

Detection of data fabrication using statistical tools

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

20 August, 2018

1 Abstract

PLACEHOLDER

2 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 Voignot; 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.

In order to stifle attempts at data fabrication, improved detection of fabricated data is considered to deter such behavior. 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., Parker & Hamblen, 1989) and are widely implemented not only at journals but also in the evaluation of student theses (e.g., with commercial services such as Turnitin). 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; Tversky & Kahneman, 1971, 1974; Wagenaar, 1972), they might create fabricated data that fail to follow the fundamentally probabilistic nature of genuine data. Whether the data and outcomes of analyses based on these data are in line with the (at least partly probabilistic) processes that are assumed to underlie them, may indicate deviations from the reported protocol, potentially even data fabrication or falsification.

Statistical methods have proven to be of importance in initiating data fabrication investigations or in assessing scope of potential data fabrication. For example, Kranke, Apfel, and Roewer skeptically perceived Fujii’s data (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” (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, Dexter, Pandit, Shafer, & Yentis, 2015; Carlisle & Loadman, 2016) 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, we need to study their diagnostic value to inform decisions about the utility of these methods. Specifically, many of the developed statistical methods to detect data fabrication are quantifications of case specific suspicions by researchers, 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 are not included. Using different statistical methods to detect fabricated data using genuine versus fabricated data could offer information 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.¹

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 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 raw 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.

3 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 can be computed (e.g., an APA reported t - or F -statistic; C. Hartgerink et al., 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

¹ Jelte, deze alinea heb ik veel laten staan ondanks dat je het ‘gemakzuchtig’ geschreven vond. Ik vind de punten die jij weghaalde wel relevant. Ik heb een aantal tekstuele wijzigingen wel meegenomen.

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 et al., 2016a; 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 still occur (Hartgerink et al., 2016a; 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 be 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. In Table 1, we simulated normal distributed measurements and t -test comparisons for statistically equivalent populations (Set 1). We also fabricated data for equivalent groups, where we determined the mean and standard deviation first and then added 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.

	Set 1			Set 2		
	Experimental	Control	P-value	Experimental	Control	P-value
	M (SD)	M (SD)		M (SD)	M (SD)	
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; 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

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

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

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

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, 2017; Carlisle et al., 2015).

As an example, we apply the reversed Fisher method to both the genuine- and fabricated results from Table 1. Using the threshold $t = 0.05$ to select only the nonsignificant results from Table 1, we retain $k = 10$ genuine p -values and $k = 10$ fabricated p -values. This results in $\chi_{2 \times 10}^2 = 18.362, p = 0.564$ for the genuine data (Set 1), and $\chi_{2 \times 10}^2 = 66.848, p = 6 \times 10^{-7}$ for the fabricated data (Set 2). Another example, from the Fujii case (Carlisle, 2012), illustrates that the reversed Fisher method may also 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, $\chi_{2 \times 5}^2 = 19.335, p = 0.036$. 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 analysis²

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

$$var(x) \sim \frac{\chi_{n_j-1}^2}{n_j - 1} \quad (3)$$

where n is the sample size of the j th group. Assuming equal variances of the J populations, this population

²Marcel, ik heb ervoor gekozen de structuur te houden, omdat deze vergelijkbaar opbouwend is als de Fisher method en elders. Ik heb mijn best gedaan het te simplificeren op punten en het allemaal wat vlotter te laten lopen. Hoop dat dit werkt.

variance is estimated by the Mean Squares within (MS_w) as

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

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 variances in group j (i.e., z_j^2) is

$$z_j^2 \sim \left(\frac{\chi_{n_j-1}^2}{n_j - 1} \right) / MS_w \quad (5)$$

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.

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., $P(SD_{z_{obs}} \leq SD_{z_{exp}})$).³ In other words, we compute how many samples of j groups 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. To recapitulate, we simulated data from a genuine probabilistic process using normally distributed measurements and t -test comparisons for statistically equivalent populations (Set 1); we fabricated data for equivalent groups (Set 2), where we determined the mean and standard deviation first and then added uniform noise to these parameters. 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. Using the variance analysis procedure with the SD_z measure of the dispersion of variances, we can quantify how extreme this difference is: Set 1 has no excessive consistency in the dispersion of the standard deviations ($p = 0.214$), whereas Set 2 does show excessive consistency in the dispersion of the standard deviations ($p = 0.006$). In words, out of 100,000 theoretically expected samples under the null model of independent groups with equal variances on a normally distributed measure, 2.142×10^4 showed less dispersion in standard deviations for Set 1, whereas only 572 showed less dispersion in standard deviations for Set 2.⁴ 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. The standard deviation of these standard deviations is 0.54 (i.e., SD_z). Such consistency in standard deviations (or even more) would only be observed in 1.21% of 100,000 simulated replications (Simonsohn, 2013).

³Marcel, bedoel je zoiets voor formule voor de test statistic?

⁴Marcel, ik voeg hier even geen plaatjes toe omdat ik standaardfuncties heb gemaakt en niet al die iteraties opsla los. Kost me op het moment teveel tijd om dat uit te vogelen met de revisies die nog liggen. Ik run deze dingen allemaal als ik het document genereer vanuit R, dus is wel reproduceerbaar :')

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 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 $r = 0.176$ (Hartgerink et al., 2017). These effect sizes can readily be recomputed based on data extracted with `statcheck` across thousands of results (Hartgerink, 2016; Nuijten et al., 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 occurring 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:

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

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).

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

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

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.

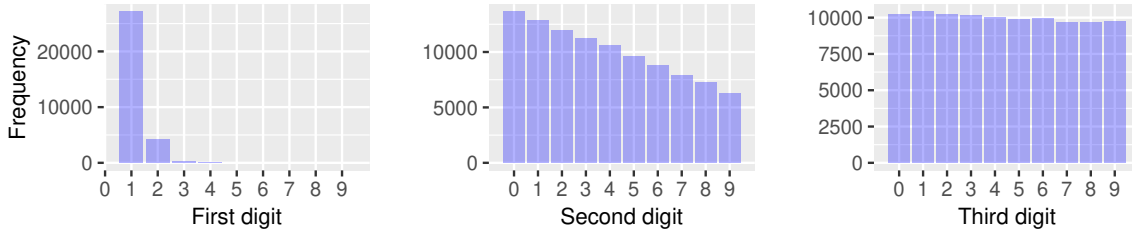


Figure 1: Frequency distributions of the first-, second-, and third digits. We sampled 100,000 values from a standard normal distribution to create these digit distributions.

Multivariate associations

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 negative 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 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).

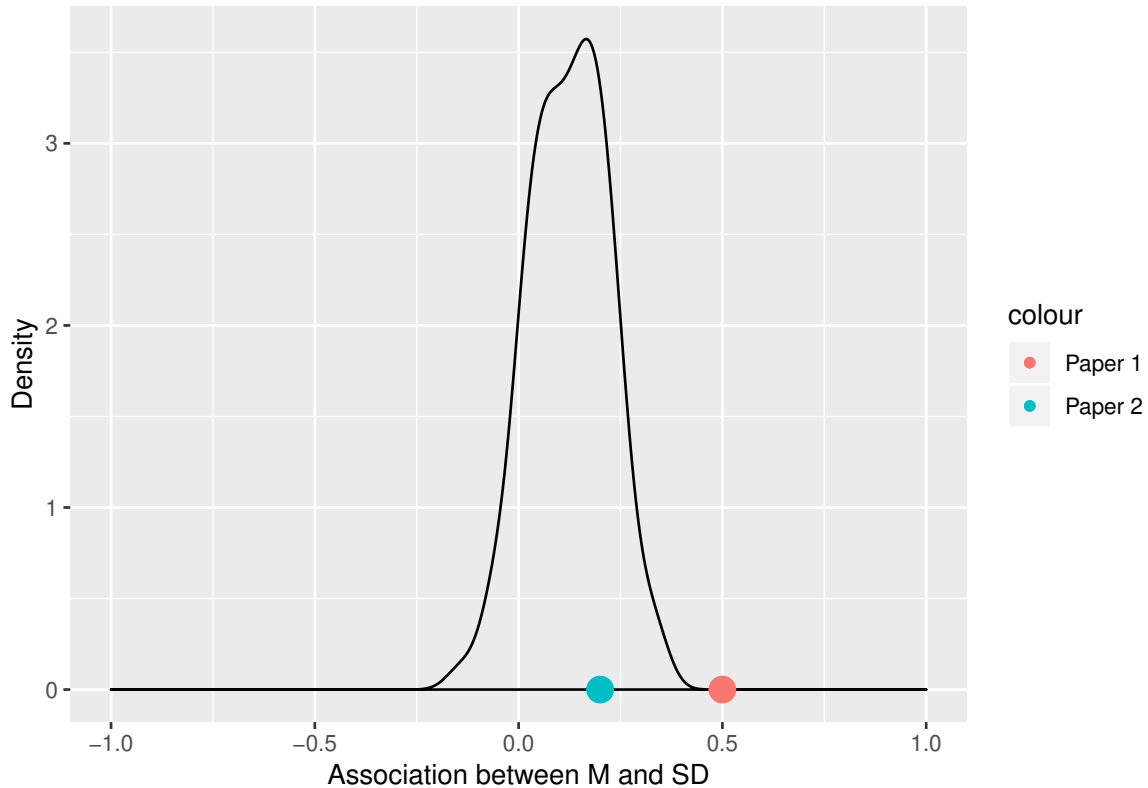


Figure 2: Distribution of 100 simulated observed associations between Ms and SDs for a response latency task; simulated under $N(.123, .1)$. The red- and blue dots indicate observed multivariate associations from fictitious papers. Paper 1 may be considered relatively extreme and of interest for further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

4 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; 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, 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 (Tversky & Kahneman, 1974). We chose this study 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 drawn 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 Ethical 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 locations we computed three summary statistics (i.e., sample sizes, means, and standard deviations) for each of the four conditions in the four anchoring studies (i.e., $3 \times 4 \times 4$; data: <https://osf.io/5xgcp/>). We computed these summary statistics from the raw ML data, which were cleaned using the original analysis scripts from the ML project.

The sampling frame for the participants asked to fabricate data consisted of 2,038 psychology researchers who published a peer-reviewed paper in 2015, as indexed in Web of Science (WoS) with the filter set to the U.S. We sampled psychology researchers to improve familiarity with the anchoring effect (Jacowitz & Kahneman, 1995; 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.⁵ 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

⁵We discovered that we included several non-U.S. researcher against our initial aim. We filtered Web of Science on U.S. origin, but found out that this meant that one of the authors on the paper was U.S. based. As such, corresponding authors might still be non-U.S. Based on a search through the open ended comments of the participant’s responses, there was no mention of issues in fabricating the data related to the metric or imperial system.

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 (the procedure for this is explained in the Data Analysis section). The provided e-mail addresses were unlinked from individual responses upon sending the bonus gift cards. The full Qualtrics survey is available at osf.io/rg3qc.

Each participant was instructed to fabricate 32 summary statistics ($4 \text{ studies} \times 2 \text{ anchoring conditions} \times 2 \text{ sexes} \times 2 \text{ statistics [mean and SD]}$) that corresponded to three hypotheses. We instructed participants to fabricate results for the following hypotheses: there is (i) a positive main effect of the anchoring condition, (ii) no effect of sex, and (iii) no interaction effect between condition and sex. We fixed the sample sizes in the 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 (fabrication) hypotheses, a checkmark appeared as depicted in Figure 3. We required participants to copy-paste the yellow cells into Qualtrics. This provided a standardized response format that could be automatically processed in the analyses. Technically, participants could provide a response that did not correspond to the instructions but 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); although a recent multilab replication contested this finding, osf.io/cwavm/). We rewarded participation with a \$30 Amazon gift card. The fabricated results that were most difficult for us to detect with the combination method (described next) as fabricated received a bonus \$50 Amazon gift card. Participants were not informed about how we planned to detect data fabrication. Using quota sampling, we collected as many responses as possible for the available 36



rewards, resulting in 39 fabricated data sets (<https://osf.io/e6zys>; 3 participants did not participate for a bonus).

Data analysis⁶

We analyzed the genuine- and fabricated data sets for each of the anchoring studies using four types of analyses. Each of these analyses is conducted per set of four anchoring studies, fabricated either by our participants or retrieved from the individual labs in the Many Labs data. First, we applied the reversed Fisher method. Second, we applied variance analyses. This  combined the individual results using the original Fisher method (a meta-analysis method; Fisher, 1925). Fourth, we used  four effect sizes of the statistically significant anchoring effect.

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  (one per study) and one for the four statistically nonsignificant interaction effects (one per study).

Specifically for the variance analyses, we substantially deviated from the preregistration (<https://osf.io/tshx8/>) and added multiple analyses. We analyzed the sample variances of  four anchoring studies per lab or participant in fourteen ways. For each of the variance analyses, we conducted them 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., $max_z - min_z$; see also the Theoretical Framework). First, we analyzed the 16 sample variances from the four anchoring studies (four per study), combining them in  the variance analysis as preregistered. However, only upon analyzing these values, we realized that the variance analyses assume that the included variances are  the same population distribution. Assuming homogeneous populations of variances is not necessarily realistic for the different anchoring conditions. Hence, we included variance  analyses based on subgroups, where we analyzed each anch  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. Of these 28 variance analyses (14 for each dispersion of variances measure), only the first one described here was preregistered.

We also combined the reversed Fisher method results with the results from the different variance analyses using the original Fisher method. More specifically, we combined the results from  reversed Fisher method analyses (one for the gender effects and one for the interaction effects) with the variance analysis combining the variances of the four anchoring studies, assuming homogeneous population variances (preregistered). We also included combinations where the variance analysis was conducted per study separately (including four variance analysis results), per anchoring condition for each study separately (including eight variance analysis results), or across all studies combined but taking into account heterogeneous variances per anchoring condition for each study (including one variance analysis result). We only conducted this combination test for the results from the variance analyses using dispersion of variance measure  on the standard deviation of the variances (i.e., SD_z ; not preregistered). Note that the performance of combining various analyses as we do here is dependent on the performance of the individual results included in the combination (e.g., if all included results perform well the combination 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 (much) larger than genuine statistically significant effects. As such, we inspected statistically significant anchoring effects four times, once for each anchoring study separately across the participants fabricating data and the original data from the separate labs in the Many Labs project (not preregistered).

For each of the previously described statistical methods to detect data fabrication, we carried out sensitivity and specificity analyses using Area Under Receiving Operator Characteristic (AUROC) curves. AUROC-analyses summarize the sensitivity (i.e., True Positive Rate [TPR]) and specificity (i.e., True Negative Rate [TNR]) for various decision criteria (e.g., $\alpha = 0, .01, .02, \dots, .99, 1$). For our purposes,

⁶Marcel, je vraagt of deze sectie compleet is, maar ik denk het wel. De eerste paragraaf geeft aan dat er (1) reversed Fisher method gebruikt word, (2) variantie analyses, (3) effect sizes, (4) combinatie van 1+2. De SD_z en $max-min_z$ zijn in de theoretisch kader al toegelicht; daar waar je om formules vroeg. Ik heb geprobeerd het te verduidelijken door de analyses elk 1 paragraaf te geven ipv alleen de variantie analyses :

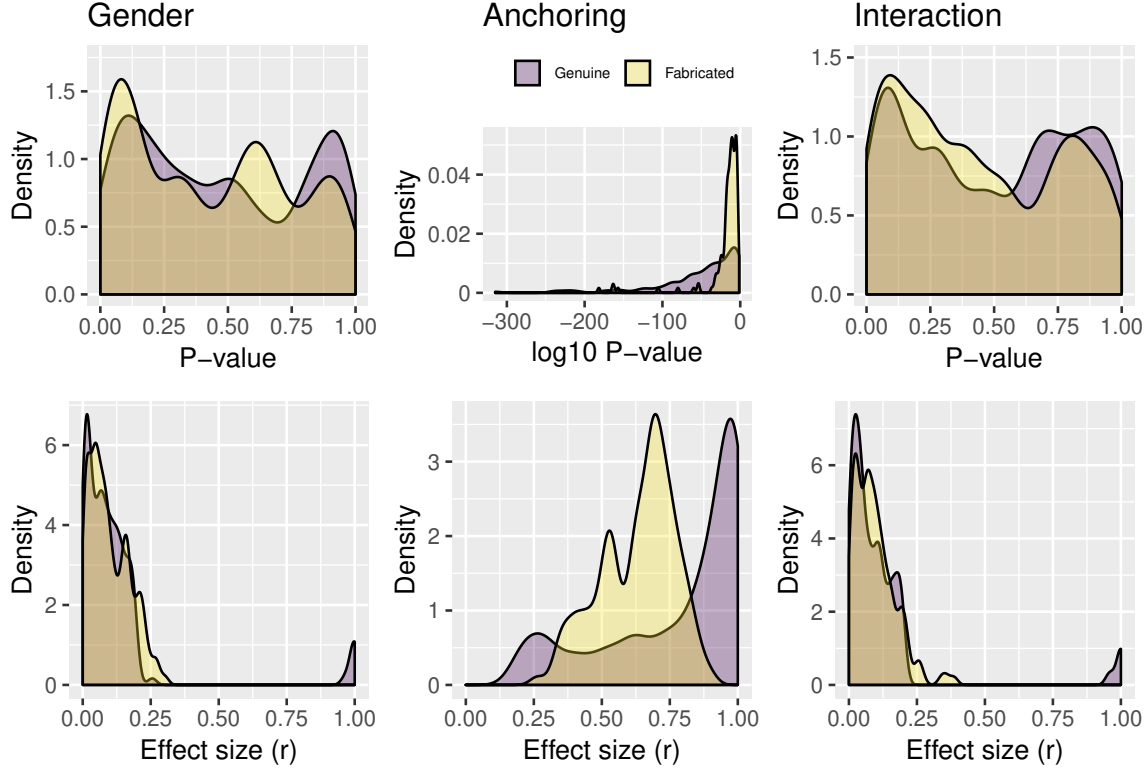


Figure 4: Density distributions of genuine- and fabricated summary statistics across four anchoring studies, per effect (gender, anchoring, or interaction) and type of result (p-value or effect size).

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

P-value analysis

When we apply the reversed Fisher method to the statistically nonsignificant effects, results indicate its performance is approximately equal to chance classification. We find $AUROC = 0.501$, 95% CI [0.468-0.535] for statistically nonsignificant gender effects and $AUROC = 0.516$, 95% CI [0.483-0.549]

⁷Labels van de plot kloppen wel; de distributies zijn transparant omdat ze over elkaar heen lopen en daardoor krijg je dat.

for statistically nonsignificant interaction effects. In other words, results from this sample indicate that detection of fabricated data using the distribution of statistically nonsignificant p -values to detect excessive amounts of high p -values does not seem promising.

Variance analysis

We expected the dispersion of variances to be lower in fabricated- as opposed to genuine data. We computed the AUROC values for the variance analyses with the directional hypothesis that genuine data would show 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, once for each dispersion of variance measure. Of these 14, 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 homogeneous population variances (<https://osf.io/tshx8/>), which are the results reported in the second row of Table 3. All other variance analyses have not been preregistered and should therefore be considered exploratory.

Table 3: Area Under Receiving Operator Characteristic (AUROC) values for each variance analysis and operationalization, including its 95 percent Confidence Interval. Heterogeneity assumes population variances differ for the low- and high anchoring conditions, whereas homogeneity assumes equal population variances across anchoring conditions. 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]	0.527 [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]

Our preregistered analysis indicates that variance analyses do not perform above chance level when the assumption of homogeneous population variances is violated. More specifically, for the dispersion of variance measure based on the standard deviation of the variances (i.e., SD_z), performance is below chance levels, $AUROC = 0.264$, 95% CI [0.235-0.293]; for the dispersion of variance measure based on the range of the variances (i.e., $max_z - min_z$) performance is around chance level, $AUROC = 0.544$, 95% CI [0.507-0.58]. This result also 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.

Our exploratory results suggest that (1) taking into account heterogeneous population variances improves the performance of the variance analyses and (2) that the dispersion of variances measured by the range of variances is consistently more robust to violations of homogeneous population variances than the standard deviation of variances. Compared to the (below) chance level performance of the preregistered variance analyses, the variance analyses that take into account heterogeneous population variances perform much better regardless of the dispersion of variance measure used. More specifically, $AUROC = 0.761$, 95% CI [0.733-0.788] for the standard deviation of the variances (i.e., SD_z) and $AUROC = 0.827$, 95% CI [0.8-0.853] for the range of the variances (i.e., $max_z - min_z$). For the analyses assuming homogeneous population variances per study (i.e., rows 3-6 of Table 3), we see that $max_z - min_z$ is consistently more

robust at detecting data fabrication when compared to SD_z . Lastly, we see that the AUROC results for variance analyses separated per study or anchoring condition within a study are quite variable (ranging from 0.373-0.798), which suggests that a combined analysis of variances across homogeneous subsets of standard deviations is preferred.

Combining p -value- and variance analyses

Results presented in Table 4 indicate that the combinations of the p -value analyses and variance analyses performs poorly in detecting fabricated data in our study. The p -value analyses of the gender- and interaction effects already performed at chance level, and the variance analyses performed reasonably poor for all but the combined method with subgroups. As such, the combinations would not be expected to work well in detecting data fabrication because little to no evidential value is added by the reversed Fisher method to the evidential value of the variance analyses.

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

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

Effect sizes

Using the statistically significant effect sizes from the anchoring studies, we are able to differentiate between the fabricated- and genuine results fairly well. Figure 4 (middle column, second row) indicates that the fabricated statistically significant effects are considerably different. If we inspect the effect size distributions (r), we see that the median fabricated effect size is 0.891 whereas the median genuine effect size is 0.661 (median difference 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, we see that the $AUROC$ for the studies approximate .75 each (0.743, 95% CI [0.712-0.774]; 0.734, 95% CI [0.702-0.767]; 0.737, 95% CI [0.706-0.768]; 0.755, 95% CI [0.724-0.786]; respectively). Figure 5 depicts the density distributions of the genuine- and fabricated effect sizes per study, which shows the extent to which the density of the fabricated effect sizes exceeds the maximum of the genuine effect sizes. In other words, results indicate that given a randomly drawn genuine- and fabricated anchoring effect size, there is approximately a 75% chance that the larger effect size is the fabricated one in this sample. Based on these results, it seems that using extreme effect sizes to detect data fabrication is a parsimonious and fairly effective method.⁸

⁸Kunnen wel meer, ik heb plaatje toegevoegd waar de distributies per studie uit elkaar worden gezet. Hier kun je zo zien hoeveel density van de fabricated boven de maximum van de genuine ligt

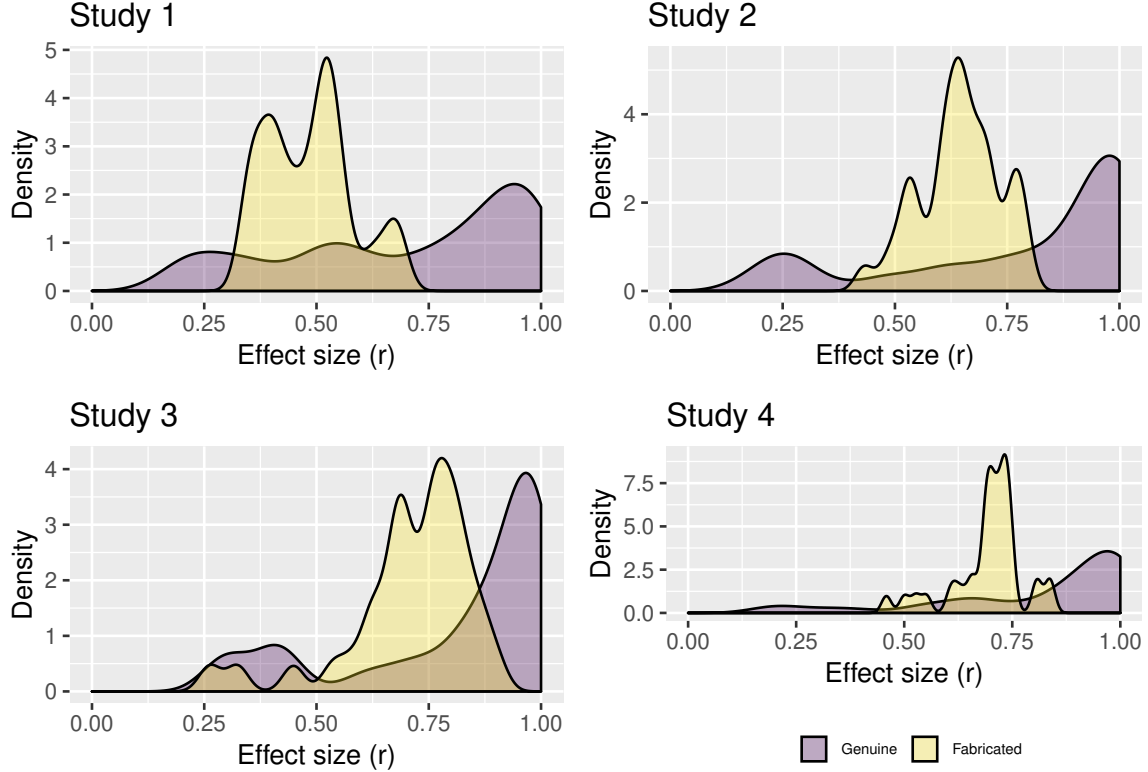


Figure 5: Density distributions of genuine- and fabricated anchoring effect sizes for each of the four anchoring studies.

Fabricating effects with Random Number Generators (RNGs)

Fabricated effects might seem more genuine when participants used Random Number Generators (RNGs). The analyses presented next are not preregistered. RNGs are typically used in computer-based simulation procedures where data is generated that are supposed to arise from probabilistic processes. Given that our framework of detecting data fabrication rests on the lack of intuitive understanding of humans at drawing values from probability distributions, those participants who used an RNG might come closer to fabricating seemingly genuine data. Hence, those data might be harder to detect.

We split 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 5, except that this time the density distributions of the fabricated data are split between these two groups.

Based on Figure 6 we conclude that using RNGs results in less exaggerated summary statistics, but still larger than genuine ones. Furthermore, it seems that the use of RNGs produced somewhat more uniformly distributed statistically nonsignificant p -values than those without RNGs, but that difference is not confirmed by the AUROC values (gender, with RNG $AUROC = 0.455$ 95% CI [0.405-0.504], without RNG $AUROC = 0.52$ 95% CI [0.482-0.557]; interaction, with RNG $AUROC = 0.601$ 95% CI [0.558-0.644], without RNG $AUROC = 0.482$ 95% CI [0.444-0.52]). For the best performing variance analysis (i.e., heterogeneity over all four anchoring studies with $max_z - min_z$ operationalization) classification performance is barely different between those data fabricated with ($AUROC = 0.78$ 95% CI [0.728-0.833]) or without RNGs ($AUROC = 0.845$ 95% CI [0.817-0.874]). For effect sizes, Table 5 specifies the differences in sample estimates of the AUROC between the groups of fabricated results with and without RNGs (as compared to the genuine data). These results indicate that the fabricated data from participants who used RNGs are relatively more difficult to detect (mean probability of 0.604 that the larger effect is fabricated if presented with one genuine and fabricated effect size), compared to data from participants who did not use a RNG (mean probability of 0.797 that the larger effect is fabricated if presented with one genuine and fabricated effect size; see also Table 5). Based on these results, it seems that only effect

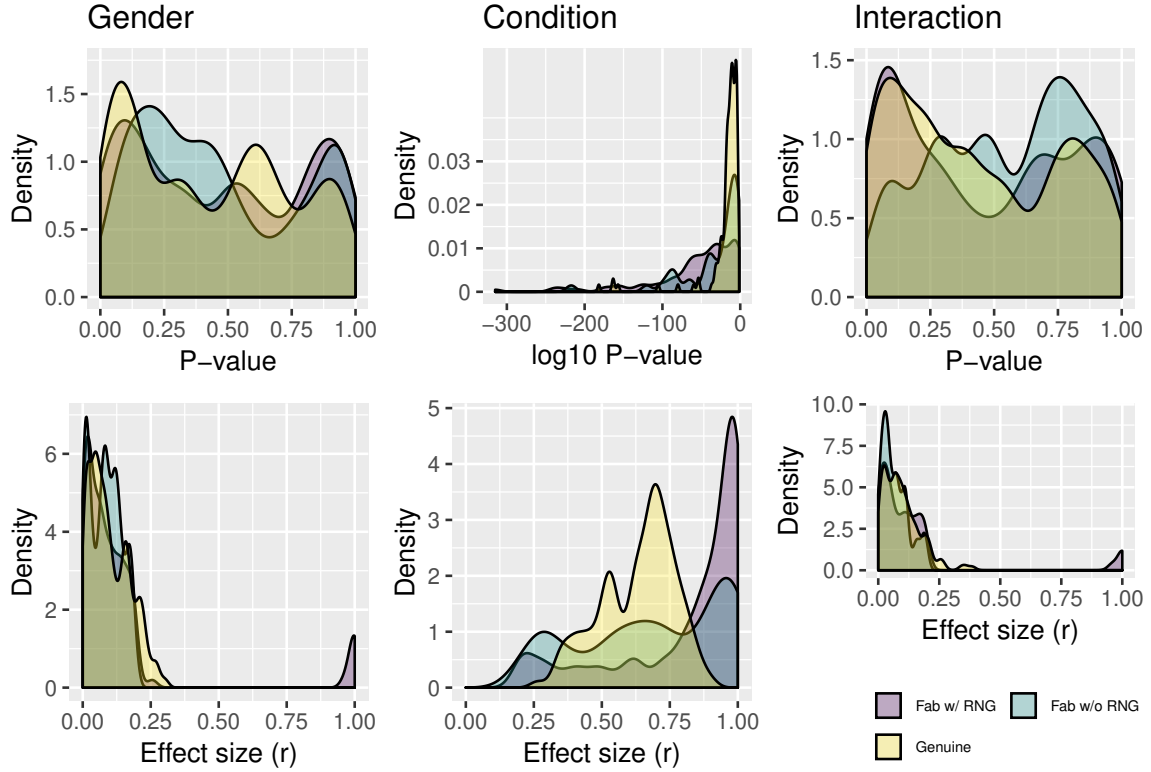


Figure 6: Density distributions of p-values and effect sizes for the gender effect, the anchoring effect, and the interaction effect across the four anchoring studies. This is a reproduction of Figure 4, except that each panel now separates the density distributions for fabricated results using a random 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 5: AUROC values for detecting data fabrication based on effect sizes for those participants who used Random Number Generators (RNGs) and those participants who did not use RNGs, including 95 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]

sizes become less effective at detecting fabricated data, but note that we did not preregister the analyses in this section.

Discussion

We presented the first controlled study on detecting data fabrication at the level of the individual data set using summary statistics. As far as we could find, previous efforts only looked at group-level comparisons of genuine- and fabricated data (Akhtar-Danesh & Dehghan-Kooshkghazi, 2003), inspected properties of individually fabricated sets of data without comparing them to genuine data, or did not contextualize these data in a realistic study with specific hypotheses (Mosimann et al., 1995). We explicitly asked researchers to fabricate results for an effect within their research domain (i.e., the anchoring effect), which was contextualized in realistic hypotheses, and was compared to genuine data on the same effect. We investigated the performance of using the reversed Fisher method, variance analyses, combinations of these two methods, and statistically significant effect sizes to detect fabricated data.

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

Using a Random Number Generator (RNG) to fabricate summary statistics could decrease the probability of detecting a fabricated dataset, depending on the type of analysis. Although we did not preregister these analyses, results suggest that using RNGs substantially decreases the performance of using effect sizes to classify fabricated- from genuine data. This indicates that data fabricated by humans without RNGs might be excessively bold. However, it also showcases that methods to detect data fabrication with effect sizes may potentially fail when RNGs are involved. On the other hand, using RNGs did not substantially decrease the performance of the variance analysis that analyzed the anchoring conditions. 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.

For the reversed Fisher method, results indicated that participants did not fabricate excessive amounts of high p -values when told to fabricate statistically nonsignificant effects. More specifically, the analysis of nonsignificant p -values appeared to perform at chance level, going 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 high p -values.

We noted that the assumption of homogeneous population variances in the variance analyses had not previously been explicated nor tested for robustness to violations. In Simonsohn (2013) it remains implicit that the variances grouped together in an analysis should arise from a homogeneous population distribution. Our results indicate that the classification performance of variance analyses strongly depends dependent on fulfilling this assumption. The alternative operationalization we included inspected the range of standard deviations ($max_z - min_z$) instead of the standard deviation of variances (SD_z). Our alternative approach seemed to be more robust to violations of the homogeneity assumption, but was not preregistered and should 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 here) and use the range of standard deviations ($max_z - min_z$), when variance analyses are applied.

We note that the 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.⁹ Second, mental fabrication strategies may be dependent on the type of effect or measurement that is being fabricated. In the anchoring studies, data needed to be fabricated for numbers that are 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, our results might be better at detecting data fabrication because of this increased lack of intuitiveness. Other kinds of studies that are easier for fabricators to think about in terms of fabricating realistic data might prove more difficult to classify. For example, we might question how results based on Likert scale items might show different kinds of results from these anchoring studies.

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

5 Study 2 - detecting fabricated raw data

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

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

Methods

Data collection

We collected twenty-one genuine data sets on the Stroop task from the Many Labs 3 project (<https://osf.io/n8xa7/>; Ebersole et al., 2016). Many Labs 3 (ML3) includes 20 participant pools from universities and one online sample (the original preregistration mentioned 20 data sets, accidentally overlooking the online sample; Hartgerink et al., 2016b). Similar to Study 1, we assumed these data to be genuine due to the minimal individual gains for fabricating data and the transparency of the project. Using the original raw data and analysis script from ML3 (<https://osf.io/qs8tp/>), we computed the mean (M) and standard deviation (SD) for each participant their response latencies in both within-subjects conditions

⁹Marcel, we hebben geen informatie over hoeveel mensen dat hier hebben gedaan; dat is eerder iets wat in de transcripten van Studie 2 staat.