is true (Fisher, 1925). If the model assumptions of the underlying process hold, the probability density function 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 in an effort to obtain significance (Simmons, Nelson, & Simonsohn, 2011; Wicherts et al., 2016), a right-skewed distribution for significant p-values can and will likely still occur (C. H. Hartgerink et al., 2016; Ulrich & Miller, 2015).

A failure of independent p-values to be right-skewed or uniformly distributed (as would be theoretically expected) can indicate potential data fabrication.

For example, in the Fujii case, baseline measurements of supposed randomly assigned groups later turned out to be fabricated. When participants are randomly assigned to conditions, measures at baseline are expected to statistically equivalent between the groups (i.e., equivalent distributions), hence, produce uniformly distributed p-values. However, in the Fujii case, Carlisle observed many large p-values, which ultimately led to the identification of potential data fabrication (Carlisle, 2012). The cause of such large p-values may be that the effect of randomness is underappreciated when fabricating statistically nonsignificant data due to (for example) widespread misunderstanding of what a p-value means (Goodman, 2008; Sijtsma, Veldkamp, & Wicherts, 2015), which results in groups of data that are too similar conditional on the null hypothesis of no differences between the groups. As an illustration, we simulated normal distributed measurements for studies and their t-test comparisons in Table 1, under statistically equivalent populations (Set 1). We also fabricated independent data for equivalent groups, where we fixed the mean and standard deviation for all studies and subsequently added (too) little uniform noise to these parameters (Set 2). 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 is randomly generated data under the null hypothesis of random assignment (assumed to be the genuine process), whereas Set 2 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 of a study.

	Set 1			Set 2		
	Experimental	Control		Experimental	Control	
	M (SD)	M (SD)	P-value	M (SD)	M (SD)	P-value
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 propose 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). The original Fisher method is computed over the individual p-values (p_i) as

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

where the null hypothesis of a zero true effect size underlying all k results is tested and is rejected for values of the test statistic that are larger than a certain value, typically the 95th percentile of χ^2_{2k}, to conclude that true effect size differs from zero for at least one of k results. The Fisher method can be adapted to test the same null hypothesis against the alternative that the results are closer to their expected values than expected under the null. The adapted test statistic of this so-called 'reversed Fisher method' is

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

where t determines the range of p-values that are selected in the method. For instance, if $t = 0$, all p-values are selected, whereas if $t = .05$ only statistically nonsignificant results are selected in the method. Note that each result's contribution (between the brackets) is in the interval (0,1), as for the original Fisher method. The reversed Fisher method is similar (but not equivalent) to Carlisle's method testing for excessive homogeneity across baseline measurements in RCTs (Carlisle, 2012, 2017a; 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 $\chi^2_{2\times10} = 18.362, p = 0.564$ for the genuine data (Set 1), and $\chi^2_{2\times10} = 66.848, p = 6 \times 10^{-7}$ for the fabricated data (Set 2). Another example, from the Fujii case (Carlisle, 2012), also illustrates that the reversed Fisher method may detect fabricated data; the p-values related to fentanyl dose (as presented in Table 3 of Carlisle, 2012) for five independent comparisons also show excessively high p-values, $\chi^2_{2\times5} = 19.335, p = 0.036$. However, based on this anecdotal evidence little can be said about the sensitivity, specificity, and utility of the reversed Fisher method.

We note that incorrectly specified 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 were 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 paper, misspecification of one-tailed hypothesis testing is not an issue because we prespecified the effect and its direction to the participants who were requested to fabricate data.

Variance analysis

In most empirical research papers, sample variance or standard deviation estimates are typically reported alongside means to indicate dispersion in the data. For example, if a sample has a reported age of $M(SD) = 21.05(2.11)$ we know this sample is both younger and more homogeneous than another sample with reported $M(SD) = 42.78(17.83)$.

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 in sample j is χ^2-distributed (p. 445; Hogg & Tanis, 2001); that is

$$var(x) \sim \frac{\chi^2_{n_j-1}}{n_j - 1} \tag{3}$$

where n is the sample size of the jth group. Assuming equal variances of the J populations, this population variance is estimated by the Mean Squares within (MS_w) as

$$MS_w = \frac{\sum_{j=1}^{k}(n_j - 1)s_j^2}{\sum_{j=1}^{k}(n_j - 1)} \tag{4}$$

where s_j^2 is the sample variance and n_j the sample size in group j. As such, under normality and equality of variances, the sampling distribution of standardized[1] variances in group j (i.e., z_j^2) is

$$z_j^2 \sim \left(\frac{\chi^2_{n_j-1}}{n_j - 1}\right) / MS_w \tag{5}$$

Using the theoretical sampling distribution of the standardized variances, we bootstrap the expected distribution of the dispersion of variances. In other words, we use the theoretical sampling distribution of the standard deviations to formulate a null model of the dispersion of variances that is in line with the probabilistic sampling processes for groups of equal population variances. First, we randomly draw standard deviations for all j groups according to Equation 3. Second, we calculate MS_w using those previously drawn values (Equation 4). Third, we standardize the standard deviations using Equation 5. Fourth, we compute the measure of dispersion across the j groups as the standard deviation of the standardized variances (denoted SD_z, Simonsohn, 2013) or as the range of the standardized variances (denoted $max_z - min_z$). This process is repeated for i iterations to generate a parametric bootstrap distribution of the dispersion of variances according to the null model of equal variances across populations.

The observed dispersion of the variances, when compared to its expected distribution, allows a test for potential data fabrication. To this end we compute the proportion of iterations that show equally- or more extreme consistency in the dispersion of the variances to compute a bootstrapped p-value (e.g., $P(X \leq SD_{obs})$),

[1] By dividing all variances by MS_w their weighted average equals 1. This is what we call standardization for this scenario.

with SD_{obs} the standard deviation of standardized variances and X the random variable corresponding to the standard deviation of standardized variances under the null model. In other words, we compute how many samples of j groups under the null show the observed consistency of the dispersion in the variances (or more consistent), to test whether the data are plausible given a genuine probabilistic sampling process (Simonsohn, 2013). Similar to the Fisher method, this could be the result of the fabricator underappreciating the higher level sampling fluctuations, resulting in generating too little randomness (i.e., error) in the standard deviations across groups (Mosimann et al., 1995).

As an example, we apply the variance analysis to the illustration from Table 1 and the Smeesters case (Simonsohn, 2013). We apply the variance analysis across the standard deviations from each set in Table 1. For the genuinely probabilistic data (Set 1), we find that the reported mean standard deviation is 9.868 with a standard deviation equal to 0.595. For the fabricated data (Set 2), we find that the reported mean standard deviation is 10.667 with a standard deviation equal to 0.456. Such means illustrate the differences, but are insufficient to test them. Using the standard deviation of variances as the dispersion of variances measure, we can quantify how extreme this difference is using the previously outlined procedure. Results indicate that 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 randomly selected samples under the null model of independent groups with equal variances on a normally distributed measure, 2.142×10^4 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 a study in the Smeesters case ($n_j = 15$) were reported to have standard deviations 25.09, 24.58, and 25.65 (Simonsohn, 2013). Here too, we can use the outlined procedure to see whether these reported standard deviations are too consistent according to sampling fluctuations of the second moment of the data according to theory. The standard deviation of these standard deviations is 0.54. Comparing this to 100,000 randomly selected replications under the theoretical null model, such consistency in standard deviations (or even more) would only be observed in 1.21% of those (Simonsohn, 2013).

Extreme effect sizes

There is sufficient evidence that data fabrication can result in (too) large effects. For example, in the misconduct investigations in the Stapel case, large effect sizes were used as an indicator of data fabrication (Levelt, 2012) with some papers showing incredibly large effect sizes that translate to explained variances of up to 95% or these effect sizes were larger than the product of the reliabilities of the related measures. Moreover, Akhtar-Danesh & Dehghan-Kooshkghazi (2003) asked faculty members from three universities to fabricate data sets and found that the fabricated data generally 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, effect sizes in psychology papers are often reported as a standardized mean difference (e.g., d) or as an explained variance (e.g., R^2). A test statistic can be transformed into a measure of effect size. A test result such as $t(59) = 3.55$ in a between-subjects design corresponds to $d = 0.924$ and $r = 0.176$ (C. 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 (Cialdini & Goldstein, 2004) 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 distribution is $r = .2$ and a reasonable number of studies on the 'foot-in-the-door' effect have been conducted, an extremely large effect might be considered 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. Here we focus on testing the distribution of leading digits based on the Newcomb-Benford Law (NBL) and testing the distribution of terminal digits based on the uniform distribution 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 occuring under certain conditions, but rather 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:

$$P(d) = log_{10} \frac{1 + d}{d} \tag{6}$$

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 using a χ^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 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 that humans might be able to fabricate data that are in line with the NBL (Burns, 2009; Diekmann, 2007), immediately undermining the applicability of the NBL in context of detecting data fabrication.

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 also conducted using a χ^2-test ($df = 10 - 1$) on the digit occurrence counts (including zero), where the observed frequencies are compared with the expected uniform frequencies. 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, standard 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 only slightly non-uniform. 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, depends on the process generating the data. In our example with $N \sim (0, 1)$, the distribution of the third and later digits seem well-approximated by the uniform distribution.

Multivariate associations

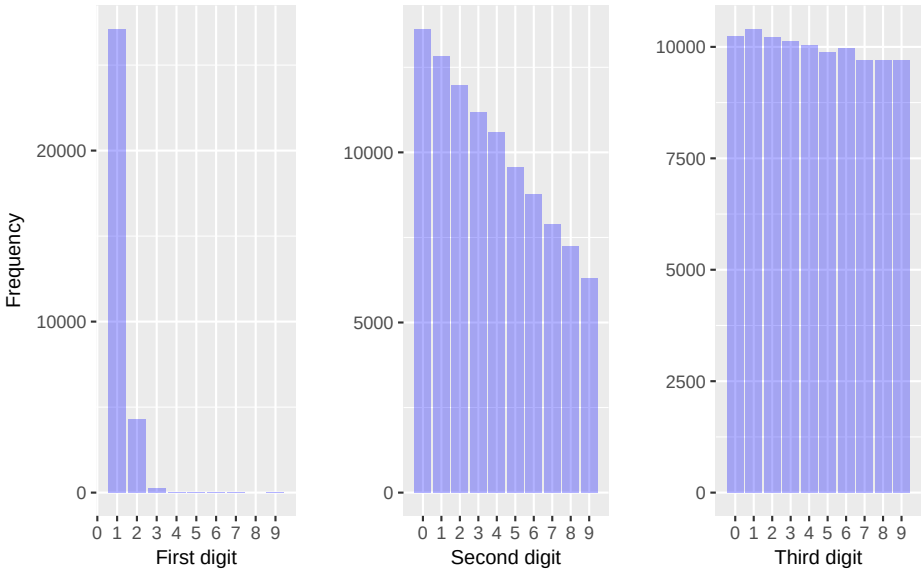

Figure 1: Frequency distributions of the first-, second-, and third digits. We sampled 100,000 values from a standard normal distribution to create these digit distributions.

Variables or measurements included in one study can have multivariate associations that might be non-obvious to researchers. Hence, such relations between variables or measurements might be overlooked by people who fabricate data. Fabricators might also simply be practically unable to fabricate data that reflect these multivariate associations, even if they are aware of these associations. For example, in response time latencies, there typically is a positive relation between mean response time and the variance of the response time. 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 for someone who 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 (arguably) 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. With these control data, a meta-analysis provides an overall estimate of the multivariate relation that can subsequently be used to verify the credibility of a set of statistics.

Specifically, the multivariate associations from the genuine data are subsequently used to estimate the extremity of an observed multivariate relation in investigated data. 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 (e.g., a correlation) between Ms and SDs from the literature ($N \sim (.123, .1)$). Assume

we have two papers, each coming from a pool of direct replications providing an equal number of Ms and corresponding SDs. Associations between these statistics are 0.5 for Paper 1 and 0.2 for Paper 2. From Figure 2 we see that the association in Paper 1 has a much higher percentile score in the distribution (i.e., 99.995th percentile) than that of Paper 2 (i.e., 78.447th percentile).

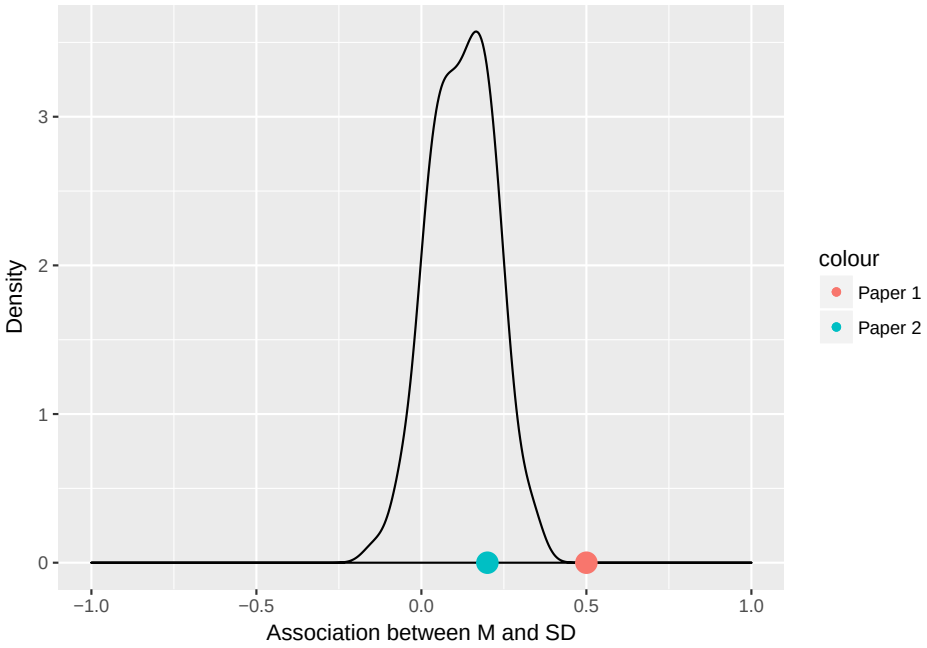

Figure 2: Distribution of 100 simulated observed associations between Ms and SDs for a response latency task; simulated under $N(.123, .1)$. The red- and blue dots indicate observed multivariate associations from fictitious papers. Paper 1 may be considered relatively extreme and of interest for further inspection; Paper 2 may be considered relatively normal.

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 with psychological data. We asked participants to fabricate data that were supposedly drawn from a study on the anchoring effect (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?

In their classic study, A. Tversky & Kahneman (1974) varied the anchor in this question between 10% and 65% and found that they yielded mean responses of 25% and 45%, respectively (A. Tversky & Kahneman, 1974). We chose the anchoring effect because it is well known and because a considerable amount of (arguably) genuine data sets on the anchoring heuristic are freely available (https://osf.io/pqf9r; Klein et al., 2014). This allowed us to compare data knowingly and openly fabricated by our participants (researchers in psychology) to actual data that can be assumed to be genuine because they were draw from a large-scale international project involving many contributing labs (a so-called Many Labs study). Our data fabrication study was approved by Tilburg University's Ethics Review Board (EC-2015.50; https://osf.io/7tg8g/).

Methods

We collected genuine summary statistics from the Many Labs study and fabricated summary statistics from our participating fabricators 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 (genuine or fabricated) studies provided us with summary statistics in a 2 (low/high anchoring) $\times$ 2 (male/female) factorial design. Our analysis of the data fabrication detection methods used the summary statistics (i.e., means, standard deviations, and test results) of the four anchoring studies fabricated by each participant or the four anchoring studies that had actually been conducted by each participating lab in the Many Labs project (Klein et al., 2014). 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)") and explain the reason of the deviation.

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 labs 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.[2] 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 our study (April 25, 2016; 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 (participants were not informed about how we planned to detect data fabrication; the procedure for this is explained in the Data Analysis section). We did not inform participants about how we planned to detect data fabrication. The provided e-mail addresses were unlinked from individual responses upon sending the bonus gift cards. The full Qualtrics survey is available at osf.io/rg3qc.

Each participant was instructed to fabricate 32 summary statistics (4 studies $\times$ 2 anchoring conditions $\times$ 2 sexes $\times$ 2 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 in the fabricated anchoring studies to 25 per cell so that 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 included means and standard deviations for the four conditions. Using these values, the spreadsheet automatically computed statistical tests and immediately showed them in the "Current result" column instantaneously. If these results supported the (fabrica-

[2]We discovered that we included several non-U.S. researchers 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.

tion) 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 none of them did.

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, shown in the column "Current result". 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).

Upon completion of the data fabrication, we debriefed respondents within Qualtrics (full survey: osf.io/rg3qc/). Respondents self-rated their statistical knowledge (1 = extremely poor, 10 = excellent), what statistical analysis programs they used frequently (i.e., at least once per week), whether they had ever conducted an anchoring study themselves, whether they used a random number generator to fabricate data in this study, whether they fabricated raw data to get summary statistics, how many combinations of means and standard deviations they created for each study (on average), and a free-text description of their fabrication procedures per study. Lastly we reminded participants 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). Using quotum 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 (36 and 39, respectively), with each data set consisting of summary statistics of four anchoring studies. The data set is the unit of analysis. Four types of analyses are conducted on each of the 75 data sets; (i) the reversed Fisher method, (ii) variance analyses, (iii) the Fisher method applied to the results of the former two, and (iv) analysis of the effect sizes of the statistically significant anchoring effect of the four anchoring studies. Per type of analysis, we examine if we can distinguish the 36 genuine from the 39 fabricated data sets, mainly using Area Under Receiving Operator

Characteristic (AUROC) curves. Below we first describe each of the four types of analyses, followed by a description of the AUROC curve analysis.

We conducted two analyses to detect data fabrication using the reversed Fisher method. More specifically, we conducted one reversed Fisher method analysis for the four statistically nonsignificant results of the gender effect (one per anchoring study) and one for the four statistically nonsignificant interaction effects (one per anchoring study). This results in two reversed Fisher method results (based on k=4) per data set.

For the variance analyses, we substantially deviated from the preregistration (https://osf.io/tshx8/) and added multiple analyses. We analyzed the 16 sample variances (four anchoring studies $\times$ four conditions per anchoring study) per lab or participant in fourteen different ways. Each of the fourteen variance analyses was conducted using two dispersion of variance measures. One measure inspects the standard deviation of the sample variances (i.e., SD_z); one measure inspects the range of the sample variances (i.e., $max_z - min_z$); we ran all 28 analyses with 100,000 iterations from which we computed the bootstrapped p-value (see also the Theoretical Framework). Of these 28 variance analyses (14 for each dispersion of variances measure), only one was preregistered. This was the variance analysis combining all 16 sample variances of the four anchoring studies. Upon analyzing the results of this preregistered variance analysis, however, we realized that the variance analyses assume that the included variances are from the same population distribution. Assuming homogeneous populations of variances is unrealistic for the four very different anchoring conditions or studies (i.e., they have outcome measures on very different scales, such as distances in miles and babies born). Hence, we included variance analyses based on subgroups, where we analyzed each anchoring study separately (four variance analyses) or analyzed each anchoring condition of each study separately (i.e., the low/high anchoring condition collapsed across gender; eight variance analyses). We also conducted one variance analysis that combined all variances across studies but takes into account the subgroups per anchoring condition per study.

We also combined the reversed Fisher method results with the results from the variance analyses using the original Fisher method. More specifically, we combined the results from the two reversed Fisher method analyses (one analysis for the four gender effects and one analysis for the four interaction effects) with the preregistered variance analysis (the result of this analysis was used to determine the three most difficult to detect fabricated datasets and subsequently to reward the 'best fabricators'). We additionally applied the Fisher method to results of the reversed Fisher method (two results) with three different combinations of results of the variance analyses; based on variance analyses per anchoring study (four results), per anchoring study $\times$ condition combination (eight results), and across all studies and conditions but taking into account heterogeneous variances per anchoring condition for each study (one result). Hence, the additional Fisher method analyses were based on six, ten, and three results, respectively. Throughout these combinations, we only use the SD_z dispersion of variance measure for parsimony. Note that the performance of the Fisher method combining results of various analyses (the reversed Fisher method and the variance analyses) as we do here is naturally dependent on the performance of the individual results included in the combination; if all included

results perform well the Fisher method is bound to perform well and vice versa.

Finally, we looked at statistically significant effect sizes. We expected fabricated statistically significant effects to be larger than genuine statistically significant effects. As such, we compared the 75 statistically significant anchoring effects for each of the four anchoring studies separately (not preregistered).

For each of the previously described statistical methods to detect data fabrication, we carried out AUROC curve analyses. 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., $\alpha = 0, .01, .02, ..., .99, 1$). 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) is equal to 50% (assuming equal prevalences). 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 ($k = 39$) datasets, which contain four p-values and relevant effect sizes (r) for each type of effect (gender, anchoring, interaction) per dataset (i.e., 75×4 data points for each plot). These group-level comparisons provide a general overview of the differences between the genuine and fabricated data. Figure 4 (right and left column) already indicates that there are few systematic 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 differences when we required participants to fabricate statistically significant summary statistics (i.e., anchoring hypothesis; middle column). We discuss results bearing on the specific tests for data fabrication next.

P-value analysis

When we applied the reversed Fisher method to the statistically nonsignificant effects, results indicated its performance is approximately equal to chance classification. We found $AUROC = 0.501$, 95% CI [0.468-0.535] for statistically nonsignificant gender effects and $AUROC = 0.516$, 95% CI [0.483-0.549] for statistically nonsignificant interaction effects. For the gender effects, we classified 12 of the 39 fabricated summary statistics as fabricated ($\alpha = .01$) and 6 of the 36 genuine summary statistics as fabricated (results per respondent available at osf.io/a6jb4). For the interaction effects, we classified 11 of the 39 fabricated summary statistics ($\alpha = .01$) and 8 of the 36 genuine summary statistics as fabricated (results per respondent available at osf.io/jz57p). In other words, results from this sample indicated that detection of fabricated data using the

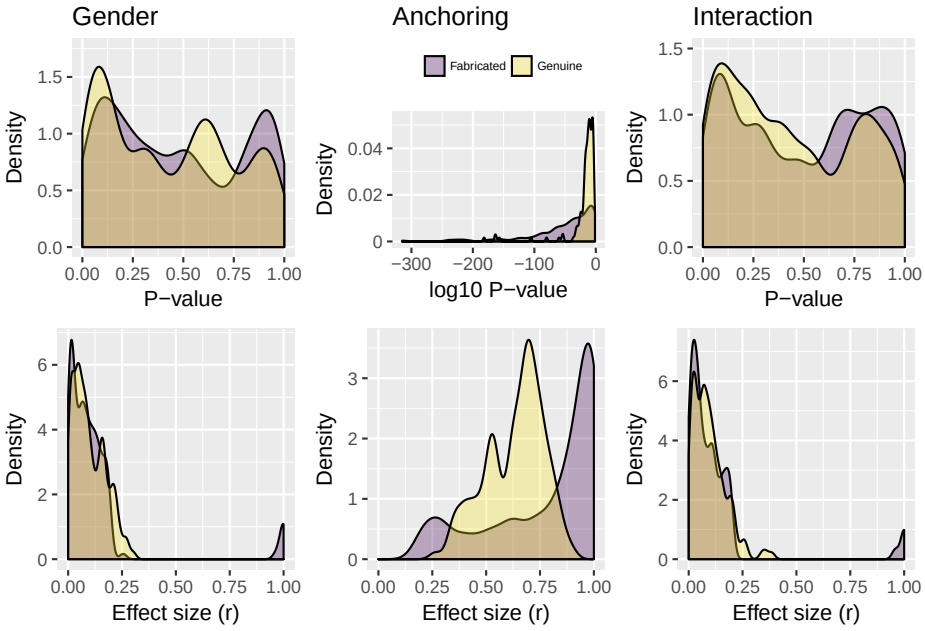

Figure 4: Density distributions of genuine and fabricated summary statistics across four anchoring studies, per effect (gender, anchoring, or interaction across columns) and type of result (p-value or effect size across rows).

distribution of statistically nonsignificant p-values to detect excessive amounts of high p-values does not seem promising.

Variance analysis

We expected the dispersion of variances to be lower in fabricated data as opposed to genuine data. We computed the AUROC values for the variance analyses with the directional hypothesis that genuine data shows more variation than fabricated data, using either the dispersion of variance as captured by the standard deviation of the variances (i.e., SD_z) or the range of the variances (i.e., $max_z - min_z$). AUROC results of all 14 analyses (as described in the Data analysis section) are presented in Table 3, one result for each dispersion of variance measure. Of these 14 results, we only preregistered the variance analysis inspecting the standardized variances across all studies under both the SD_z and $max_z - min_z$ operationalizations, assuming unrealistically homogeneous population variances (https://osf.io/tshx8/; second row of Table 3). As we did not preregister the other variance analyses, these should be considered exploratory.

Under the (in hindsight unrealistic) assumption of homogeneous population variances, our preregistered variance analyses did not perform above chance level. Using the standard deviation of the variances (i.e., SD_z) as dispersion of variance measure, the results are: $AUROC = 0.264$, 95% CI [0.235-0.293]. With this statistic, we classified 0 of the 39 fabricated summary statistics ($\alpha = .01$) and 0 of the 36 genuine summary statistics as fabricated (results per respondent

Table 3: Area Under Receiving Operator Characteristic (AUROC) values of each variance analysis and operationalization, including its 95 percent Confidence Interval. 'Heterogeneity' assumes unequal population variances for the low- and high anchoring conditions, whereas 'homogeneity' assumes equal population variances across anchoring conditions in the same study. We preregistered only the analyses in the second row.

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]

available at osf.io/9cjdh). Using the range of the variances (i.e., $max_z - min_z$) as dispersion of variance, the results are: $AUROC = 0.544$, 95% CI [0.507-0.58]. With this statistic, we detected 39 of the 39 fabricated summary statistics as fabricated ($\alpha = .01$) and 36 of the 36 genuine summary statistics as fabricated (results per respondent available at osf.io/2ts6b). Comparing the results between SD_z and $max_z - min_z$ indicates that the range of the variances measure seems more robust to the violations of the assumption of homogeneous variances than the standard deviation of the variances measure. Overall these results indicate that a violation of the homogeneity assumption may undermine analyses on heterogeneous variances. These assumptions should be made more explicit and checked whenever possible, to prevent improper use.

We conducted exploratory analyses that take into account the heterogeneity of variances across conditions and studies, which sometimes also resulted in improved performance to detect data fabrication. Analyses separated per study or anchoring condition show variable $AUROC$ results (ranging from 0.373-0.798; rows 3-14 in Table 3). Using the standard deviation of variances (i.e., SD_z; row 1 in Table 3) in a heterogeneous manner across the conditions and studies, $AUROC = 0.761$, 95% CI [0.733-0.788]. With this statistic, we classified 9 of the 39 fabricated summary statistics as fabricated ($\alpha = .01$) and 0 of the 36 genuine summary statistics (results per respondent available at osf.io/srpg9). Using the range of variances (i.e., $max_z - min_z$) in a heterogeneous manner across the conditions and studies, $AUROC = 0.827$, 95% CI [0.8-0.853]. With this statistic, we classified the same 9 of the 39 fabricated summary statistics as fabricated ($\alpha = .01$) and 0 of the 36 genuine summary statistics (results per respondent available at osf.io/93rek).

Combining p-value and variance analyses

Our preregistered analysis combined the homogeneous variance analysis across studies and conditions with the p-value analyses of the gender and interaction effects. This combined analysis yielded $AUROC = 0.58$, 95% CI [0.548-0.611]. With this statistic, we classified 19 of the 39 fabricated summary statistics as fabricated ($\alpha = .01$) and 16 of the 36 genuine summary statistics (results per respondent available at osf.io/hq29t). Given that the combination method would be expected to perform not much better than its constituent results it logically follows that the combination of p-values and variance analyses performs this poorly.

The poor performance is in part is due to the unrealistic assumption of homogeneous variances in the variance analysis; we explored the efficacy of other combinations that loosen this assumption. First, we split the variance analyses per study and included four variance analysis results instead of one when we analyzed them overall; $AUROC = 0.605$, 95% CI [0.573-0.636]. With this statistic, we classified 20 of the 39 fabricated summary statistics as fabricated ($\alpha = .01$) and 13 of the 36 genuine summary statistics (results per respondent available at osf.io/r8pf5). Second, we split the variance analyses further, splitting across conditions within studies. This adds another four variance analyses (a total of eight); $AUROC = 0.684$, 95% CI [0.655-0.714]. With this statistic, we classified 25 of the 39 fabricated summary statistics as fabricated ($\alpha = .01$) and 15 of the 36 genuine summary statistics (results per respondent available at osf.io/sv35k). Finally, we replaced the original homogeneous variance analysis (row 2 in Table 3) with the overall and heterogeneous variance analysis (row 1 in Table 3); $AUROC = 0.647$, 95% CI [0.616-0.677]. With this statistic, we classified 23 of the 39 fabricated summary statistics as fabricated ($\alpha = .01$) and 16 of the 36 genuine summary statistics (results per respondent available at osf.io/zt3nk). As the $AUROC$s of the combination method did not exceed that of the variance analyses alone, we conclude that the combination method failed to outperform the variance analyses.

Extreme effect sizes

Using the statistically significant effect sizes from the anchoring studies, we differentiated between the fabricated and genuine results fairly well. Figure 4 (middle column, second row) indicates that the fabricated statistically significant effects were considerably different from the genuine ones. When we inspected the effect size distributions (r), we saw that the median fabricated effect size across the four studies was 0.891 whereas the median genuine effect size was considerably smaller (0.661; median difference across the four anchoring effects 0.23). In contrast to the fabricated nonsignificant effects, which resembled the genuine data quite well, the statistically significant effects seem to have been harder to fabricate for the participants. More specifically, 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). Figure 5 depicts the density distributions of the genuine and fabricated effect sizes per anchoring study, which shows the extent to which the density of the fabricated effect sizes exceeds the maximum of the genuine effect sizes. For instance, the percentage of fabricated statistically significant anchoring effect sizes that is larger than all 36 genuine statistically significant anchoring effect

sizes is 59% in Study 1, 64.1% in Study 2, 53.8% in Study 3, and 66.7% in Study 4. Based on these results, it seems that using extreme effect sizes to detect potential data fabrication may be is a parsimonious and fairly effective method.

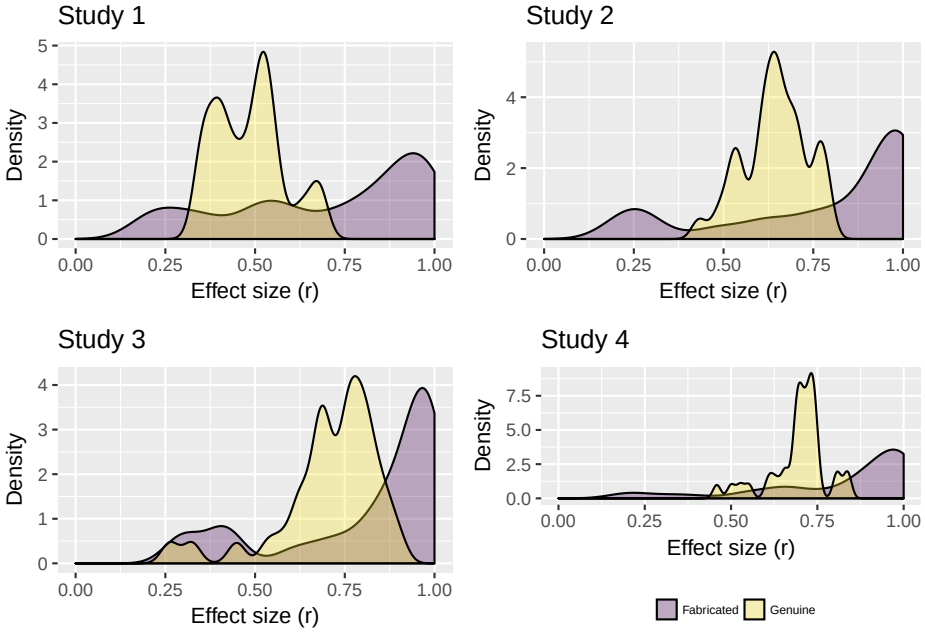

Figure 5: Density distributions of genuine and fabricated anchoring effect sizes for each of the four anchoring studies.

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 are 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, leading to more difficult to detect fabricated data. The analyses presented next were not preregistered.

We split our analyses for those 11 participants who indicated using RNGs and the remaining 28 participants who indicated not to have used RNGs. Figure 6 shows the same density distributions as in Figure 4, except that this time the density distributions of the fabricated data are split between these two groups.

Figure 6 suggests that using RNGs may have resulted in less exaggerated anchoring effect sizes, but still larger than genuine ones. Furthermore, it seems that the use of RNGs produced somewhat more uniformly distributed statistically nonsignficant p-values than those without RNGs. For effect sizes, Table 4 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).

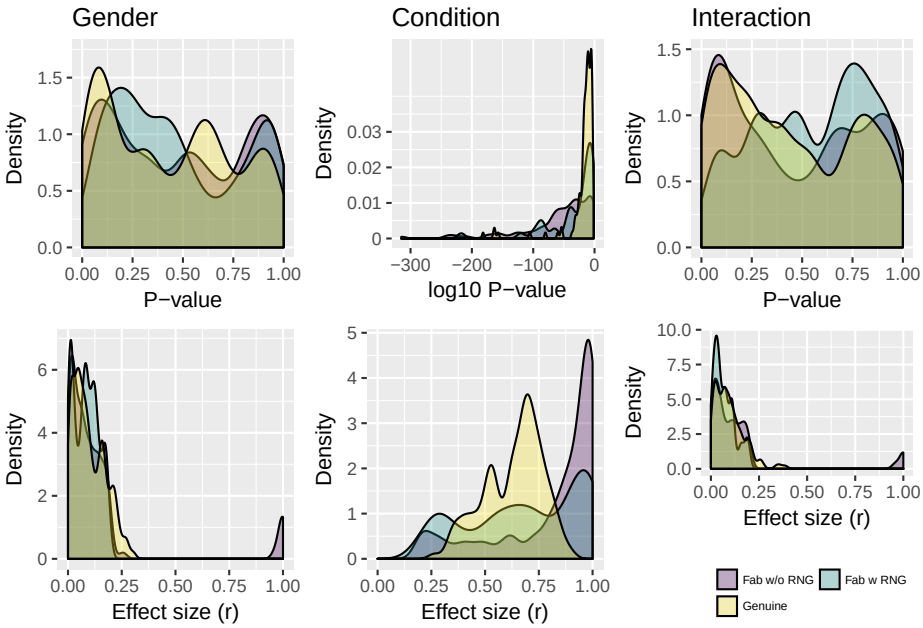

Figure 6: Density distributions of p-values and effect sizes for the gender effect, the anchoring effect, and the interaction effect across the four anchoring studies. This figure is similar to Figure XXX, except that each panel now separates the density distributions for fabricated results using a random number generator (RNG), fabricated results without using a RNG, and genuine effects. Respondents self-selected to use (or not use) RNGs in their fabrication process.

Table 4: 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 percent 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]

These results indicate that the fabricated effect sizes from participants who used RNGs are relatively more difficult to detect compared to data from participants who did not use a RNG (illustratively, the simple mean of the left column of Table 4 is 0.604 compared to the right column simple mean of 0.797). The numbers presented inTable 4 can be interpreted as the probability that the larger effect is fabricated, when presented with one genuine and fabricated effect size. For nonsignificant p-values, we obtained the following $AUROC$ 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 $max_z - min_z$) classification performance does not seem to be systematically 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]).

Note that participants self-selected the use of RNGs or not, and that we did not preregister these analyses. Given the small number of results (11 versus 28), we did not statistically test the differences due to lack of statistical power, and only present descriptive results.

Discussion

We presented the first controlled study on detecting data fabrication at the level summary statistics. As far as we could tell, 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 effect within their research domain (i.e., the anchoring effect), which was contextualized in realistic hypotheses, and compared them to genuine data on the same effect. We investigated the performance of the reversed Fisher method, variance analyses, combinations of these two methods, and statistically significant effect sizes to detect fabricated data.

Methods related to classifying statistically significant summary statistics (i.e., effect sizes and variance analyses) performed fairly well, whereas those relating to statistically nonsignificant summary statistics (i.e., p-value analyses) performed

poorly. Non-preregistered results suggest that variance analyses performed similarly or marginally better than using statistically significant effect sizes in this sample. Hence, we recommend using methods that investigate statistically significant effects to detect potential data fabrication, but prior to their application their assumptions should be well understood and tested.

We noted that the assumption of homogeneous population variances in the variance analyses has 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 indicated that the classification performance of variance analyses may strongly depend on satisfying this assumption, that is, the performance of the method is not robust to violations of the homogeneity assumption. The alternative approach to variance analyses using the range of variances instead of their standard deviation (i.e., $max_z - min_z$ rather than SD_z) seemed to be more robust to violations of the homogeneity assumption. This comparison was not preregistered and its performance could be studied further. Nonetheless, based on the success of using the dispersion of variances, 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 in the datasets studied here) and also consider using the range of standard deviations $max_z - min_z$).

Of all methods we applied, we obtained the best performance using the heterogeneous variance analyses, which resulted in detecting 9 out of 39 fabricated data sets (23%) and no false positives (0; $\alpha = .01$). Performance using (only) the statistically effect sizes was comparably good. Consequently, we failed to detect the majority of the fabricated datasets using statistical methods based on nonsignificant p-values, consistency of variances, and effect sizes. More worrisome is that for many methods the false positive rate was high, in one case even 100% (using $max_z - min_z$ based on the assumption of homogeneity of all variances).

Our finding that statistical analyses of data with fabrication detection tools may not be robust to violations of their assumptions has implications for investigations of research misconduct. Our results demonstrate that improper model specification can result in classifying anything as potentially fabricated (i.e., high false positive rate), which comes at high costs for all parties involved. Moreover, improper model specification may also result in a high false negative rate, as in our homogeneous variance analyses, resulting in a much too low $AUROC$ values (e.g., $AUROC = .264$). Our sometimes high false positive and false negative rates are especially worrisome in light of widespread application of statistical methods to screen for potential problematic studies (e.g., Carlisle, 2017a; Loadsman & McCulloch, 2017), when their validation is based on the criterion that the methods proved useful to detect problematic data in isolated research misconduct cases the past (e.g., Carlisle, 2012; Carlisle & Loadsman, 2016; D. R. Miller, 2015). For instance, the usefulness of the reversed Fisher method to detect problematic data in the past (Anonymous, 2012; Levelt, 2012) should not be taken as evidence of its validity for general application. Our study highlights the importance of validating methods with genuine reference data, before using these tools to flag potential problematic papers. Note that concerns like this have been expressed before (Evan D. Kharasch & Houle, 2017;

E. D. Kharasch & Houle, 2017; Mascha, Vetter, & Pittet, 2017; Moppett, 2017; Piraino, 2017).

Our results warrant further research on the underlying assumptions and validity of statistical approaches to detect potential data fabrication using summary statistics. This further research can help determine or prevent model misspecification, both in the assumptions of the statistical models and the psychology theory for specific ways of fabricating data before standard application of these methods in practice (see also Carlisle, 2017b).

For the reversed Fisher method that focused on the overly consistent results for effects that are expected to follow the null hypothesis, results indicated that participants did not fabricate excessive amounts of high p-values (i.e., closer to 1 than expected by chance) when told to fabricate statistically nonsignificant effects. This ran against our prediction that the absence of a true effect would prompt fabricators to fabricate results that do not contain enough randomness, resulting in too many high p-values. This is particularly noteworthy because this tenet has been helpful or even central to several known cases of research misconduct (Anonymous, 2012; Levelt, 2012). However, different from these specific cases, the results we asked participants to fabricate were first-order results (i.e., those immediately observable to the participants), whereas in the Stapel and Förster case, the reversed Fisher method showed potential data fabrication across second order results (i.e., similarity of means of experiments of different papers in the case of Stapel, or linearity test of first-order results in case of Förster). Hence, although our results indicate that the reversed Fisher method often does not perform well for inspecting first-order results, it may still perform well in isolated cases, particularly when applied to higher order results (see also Haldane, 1948).

Results of our reversed Fisher method are inexact because we used dependent fabricated results, which we did not take into account in our analyses. More specifically, for the p-value analyses we analyzed the four p-values from (for example) the gender effect across the four fabricated studies for one participant. This might have violated the assumption of independence, hence may have resulted in biased results of this test. Neither our analyses of the effect sizes nor our variance analyses suffer from this issue.

Analyses combining different data fabrication tools may not perform better than analyses based on a single tool, which also has implications for research misconduct investigations. First, a fabricated dataset does not imply that all tools should hint at data fabrication; fabricated data may resemble genuine data in some respects but not in others. Second, focusing on one aspect that best distinguishes fabricated from genuine data may perform best. The problem is then to identify that aspect, preferably before conducting the investigation. Our study suggests to focus on the analysis of properties of statistically significant effect sizes, whereas some fraud cases suggested to focus on properties of statistically nonsignificant effect sizes. We recommend, in cases of multiple independent possibly fabricated studies, to use several tools to identify possible fabrication in one study, and then apply and test the tools that worked to the other possibly fabricated studies (cross-validation). Importantly, we wish to emphasize that it does not make sense to require that *all* tools signal fabrication; as fabricated data may resemble genuine data in some respects, absence of one or several

signals should not be considered as evidence of no fabrication.

We also considered the possibility that the use of a Random Number Generator (RNG) to fabricate summary statistics could decrease the probability of detecting a fabricated dataset. Although we did not preregister these analyses, descriptive results suggest that using RNGs decreases the performance of using effect sizes to classify fabricated from genuine data. On the other hand, using RNGs did not substantially decrease the performance of the variance analysis that analyzed the effect sizes bearing on anchoring. Note that our results are solely descriptive due to too small group sizes for meaningful comparisons. We will investigate in Study 2 whether using RNGs affects the performance of detecting data fabrication in a similar fashion and revisit this issue in the general discussion.

We note that our presented results might be particular to the anchoring effect and not replicable with other effects. First, as opposed to many other effects in psychology, many data on the anchoring effect are already available and fabricators may have used these data when fabricating theirs. Second, 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 in the hundreds or thousands. Such relatively large values might feel more unintuitive to think about than smaller numbers in the singles or tens that might appear in other research contexts. Hence, we might be better at detecting potential data fabrication in data of our study compared to most other studies 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, fabrication of data of Likert scales may be more difficult (or easier) to detect than fabrication of continuous data.

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 problematic 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 individual level data

In Study 2 we tested the performance of statistical methods to detect fabrication of individual level (or raw) data. Our procedure is comparable to that used in Study 1: We again 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., individual level data) and included a face-to-face interview in which we debriefed participants on how they fabricated their data (C. 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 Ethics Review Board (EC-2015.50; https://osf.io/7tg8g/).

To test the validity of statistical methods to detect data fabrication in individual level data, we investigated individual level data of the classic Stroop experiment (Stroop, 1935). In a Stroop experiment, participants were 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 individual level 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 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) of response latencies for each participant 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 7 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 28 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. Two of these 80 researchers (2.5%) we contacted actually ended up participating in our study. Subsequently, we implemented a second, unplanned sampling stage where we collected e-mails from all PhD-candidates, teachers, and professors of psychology-related departments at Dutch universities. This resulted in 1,659

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 7: 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 confirmed the hypotheses, a checkmark appeared (upper right). This template is available at https://osf.io/2qrbs.

additional unique e-mails that we subsequently invited to participate in this study. Due to a malfunction in Qualtrics' quotum sampling, we oversampled, resulting in 28 participants instead of the originally intended 20 participants. The second sampling scheme was not part of the original ethics application, but was considered crucial to obtain a sufficiently large sample.

Each participant received instructions on the data fabrication task via Qualtrics and 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 7; 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 (see Data Analysis section for exact method used). Participants were not informed about how we planned to detect data fabrication; we used the combined Fisher method (described next). JGV transcribed the contents of the interview and 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).

Data analysis

To detect data fabrication in individual level data using statistical tools, we performed a total of sixteen analyses per dataset (preregistration: https://osf.io/ecxvn/) for each of the 21 genuine datasets and 28 fabricated datasets. These sixteen analyses consisted of four Newcomb-Benford Law (NBL) digit analyses, four terminal digit analyses, two variance analyses, four multivariate association analyses (deviated from preregistration in that we used a parametric approach instead of the planned non-parametric approach), a combination test of these methods, and effect sizes at the summary statistics level (the latter test replicated Study 1 and was not preregistered). We had one dataset for each participant fabricating data and for each lab in the Many Labs study, amounting to 49 datasets.

For the digit analyses (NBL and terminal), we separated the 25 Ms and 25 SDs per within-subjects condition and conducted χ^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 25 standard deviations of the response latencies in the congruent condition for excessive consistency separately from the

25 standard deviations of the incongruent condition. We conducted this analysis for each genuine and fabricated dataset, using the $max_z - min_z$ operationalization (not preregistered; based on results from Study 1 indicating that it is more robust to violations of the assumption of equal variances).

For the multivariate association analyses, we analyzed four correlations between 25 pairs of fabricated statistics (both Ms and SDs) and compared this correlation to the corresponding distribution of correlations for genuine data. More specifically, we did this for the (i) correlation between the means of congruent- and incongruent conditions, (ii) standard deviations of both conditions, (iii) means and standard deviations within the congruent condition, and (iv) means and standard deviations within the incongruent condition. We compared these correlations to the corresponding correlations for the genuine data after 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 for each dataset. More specifically, we included the p-values of ten statistical tests; four terminal digit analyses, two variance analyses, and four analyses of the multivariate associations. The results of this test served as the basis for selecting the top three fabricators. 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 (Diekmann, 2007), 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).

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 (C. 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.

Similar to Study 1, we computed the AUROC for each of these statistical methods to detect data fabrication. We again conducted all analyses using the `pROC` package (Robin et al., 2011). We also explored whether using Random Number Generators (RNGs) may have affected the detection of fabricated data in our sample by running AUROC analyses comparing genuine data and fabricated data with RNGs, or by comparing genuine data and fabricated data without RNGs.

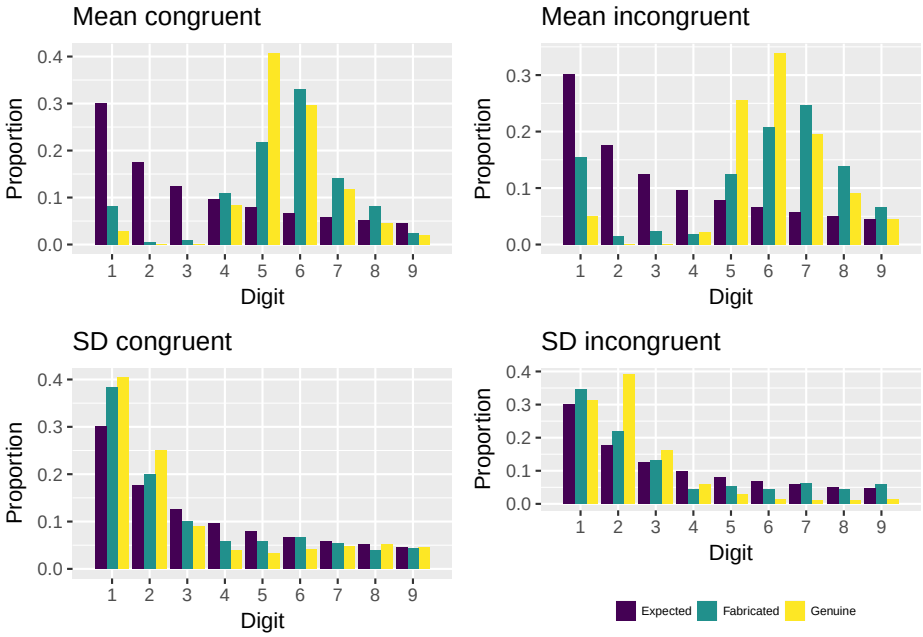

Figure 8: First (Benford) digit distributions of the (in)congruent means and standard deviations, aggregated across all Many Labs 3 datasets, across the datasets fabricated by the participants, and the theoretically expected proportions.

Results

Digit analyses

Figure 8 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 for means and probably also not appropriate for standard deviations. Figure 8 also shows that fabricated means and standard deviations differ from genuine means and SDs. Fabricated means seem systematically larger, with more dispersion than their genuine counterparts. Fabricated incronguent SDs seem smaller than those of genuine SDs. Note, however, that we did not plan to detect fabricated data using values or distributions of means and SDs directly (but see also the Variance analyis section next).

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]). Using the congruent standard deviations, we detected 19 of the 28 fabricated ones as fabricated ($\alpha = .01$) and 13 of the 21 genuine ones as fabricated (results per respondent available at osf.io/dsbge). Values from other measures showcase that the fabricated data are actually *more* in line with the NBL than the genuine data. Consequently, the genuine data and fabricated data are often wrongly classified. This is reflected by the AUROC values that are significantly smaller than .5. For congruent means, $AUROC = 0.039$, 95% CI [0-0.087]; Using the congruent means, we detected 28 of the 28 fabricated ones as fabricated ($\alpha = .01$) and 21 of the 21 genuine ones as fabricated (results per respondent available at osf.io/sgda8). For incongruent means, $AUROC = 0.024$, 95% CI [0-0.059]; Using the incongruent means, we detected 28 of the 28 fabricated ones as fabricated ($\alpha = .01$) and 21 of the 21 genuine ones as fabricated (results per respondent available at osf.io/xjsd6). For incongruent standard deviations, $AUROC = 0.156$, 95% CI [0.045-0.268]; Using the incongruent standard deviations, we detected 18 of the 28 fabricated ones as fabricated ($\alpha = .01$) and 21 of the 21 genuine ones as fabricated (results per respondent available at osf.io/2sd7w).

Figure 9 shows the aggregated terminal digit distributions of the genuine and fabricated data side-by-side with the expected terminal digit distributions. The first row depicts the terminal digit distributions of the means, 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 indeed show that terminal digit analyses perform close to chance level classification of genuine and fabricated data. More specifically, for the incongruent standard deviations, $AUROC = 0.511$, 95% CI [0.343-0.679]; 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]. The terminal digit analysis classified at most 2 of the 28 fabricated datasets as being fabricated (and 2 of the 21 genuine data as being fabricated; $\alpha = .05$).

Variance analysis

Figure 10 indicates that the standard deviations of genuine data are larger on average and more dispersed. Results indicate that the fabricated and genuine data can be perfectly separated based on results from the variance analyses ($max_z - min_z$). 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

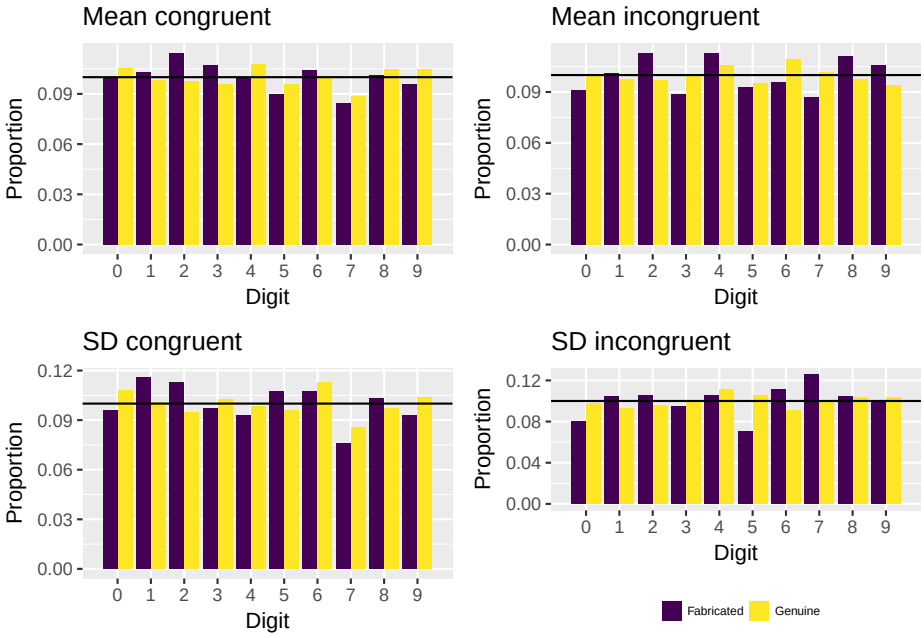

Figure 9: Terminal digit distributions for the (in)congruent means and standard deviations, aggregated across all Many Labs 3 datasets or across the datasets fabricated by the participants.

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. This method is particularly difficult to apply if no reference distribution of (arguably) genuine data is available.

Upon closer inspection of the individual level results of the variance analyses per data set, all p-values are statistically significant if compared to traditional α levels (i.e., .05; maximum 0.006 across both the genuine- and the fabricated data). As a result, we recommend that variance analyses are only used when a reference model is available (in line with the results from Study 1).

Multivariate associations

We expected that fabricated multivariate associations would be different from genuine multivariate associations. Using the parametric test of multivariate associations, results indicate classification is fair to good in the current sample. Figure 11 shows the density distributions of the various multivariate associations (rows 1-2), which already indicates that genuine data are less dispersed 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,

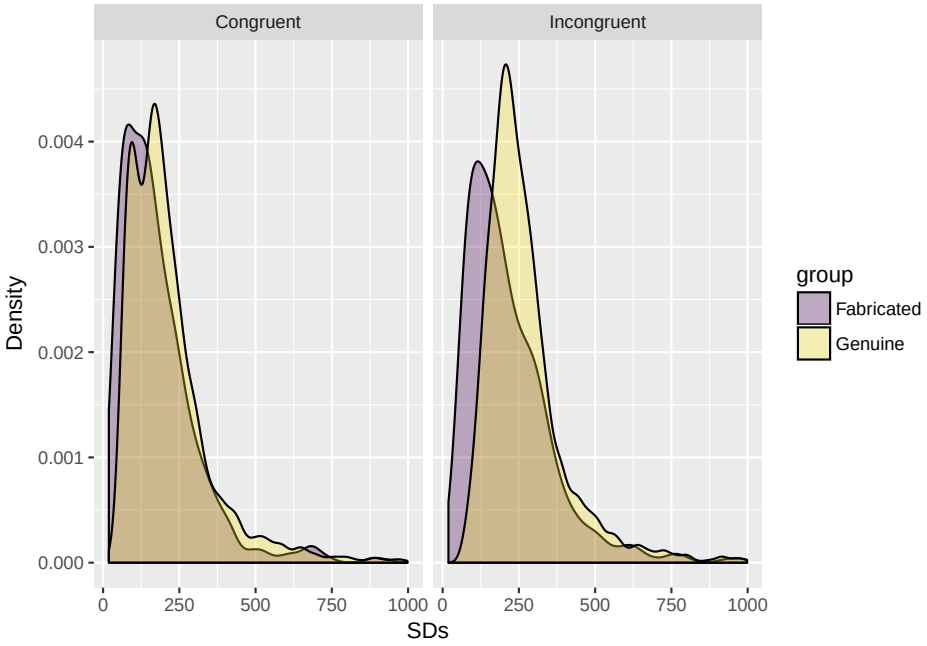

Figure 10: Density distributions of the standard deviations of the response times in the congruent conditions (left) and the incongruent conditions (right), split for the genuine and fabricated data. X-axis truncated at 1000.

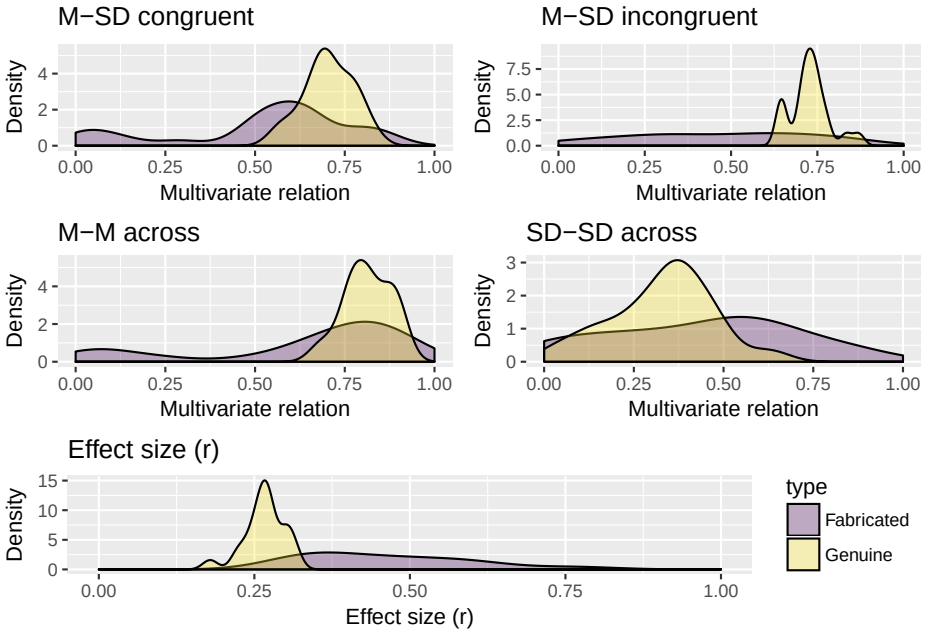

Figure 11: Density distributions of the multivariate relations (first two rows) and the effect sizes (final row), split for the genuine and fabricated data.

$AUROC$ values range from 0.549 through 0.842. More specifically, the $AUROC$ for the various sets of relations (going clockwise with the first four figures in Figure 11) 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, $AUROC = 0.549$, 95% CI [0.379-0.72] for SD-SD across conditions. The percentage of fabricated multivariate relations that is larger than all 21 genuine multivariate relations is 7.1% for M-SD congruent, 0% for M-SD incongruent, 7.1% for M-M across, and 14.3% for SD-SD across. Overall, it seems that comparing multivariate associations to known genuine ones is a good way to detect (potential) data fabrication, with the connotation that a reference distribution is needed.

Combining variance, terminal digit, and associational analyses

As preregistered, we combined both variance analyses, the terminal digit analyses, and the tests of the multivariate associations with the Fisher method (10 results in total). 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. The maximum p-value of the combination

of these tests for either the genuine or fabricated data is 0.003 (results per respondent available at osf.io/rke9q), indicating that all datasets would be classified as fabricated if we did not compare the results from the genuine and fabricated data.

Extreme effect sizes

Figure 11 (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), with fabricated effect sizes generally being larger than genuine effect sizes. 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 (results per respondent available at osf.io/). As such, this is a particularly good result within this sample (we did not preregister this analysis).

Fabricating effects with Random Number Generators (RNGs)

Using Random Number Generators (RNGs) in the individual level data fabrication procedure did not seem to have a substantial effect on how genuine the fabricated results appeared. We explored this in our data (i.e., not preregistered) and Table 5 presents the AUROC values split on participating researchers who said they used ($k = 19$) or did not use RNGs ($k = 9$) to fabricate data (based on manual coding of the interview transcripts). Noteworthy from our exploration is that the effect size distribution seems approximately similar for both data fabricated with and without RNGs (Figure 12). 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 that use individual-level (raw) data can distinguish fabricated data from genuine data. To this end, we replicated the procedure from Study 1 and asked researchers to fabricate data for individual participants for the classic Stroop task. We also collected (arguably) genuine data from the labs involved in the Many Labs study, which included the classic Stroop task. As such, we had both genuine and fabricated data sets on the same effect.

Using these data sets we attempted to classify genuine and fabricated individual level data using digit analyses, variance analyses, multivariate associations, and effect sizes. Results of preregistered analyses indicate that digit analyses of raw data performed at chance level, variance analyses of individual level data performed excellently, and analyses of multivariate relations between variables in the individual level data performed fairly to excellently. Moreover, the summary statistic effect size appeared to strike a surprisingly good balance between efficacy and parsimony for classifying fabricated- from genuine individual level

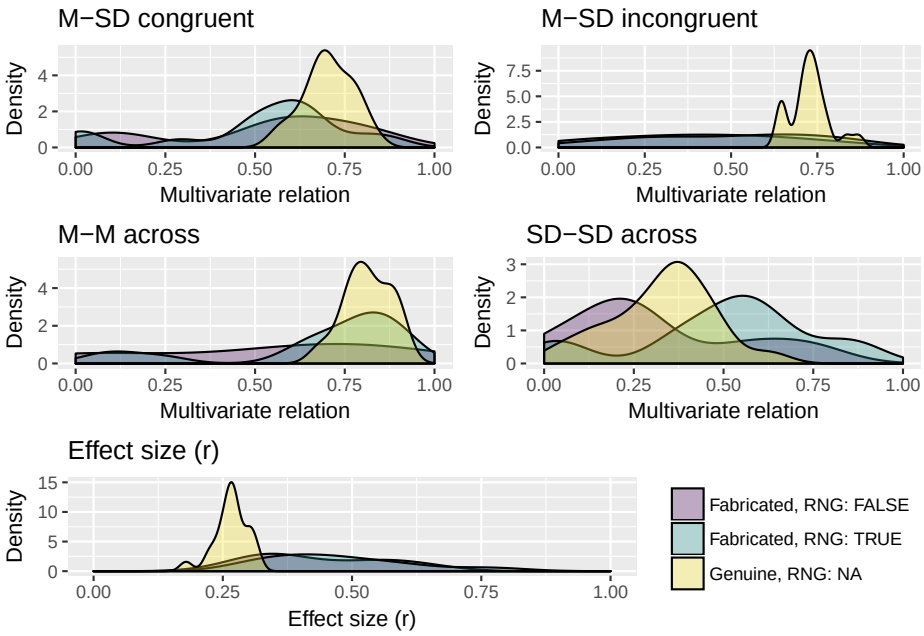

Figure 12: Density distributions of the multivariate relations (first two rows) and the effect sizes (final row), split for the genuine data, the fabricated data without using Random Number Generators RNGs), and fabricated data with using RNGs.

Table 5: AUROC values with 95 percent 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 with 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]
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]

data (only superseded in performance by the more complex variance analyses). This replicates the finding from Study 1 that effect sizes are a valuable piece of information to discern genuine from fabricated data. Fabricators' use of Random Number Generators (RNGs) did not appear to have a consistent relation with classification performance with individual level 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 8 (clearly) indicates this is not the case for summary statistics of response times. Response times are untruncated ratio measures in theory that technically 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). If the NBL is being considered for applications to detect (potential) misconduct, there need to be indications that the data generation process is in line with the requirements of the NBL, but we consider that this is hardly the case for experimental studies in the social sciences.

Going against our predictions, participants fabricated individual level data that was almost indistinguishable from the genuine individual level data when looking at terminal digit analyses. Given the theoretical framework we use, wherein humans are expected to be poor at fabricating stochastic processes that underlie data collection procedures, we expected that our participants would be unable to fabricate uniformly distributed terminal digits. Our sample 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 human's lack of intuitive understanding of uniform probabilities manifests itself in fabricated individual level data, and if so, under which conditions.

Study 2 replicated the effectiveness of variance analyses (preregistered) and effect sizes (not preregistered) to detect data fabrication, but failed to replicate the potential effect of RNGs on detection rates (not preregistered). These mixed results with respect to the effect of RNGs on the fabricated data suggests that a lack of intuitions for probabilities does not necessarily manifest itself in fabricated data. Hence, further research might look into correlating the (lack of) expertise on probabilities and the kind of data being fabricated. With respect to variance analyses and effect sizes, our results suggest that these are the most promising methods when genuine data are available (we further discuss this in the General Discussion).

Study 2 substantiates two conclusions from Study 1: (1) As methods may not be robust to violations of its assumptions (e.g., NBL in Study 2), these methods should be validated with genuine reference data if available, before using these tools to flag potential problematic papers. This dependence on assumptions also questions the validity of automatic large-scale scrutiny for data fabrication. (2) Although some methods did not perform well in Study 2, these methods have shown to work well to detect data fabrication in some isolated cases of misconduct. For instance, both the NBL (Cho & Gaines, 2007) and the analysis of terminal digits (Mosimann et al., 1995) have shown their usefulness in some cases. Similarly, although some methods worked well in Study 2 (i.e. variance analyses, effect size distributions, multivariate associations), this does not mean that they always work well in detecting fabricated data, or that they could exonerate anyone when these methods fail to flag any fabrication.

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 individual level data from two well-known psychology research topics, it seems like classification based on statistically significant effect sizes strikes the best balance between parsimony, effectiveness, and usability. On the other hand, variance analyses are a good option that is somewhat more complex in its application because one has to identify the sets of variances that can be expected to be homogeneous. The digit analyses based on the Newcomb-Benford law and the terminal digit principle did not perform well. We bundled our functions for the variance and digit analyses and the (reversed) Fisher method in the `ddfab` (short for detecting data fabrication) package for `R`, which is available through GitHub (https://github.com/chartgerink/ddfab) for application in further research and development.

We designed the current studies to have sufficient information to detect data fabrication within a given set of data, but not necessarily to generalize our results to a larger population. As such, the sample sizes of the presented studies and the type of effect we chose as the empirical context necessarily restrict the drawing of more general inferences. Further research should consider whether these results also apply to other types of data or effects. Nevertheless, our studies have highlighted that variance- and effect size analysis and multivariate

associations 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 (although we note those of the Fisher method are incorrect due to dependent p-values). 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 depends on the availability of unbiased genuine data to compare to and cannot readily be done by using generic decision criteria such as α-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 impossible or problematic at best. More specifically, we would have classified all datasets as fabricated if we had used the traditional hypothesis testing approach. Hence, we agree with the call to always include a control sample when applying these statistical tools to studies that look suspicious (Simonsohn, 2013). It is for exactly this reason that we refrain from formulating general decision rules for the methods presented in this paper. This might also have implications for systematic applications of statistical methods to detect potentially problematic data, such as the recent application by Carlisle (2017a). Carlisle (2017a) used the same method applied in the Fujii case to approximately 5,000 clinical trials without any further validation of the methods (Bolland, Gamble, Avenell, & Grey, 2019). Our results suggest that in practice aberrant effects are best detected in relative fashion, for example in a meta-analysis (corroborating our own anecdotal experience), or to look for excessively large effect sizes (e.g., $r > .95$) as an initial screening of a set of effects (especially when that effect size is larger than the reliability of the product of the measures involved). Using absolute classification (i.e., $p < \alpha$) can be problematic, considering that many of the methods we tested (e.g., variance analyses, digit analyses) are not specific enough and rely on models with strong assumptions, potentially flagging both genuine and fabricated data as problematic.

Because we included the Many Labs data (Ebersole et al., 2016; Klein et al., 2014) we had (arguably) 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 have found inflated effect size estimates of the anchoring- or Stroop effects due to publication bias. These inflated effect size estimates could have resulted 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 publication bias and fabricating data might have similar effects in turn conflates the detection of fabricated data. Collecting an unbiased genuine effect distribution thus requires careful attention; when arguably genuine effects are collected from a literature ridden with publication bias and related biases, detection of data fabrication may be undermined. 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 Hartgerink & Zelst, 2018).

Our results depend on the (majority of the) Many Labs data being genuine. We

remain confident that (most of) the Many Labs data are genuine for a variety of reasons. First, the sheer number of people involved in these projects 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 of course 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 in screening for problems in the published literature (e.g., Bik et al., 2016; Carlisle, 2017a) or during peer review is that the diagnostic value of any instrument is dependent on the base rate of afflicted cases (here: fabricated data). In our study design, we built in 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 strongly 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 (see also Stricker & Günther, 2019) that would cause approximately 1 out of 3 cases of 'detected data fabrication' to be false. Hence, we do not recommend attempting to detect data fabrication on statistical methods alone.

We do advise to use some of the more successful statistical methods as screening tools in review processes and as additional tools in formal misconduct investigations where prevalence is supposedly higher than in the general population of research results. We note that this should only happen in combination with evidence from other sources than statistical methods (e.g., focusing on practical, methodological, or substantive aspects). As we mentioned before, excessively large effect sizes might be used as a screening approach for further manual or in-depth investigation, but we warn against the potential for confirmation bias that results from these earlier tests might create. 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 and using a rigorous protocol that involves due care and due process.

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 appropriate for several 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 such suspect data. 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 any negative effect by using findings from psychology research to decrease potential carry-over of this controlled misbehavior (Mazar et al., 2008; although a recent multilab replication contested this effect, Verschuere et al., 2018). 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 shy away from conducting this kind of research to begin with. However, we argue that the potential ethical uses of these methods are substantial (improved detection of fabricated data by a potential many) and 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 from preregistered projects improves the unbiased estimation of effect sizes and multivariate associations, where the peer-reviewed literature inflates estimated effect sizes due to publication bias and often lacks the required information to compute these multivariate associations. 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 were fabricated based on existing data (LaCour & Green, 2014; McNutt, 2015). Researchers detected that this data had been fabricated because it seemed to be a(n almost) linear transformation

of variables because they could access the relevant dataset (Broockman, Kalla, & Aronow, 2015). 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. As such, beyond being fruitful for examining reproducibility (Munafò et al., 2017) and facilitating new research, open data may also facilitate the improvement of detecting potential data fabrication. 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 causes of data fabrication possible. One 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. 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 more sustainable and effective. Nonetheless, we also recognize that there will always be dishonesty involved for some researchers, and we recommend that research engage in more penetration testing of how those with dishonesty can fool a system.

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

Anonymous. (2012). Suspicion of scientific misconduct by Dr. Jens Foerster. Retrieved from http://wayback.archive.org/web/20170511084213/https://retractionwatch.files.wordpress.com/2014/04/report_foerster.pdf

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

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

Bolland, M. J., Gamble, G. D., Avenell, A., & Grey, A. (2019). Rounding, but not randomization method, non-normality, or correlation, affected baseline p-value distributions in randomized trials. *Journal of Clinical Epidemiology*, *110*, 50–62. http://doi.org/10.1016/j.jclinepi.2019.03.001

Broockman, D., Kalla, J., & Aronow, P. (2015). Irregularities in LaCour (2014). Retrieved from https://wayback.archive.org/web/20180823093137/http://stanford.edu/~dbroock/broockman_kalla_aronow_lg_irregularities.pdf

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

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. (2017a). 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. (2017b). Seeking and reporting apparent research misconduct: Errors and integrity - a reply. *Anaesthesia*, *73*(1), 126–128. http://doi.org/10.1111/anae.14148

Carlisle, J. B., & Loadsman, J. A. (2016). Evidence for non-random sampling in randomised, controlled trials by yuhji saitoh. *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

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology*, *55*(1), 591–621. http://doi.org/10.1146/annurev.psych.55.090902.142015

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., 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., & Zelst, M. van. (2018). As-you-go instead of After-the-fact: A network approach to scholarly communication and evaluation. *Publications*, *6*(2), 21. http://doi.org/10.3390/publications6020021

Hartgerink, C., 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., 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 Presss.

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 Anaesthesist*, *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.

Kharasch, E. D., & Houle, T. T. (2017). Errors and integrity in seeking and reporting apparent research misconduct. *Anesthesiology*, *127*(5), 733–737. http://doi.org/10.1097/aln.0000000000001875

Kharasch, E. D., & Houle, T. T. (2017). Seeking and reporting apparent research misconduct: Errors and integrity. *Anaesthesia*, *73*(1), 125–126. http://doi.org/10.1111/anae.14147

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

LaCour, M. J., & Green, D. P. (2014). When contact changes minds: An experiment on transmission of support for gay equality. *Science*, *346*(6215),

1366–1369. http://doi.org/10.1126/science.1256151

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/

Loadsman, J. A., & McCulloch, T. J. (2017). Widening the search for suspect data - is the flood of retractions about to become a tsunami? *Anaesthesia*, *72*(8), 931–935. http://doi.org/10.1111/anae.13962

Mascha, E. J., Vetter, T. R., & Pittet, J.-F. (2017). An appraisal of the carlisle-stouffer-fisher method for assessing study data integrity and fraud. *Anesthesia & Analgesia*, *125*(4), 1381–1385. http://doi.org/10.1213/ane.0000000000002415

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

McNutt, M. (2015). Editorial expression of concern. *Science*, *348*(6239), 1100–1100. http://doi.org/10.1126/science.aac6184

Miller, D. R. (2015). Probability screening in manuscripts submitted to biomedical journals - an effective tool or a statistical quagmire? *Anaesthesia*, *70*(7), 765–768. http://doi.org/10.1111/anae.13165

Moppett, I. K. (2017). Errors in published papers are multifactorial. *Anaesthesia*, *72*(11), 1415–1416. http://doi.org/10.1111/anae.14048

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

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Sert, N. P. du, … Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1). http://doi.org/10.1038/s41562-016-0021

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

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 (19852013). *Behavior Research Methods*, *48*(4), 1205–1226. http://doi.org/10.3758/s13428-015-0664-2

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

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

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

Piraino, S. W. (2017). Issues in the statistical detection of data fabrication and data errors in the scientific literature: Simulation study and reanalysis of carlisle, 2017. http://doi.org/10.1101/179135

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

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology. *Psychological Science*, *22*(11), 1359–1366. http://doi.org/10.1177/0956797611417632

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

Stricker, J., & Günther, A. (2019). Scientific misconduct in psychology. *Zeitschrift Für Psychologie*, *227*(1), 53–63. http://doi.org/10.1027/2151-2604/a000356

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

Verschuere, B., Meijer, E. H., Jim, A., Hoogesteyn, K., Orthey, R., McCarthy, R. J., ... Yildiz, E. (2018). Registered replication report on mazar, amir, and ariely (2008). *Advances in Methods and Practices in Psychological Science*, 251524591878103. http://doi.org/10.1177/2515245918781032

Wagenaar, W. A. (1972). Generation of random sequences by human subjects: A critical survey of literature. *Psychological Bulletin*, *77*(1), 65–72. http://doi.org/10.1037/h0032060

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., Aert, R. C. M. van, & Assen, M. A. L. M. van. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, *7*. http://doi.org/10.3389/fpsyg.2016.01832

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