Ecological performance of detecting data fabrication

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

24 June, 2017

Contents

Theoretical framework **3**
Detecting data fabrication in summary results . . . 3
Detecting data fabrication in raw data . . . 5

Study 1 **7**
Methods . . . 7
Results . . . 9

Study 2 **12**
Methods . . . 13
Results . . . 13
Discussion . . . 13

General discussion **13**
`ddfab` package . . . 13
Study fallout . . . 15

Session info **15**

References **15**

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

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

In the last decade, detecting image manipulation has become one of the few forms of detecting scientific misconduct (alongside plagiarism). The Journal of Cell Biology scans each submitted image for potential manipulation (The Journal of Cell Biology, 2015), which greatly increases the risk of detecting (blatant) image manipulation. More recently, algorithms have been developed to automate the scanning of images for (subtle) manipulations (Koppers, Wormer, & Ickstadt, 2016). These developments in detecting image manipulation have increased detection risk during the pre-publication and post-publication phase by improving detection mechanisms and increasing the understanding of how images might be manipulated. Moreover,

their application also helps researchers systematically evaluate research articles to estimate the extent of the problem of image manipulation (4% of all papers are estimated to contain manipulated images, Bik, Casadevall, & Fang, 2016).

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

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

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

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

Theoretical framework

Throughout this paper, we inspect statistical methods to detect potential data fabrication that can be applied to (1) summary results or (2) raw data. Even though structure and contents of data can look different depending on the structure of a study and the measures, there are certain common characteristics of empirical results and the underlying raw data that can be inspected. For example, summary results frequently include means, standard deviations, test-statistics, and p-values. Raw data frequently contain at least some variables measured at a interval- or ratio scale (Stevens, 1946). Such common characteristics allow for the development of generic statistical methods that can be applied across a varied set of results to screen for problematic data. We review the theoretical framework of the specific methods we apply throughout this paper, but these are not exhaustive of all methods available to test for potential problems in empirical data (Anaya, 2016; Brown & Heathers, 2016; see also Buyse et al., 1999; James Heathers, 2017).

Detecting data fabrication in summary results

P-value analysis

The distribution of one p-value depends on various parameters, but there are specific theoretical boundary conditions to this distribution (Fisher, 1925). Based on the population effect size, the precision of the estimate, and the observed effect size, the p-value distribution can be uniform (when then null hypothesis is true) or right-skewed (when the alternative hypothesis is true). By extension, any other distribution, after taking into account sampling error, is theoretically suspect. For example, a distribution might show a bump when authors p-hack (e.g., Figure 1 in C. H. Hartgerink, Aert, Nuijten, Wicherts, & Assen, 2016), but fabricators might create other non-uniform or non-right-skewed distributions, failing to take into account the theoretical boundary conditions.

In order to test whether observed p-values transgress the boundary condition, we previously proposed an adaptation of Fisher's method (S. P. O'Brien et al., 2016). This adaptation is a simple reversal of the Fisher method (Fisher, 1925), which was originally introduced as a simple meta-analytic test to indicate presence of an effect; this test is computed as

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

where it tests for right-skew (i.e., more smaller p-values than larger p-values) across the k number of p-values. Reversing the Fisher method results in

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

where it now tests for left-skew (i.e., more larger p-values than smaller p-values) across the k number of p-values that falls above the threshold t. This threshold is added and can be set to anything. Given the frequent misinterpretation of p-values (D. G. Altman & Bland, 1995; e.g., as the probability of an effect, S. Goodman, 2008), researchers who fabricate nonsignificant data might forget to fabricate a uniform p-value distribution and generate only large p-values when fabricating nonsignificant results. In that case, it makes sense to set $t = .05$ or $t = .10$, in order to include only nonsignificant test results. Upon writing this paper, it became clear to us that this is similar to the operating principle of Carlisle's method testing for excessive homogeneity across baseline measurements in Randomized Clinical Trials (Carlisle, 2012, RCTs; 2017; Carlisle et al., 2015).

Despite this test being useful for detecting data anomalies in nonsignificant p-values, two clear exceptions should be taken into account: (1) model-fit tests and (2) wrongly specified one-tailed tests. When model-fit tests are used, these can quickly result in high p-values because of model saturation (i.e., where the null hypothesis is good fit and the alternative is bad fit). For properly specified one-tailed tests, the p-value distribution is right-skewed. When wrongly specified, this distribution is reversed and becomes left-skew.

These two exceptional cases should not be forgotten during the application of this test, but are irrelevant for the rest of this paper due to our study design (see methods of both studies).

Variance analysis

Variance- or standard deviation estimates are typically reported to indicate spread, but just like the mean there should be sampling error in this estimate proportionate to the sample size (i.e., the variance of the variances). A variance estimate follows a χ^2-distribution, which is dependent on the sample size (p. 445; Hogg & Tanis, 2001); that is

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

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

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

where s_j^2 is the variance in the jth group. If simulating (see next paragraph), the simulated variances are used to compute MS_w.

Assuming that the observed variances are from the same population distribution, the expected spread of reported standardized variances can be simulated. By repeatedly drawing (standardized) variance estimates for j groups and then computing their spread allows for approximation of the sampling distribution. Spread can be operationalized in various ways, such as the standard deviation of the variances (denoted in this paper as SD_z) or as the range of the variances (denoted as $max - min_z$). If there is reason to believe that the observed variances come from heterogeneous populations, subgrouping (i.e., one variance analysis per subgroup) is necessary to prevent false results.

Subsequently, the observed sampling variance (or range) can be compared to the expected sampling distribution. Too consistent results would indicate potential anomalies in the reported data (Simonsohn, 2013). For example, if four independent samples all yield the variance 2.22, this could be considered excessively consistent when the probability that this amount of consistency (or more) is less than 1 out of 1000 in truly random samples.

Effect size analysis

A fabricator might present unrealistically large effects, forgetting the exponential relation between the effect size measure and the implied correlation. From our own experience, and anecdotal evidence elsewhere (Bailey, 1991), large effects have previously raised initial suspicions. The size of a large effect can quickly become unrealistic; $d = 1.2$ already implies a correlation of 0.51 between the dependent- and independent measures. More generally, the relation between the implied correlation and the effect size tends to be exponential (see Figure 1) depicts.

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

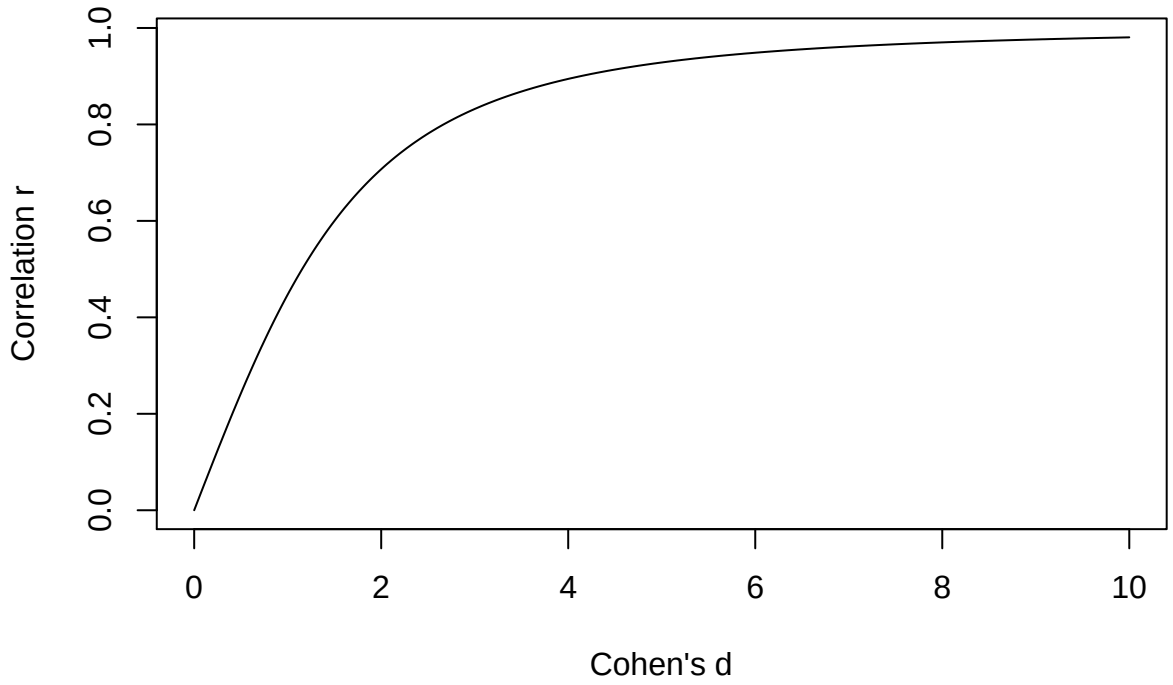

Figure 1: The relation between Cohen's d effect size and its direct transformation, the correlation r (or vice versa). If effects d>2 are fabricated, the implied correlation between the dependent- and independent measure(s) is approximately .6.

Detecting data fabrication in raw data

Digit analysis

Raw data can contain ratio- or interval scale measures that, under specific conditions, are subject to mathematical properties. These properties pertain either to the leading (first) digit (e.g., the 1 in 123.45) or the terminal (last) digit (e.g., the 5 in 123.45). By analyzing these leading- and terminal digits for deviations from such mathematical properties, it might be possible to screen for problematic data. These properties can be extended to sequences of digits, but in this article we focus on leading digit analysis (i.e., Benford's law) and terminal digit analysis to detect potentially problematic data.

Newcomb-Benford law

The Newcomb-Benford law (Benford, 1938; NBL; Newcomb, 1881) states that leading digits do not have an equal probability of occuring in various cases. A leading digit is the left-most digit of a numeric value, where a digit is any of the nine natural numbers $(1, 2, 3, ..., 9)$. The distribution of the leading digit, according to the NBL is

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

where d is the natural number of the leading digit and $P(d)$ is the probability of d occurring. This law has been empirically observed in a variety of cases, ranging from population data to the number of rivers in an area (Benford, 1938). Models that delineate what kinds of measures do- and do not adhere to NBL are non-trivial to compose.

For ratio scale measures that are scale and base invariant, the NBL often applies empirically but for measures other than those it is non-evident whether they follow the NBL [Hill (1995);@ 10.1214/11-PS175]. Scale- and base invariance indicate that the results of the measure do not change if the measure is multiplied by

constant c or transformed into a log scale of variable base. However, by extension, deviations from a ratio scale that is base- and scale invariant are likely to cause the NBL to not be applicable. For example, if a ratio scale is truncated (either with a minimum or maximum) it is already sufficient to cause a violation of the NBL (Nigrini, 2015). Moreover, a base- and scale invariant ratio measure implies adherence to the NBL (**???**), but does not necessitate that the leading digits follow the NBL. As such, empirical support that the numbers under investigation, in fact, follow the NBL is of importance before using the NBL to detect potential problems.

Despite the limitations of the NBL, this property has been applied to detect financial fraud (e.g., Cho & Gaines, 2007) or voting fraud (e.g., Durtschi, Hillison, & Pacini, 2004) and also to detect problems in scientific data. Previously, the NBL has been tested on 20 falsified anesthesia papers, detecting 18 as problematic (Hein, Zobrist, Konrad, & Schuepfer, 2012), or to test coefficients reported in economics journals (**???**). The NBL is typically used to compute the expected values, which are then compared to the observed values using a χ^2-test. However, the outcomes have not been validated to test the performance of the method.

The performance of the NBL to detect problematic data is not evidential. Based on the NBL, Obama's 2008 election was rigged (Deckert, Myagkov, & Ordeshook, 2010). Financial data is likely to be ridden with digit preferences, for example due to price setting, causing false flags. Additionally, scientific data do not necessarily follow the NBL, especially in the social sciences, because data are relatively sparse and often the result of human thought, which has previously been addressed as a reason why digits do not follow the NBL (Durtschi et al., 2004). Moreover, people might even be sensitive to fabricating data that are in line with the NBL (**???**), potentially depending on whether fabricators generate or select values (Burns, 2009).

Terminal digit analysis

Terminal digit analysis is based on the principle that the rightmost digit contains most of the measurement error in the number (**???**, **???**). For example, when someone's height is measured in micrometers, measurement one might be 1.848 metres, and a second measurement 1.841, etc. If the true height is 1.845 metres, it is a logical consequent that the terminal digit is more likely to be adjusted by measurement error (i.e., someone is unlikely to be two or zero metres tall due to measurement error). This is in essence a consequent of classical test theory, where each measurement is thought of as measuring a true score and random error (**???** (66)90002-2).

As such, the rightmost digit can be expected to be uniformly distributed if sufficient precision is provided (**???**). For our purposes, sufficient precision is determined as the terminal digit being at least the third leading digit. As such, if height measurements are reported in 1.8 metres, there are only three leading digits and there is insufficient precision. However, "1.84" contains three leading digits and is regarded as sufficient. Note also that this implies that Likert scales are wholly inappropriate for terminal digit analysis.

Terminal digit analysis is conducted with a χ^2-test on the digit occurrence counts. As such, it compares the observed frequencies with the expected uniform frequencies. Before applying this method is important to consider the applicability of the principle that the last digit contains the most measurement error (**???**) because it is central to the analysis. For example, if it is reasonable to expect there are digit preferences in genuine data, this method is unlikely to differentiate between potentially problematic and genuine data.

Multivariate associations

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

The multivariate associations between various variables can be estimated from control data that are (assumably) genuine. For example, the multivariate relation between the means (Ms) and standard deviations (SDs) can be collected from comparable studies and measures in the literature. It is important to collect information from homogeneous studies and measures, to limit the influence of confounders that either suppress or

exacerbate the relation. If the study under investigation for example uses a Hot Sauce measure for aggression [(**???**)::AID-AB2>3.0.CO;2-1], it is important to collect data on that Hot Sauce measure's properties in other studies that use it, not in studies that use various measures for aggression.

Subsequently, the (non-)parametric estimates of the multivariate relations can be used to determine how extreme the observed multivariate relations are. Consider the following example, with parametric estimates of the multivariate association between Ms and SDs. For example, imagine a set of studies that use the Hot Sauce aggression measure. When meta-analyzed, the Fisher transformed correlation (**???**) between the Ms and SDs across studies is estimated to be 0.123, with a standard deviation of 0.1. When an investigated paper presents results from a Hot Sauce aggression measure that correlate highly (Fisher transformed: .5), this indicates there might be an anomaly, considering only 8.1623774×10^{-5} of studies are expected to show such a strong relation (or more extreme) between Ms and SDs if the control data are representative of the population effect.

Contrary to the aforementioned methods, this method is less generic because it requires the investigator to collect control data. However, because of this, it also has the potential to be more sensitive then the generic methods. Additionally, multivariate associations are always present in data and therefore the applicability of this method would be greater than the generic analyses that have a specific set of conditions that need to be fulfilled before the methods are applicable.

Study 1

We tested the performance of statistical methods to detect data fabrication in summary results with genuine- and fabricated summary results from four anchoring studies (Jacowitz & Kahneman, 1995; A. Tversky & Kahneman, 1974). The anchoring effect is a well-known psychological heuristic that uses the information in the question as the starting point for the answer, which is then adjusted to yield a final estimate of a quantity. For example 'Is the percentage of African countries in the United Nations more or less than [10% or 65%]?'. These questions yield mean responses of 25% and 45%, respectively (A. Tversky & Kahneman, 1974), despite essentially posing the same factual question. A considerable amount of genuine datasets on this heuristic are freely available and we collected fabricated datasets within this study. This study was approved by the Tilburg Ethical Review Board (EC-2015.50).

Methods

We collected summary results for four anchoring studies: (i) distance from San Francisco to New York, (ii) population of Chicago, (iii) height of the Mount Everest, and (iv) the number of babies born per day in the United States (Jacowitz & Kahneman, 1995). Each of the four studies provided us with summary results for a 2 (low/high anchoring) $\times$ 2 (male/female) factorial design. Throughout this study, the unit of analysis is a set of summary statistics (i.e., means, standard deviations, and test results) for the four anchoring studies from one respondent. For current purposes, a respondent is defined as researcher/lab where the four anchoring studies' summary statistics originate from. All materials, data, and analyses scripts are freely available on the OSF (https://osf.io/b24pq) and a preregistration is available at https://osf.io/ejf5x (deviations are explicated in this report).

Data collection

We downloaded thirty-six genuine datasets from the publicly available Many Labs (ML) project (https://osf.io/pqf9r; Klein et al., 2014). The ML project replicated several effects across thirty-six locations, including the anchoring effect in the four studies mentioned previously. Considering the size of the ML project, the transparency of research results, and minimal individual gain for fabricating data, we assumed these data to be genuine. For each of the thirty-six locations we computed sample sizes, means, and standard deviations for each of the four conditions in the four anchoring studies (i.e., $3 \times 4 \times 4$) for each of the thirty-six locations.

We computed these summary statistics from the raw ML data, which were cleaned using the original analysis scripts from the ML project.

Using quotum sampling, we collected thirty-six fabricated datasets of summary results for the same four anchoring studies. Quotum sampling was used to sample as many responses as possible for the available 36 rewards (i.e., not all respondents might request the gift card and count towards the quotum; one participant did not request a reward). The sampling frame consisted of 2,038 psychology researchers who published a peer-reviewed paper in 2015, as indexed in Web of Science (WoS) with the filter set to the U.S. We sampled psychology researchers to improve familiarity with the anchoring effect (Jacowitz & Kahneman, 1995; A. Tversky & Kahneman, 1974), for which summary results were fabricated. We filtered for U.S. researchers to ensure familiarity with the imperial measurement system, which is the scale of some of the anchoring studies (note: we found out several non-U.S. researchers were included because the WoS filter also retained papers with co-authors from the U.S.). WoS was searched on October 13, 2015. In total, 2,038 unique corresponding e-mails were extracted from 2,014 papers (due to multiple corresponding authors).

We invited a random sample of 1,000 researchers via e-mail to participate in this study on April 25, 2016 (invitation: https://osf.io/s4w8r). The study took place via Qualtrics with anonimization procedures in place (e.g., no IP-addresses saved). We informed the participating researchers that the study would require them to fabricate data and explicitly mentioned that we would investigate these data with statistical methods to detect data fabrication. We also clarified to the respondents that they could stop at any time without providing a reason. If they wanted, respondents received a $30 Amazon gift card as compensation for their participation if they were willing to enter their email address. They could win an additional $50 Amazon gift card if they were one of three top fabricators. The provided e-mail addresses were unlinked from individual responses upon sending the bonus gift cards. The full text of the Qualtrics survey is available at https://osf.io/w984b.

Each respondent was instructed to fabricate 32 summary statistics (4 studies $\times$ 2 conditions $\times$ 2 sexes $\times$ 2 statistics [mean and sd]) that fulfilled three hypotheses. We instructed respondents to fabricate results for the following hypotheses: there is (i) a main effect of condition, (ii) no effect of sex, and (iii) no interaction effect between condition and sex. We fixed the sample sizes to 25 per cell; respondents did not need to fabricate sample sizes. The fabricated summary statistics and their accompanying test results for these three hypotheses serve as the data to examine the properties of statistical tools to detect data fabrication.

We provided respondents with a template spreadsheet to fill out the fabricated data, in order to standardize the fabrication process without restraining the participant in how they chose to fabricate data. Figure 2 depicts an example of this spreadsheet (original: https://osf.io/w6v4u). We requested respondents to fill in the yellow cells with fabricated data, which includes means and the standard deviations for four conditions. Using these values, statistical tests are computed and shown in the "Current result" column instantaneously. If these results confirmed the hypotheses, a checkmark appeared as depicted in Figure 2. We required respondents to copy-paste the yellow cells into Qualtrics, to provide a standardized response format that could be automatically processed in the analyses.

Upon completing the fabrication of the data, respondents were debriefed. Respondents answered several questions about their statistical knowledge and approach to data fabrication and finally we reminded them that data fabrication is widely condemned by professional organizations, institutions, and funding agencies alike. We rewarded participation with a $30 Amazon gift card and the fabricated results that were most difficult to detect received a bonus $50 Amazon gift card.

Data analysis

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

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 2: Example of a filled in template spreadsheet used in the fabrication process of Study 1. Respondents fabricated data in the yellow cells, which were used to compute the results of the hypothesis tests. If the fabricated data confirm the hypotheses, a checkmark appeared in a green cell (one of four template spreadsheets available at https://osf.io/w6v4u/).

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

For each of these statistical tests to detect data fabrication we carried out sensitivity and specificity analyses using Area Under Receiving Operator Characteristic (AUROC) curves. AUROC-analyses indicate the sensitivity (i.e., True Positive Rate [TPR]) and specificity (i.e., True Negative Rate [TNR]) for various decision criteria (e.g., $\alpha = 0, .01, .02, ..., .99, 1$). With these AUROC-curves, informed decisions about optimal alpha levels can be made based on various criteria. In this case, we determine the optimal alpha level by finding that alpha level for which the combination of TPR and TNR were highest. For example, if $\alpha = .04$ results in $TPR = .30$ and $TNR = .70$, but $\alpha = .05$ results in $TPR = .5$ and $TNR = .5$, .05 was chosen as an optimal decision criterion based on the sample.

AUROC values indicates the probability that a randomly drawn fabricated- and genuine dataset can be correctly classified as fabricated and genuine (Hanley & McNeil, 1982). In other words, if $AUROC = .5$, correctly classifying a randomly drawn dataset in this sample is equal to a coin flip. For this setting, we will regard any AUROC $< .6$ as plainly insufficient for detecting data fabrication, $.6 \leq$ AUROC $< .7$ as failed, $.7 \leq$ AUROC $< .8$ as sufficient, $.8 \leq$ AUROC $< .9$ as good, and $.9 \leq$ AUROC ≤ 1 as excellent.

Results

The collected data included 36 genuine data from Many Labs 1 (https://osf.io/pqf9r; Klein et al., 2014) and 39 fabricated datasets (https://osf.io/e6zys; 3 participants did not participate for a bonus).

Figure 3 shows a group-level comparison of the genuine- and fabricated p-values and effect sizes (r). These group-level comparisons provide an overview of the differences between the genuine- and fabricated data (see also Akhtar-Danesh & Dehghan-Kooshkghazi, 2003). These distributions indicate little group differences between genuine- and fabricated data when nonsignificant effects are inspected (i.e., gender and interaction hypotheses). However, there seem to be large group differences when we required subjects to fabricate significant data (i.e., condition hypothesis). Considering this, we also investigated how well effect sizes perform in detecting data fabrication (not preregistered). In the following sections, we investigate the performance of such statistical methods to detect data fabrication on an respondent-level basis.

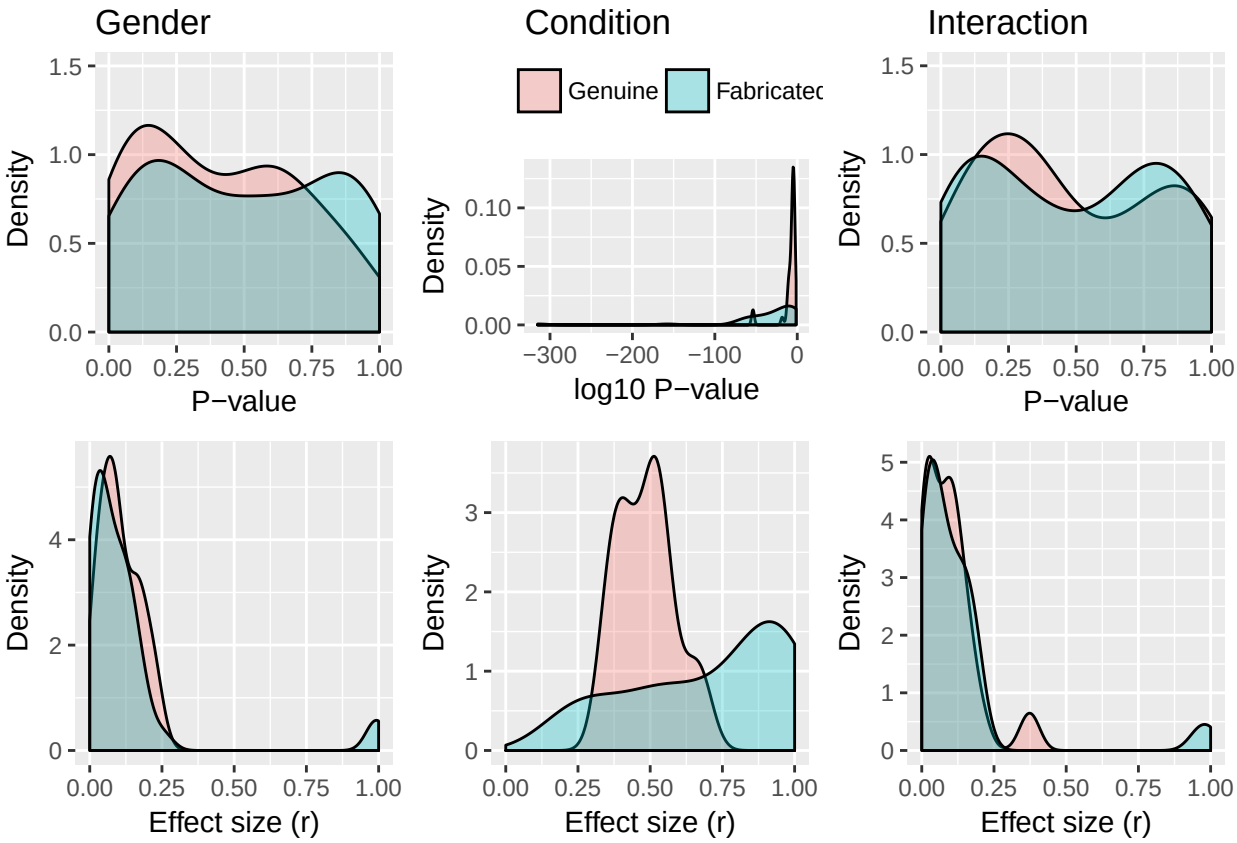

Figure 3: Overlay of density distributions for both genuine and fabricated data, per effect and type of result. We instructed respondents to fabricate nonsignificant data for the gender and interaction effects, and a significant effect for the condition effect.

Performance of variance analysis to detect data fabrication

Table 1 indicates that both operationalizations (i.e., SD_z and $max - min_z$) show similar performance based on the AUROC. All in all, their performance ranges from 0.303 through 0.796. As such, there is considerable variation for the various applications of the variance analyses.

Table 1: *Table XX.* Diagnosticity of using variance analyses to detect data fabrication, depicted with the AUROC-value.

Method	AUROC SD_z	AUROC $max - min_z$
Homogeneous, all studies combined	0.423	0.303
Homogeneous, study 1	0.367	0.374
Homogeneous, study 2	0.421	0.446
Homogeneous, study 3	0.510	0.520
Homogeneous, study 4	0.540	0.542
Heterogeneous, all studies combined	0.770	0.758
Heterogeneous study 1, low anchor condition	0.644	0.644
Heterogeneous study 1, high anchor condition	0.438	0.438
Heterogeneous study 2, low anchor condition	0.750	0.750
Heterogeneous study 2, high anchor condition	0.614	0.614
Heterogeneous study 3, low anchor condition	0.667	0.667
Heterogeneous study 3, high anchor condition	0.650	0.650
Heterogeneous study 4, low anchor condition	0.796	0.796
Heterogeneous study 4, high anchor condition	0.556	0.556

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

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

Performance of *p*-values analysis to detect data fabrication

Table 2 indicates that methods using nonsignificant *p*-values to detect data fabrication are hardly better than chance level in the current sample. We asked researchers to fabricate data for nonsignificant effect sizes, thinking they might be unable to produce uniformly distributed *p*-values. However, these results (and the density plot in Figure XX) indicate that widespread detection based on this is not promising.

Table 2: *Table XX.* Diagnosticity of using *p*-value analyses to detect data fabrication, depicted with the AUROC-value.

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

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

Table 3 indicates that combining the variance- and p-value methods provides little beyond the methods separately. The overall performance of this combination is driven by the variance analyses, given that the p-value analysis yields little more than chance classification. When combining the results from variance analyses per anchoring condition (i.e., 10 results, SD_z heterogeneous) and the p-value analyses, a minor improvement occurs over the heterogeneous variance analysis (all studies combined, $AUROC = 0.77$). However, this difference is negligible and potentially due to sampling error.

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

Method	AUROC
Combined Fisher test (3 results, SD_z homogeneous)	0.602
Combined Fisher test (3 results, SD_z heterogeneous)	0.736
Combined Fisher test (6 results, SD_z homogeneous)	0.643
Combined Fisher test (10 results, SD_z heterogeneous)	0.771

Performance of extreme effects to detect data fabrication

Table 4 indicates that using effect sizes (i.e., $1 - r$) is a simple but effective way to detect data fabrication ($AUROC = 0.744$). Compared to the variance analyses several sections ago, its performance in this sample is a bit worse (i.e., 0.744 compared to 0.77). However, it makes up for this by computational parsimony. Whereas the variance analyses require a considerable amount of effort to implement, computing the correlation and taking the inverse is a relatively simple task. Further inspecting the effect size approach to detecting data fabrication indicates that no false positives occur until $\alpha = 0.31$ (i.e., $r > 0.69$), making this the optimal alpha level based on this sample (but note the small sample).

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

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

Study 2

We investigated detecting data fabrication in raw data as an extension of Study 1. In essence, the procedure is similar: we asked actual researchers to fabricate data that they thought would go undetected. For Study 2 we included a face-to-face interview to qualitatively assess how data fabrication occurs. A preregistration of this study occurred during the seeking of funding (**???**) and during data collection (https://osf.io/fc35g).

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

Methods

Data collection

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

We collected twenty-eight faked datasets on the Stroop task experimentally in a two-stage sampling procedure. First, we invited 80 Dutch and Flemish psychology researchers who published a peer-reviewed paper on the Stroop task between 2005-2015 as available in the Thomson Reuters' Web of Science database. We selected Dutch and Flemish researchers to allow for a face-to-face interview on how the data were fabricated. We chose the period 2005-2015 to prevent a drastic decrease in the probability that the corresponding author would still be addressable via the given email. The database was searched on October 10, 2016 and 80 unique e-mails were retrieved from 90 publications. Only two of these 80 participated in the study; we subsequently implemented a second sampling stage where we collected e-mails from all PhD-candidates, teachers, and professors of psychology related departments at Dutch universities. This resulted in 1659 additional unique e-mails that we subsequently invited to participate in this study. Due to a malfunction in Qualtrics' quotum sampling, we oversampled, resulting in 28 participants instead of the originally intended 20 participants.

Each participant received instructions on the data fabrication task via Qualtrics but was allowed to fabricate data until the face-to-face interview took place. In other words, each participant could take the time they wanted/needed to fabricate the data as extensively as they liked. Each participant received downloadable instructions (original: https://osf.io/7qhy8/) and the template spreadsheet via Qualtrics (see Figure X; https://osf.io/2qrbs/). The interview was scheduled via Qualtrics with JGV, who blinded the rest of the research team from the identifying information of each participant and the date of the interview. All interviews took place between January 31 and March 3, 2017. To incentivize researchers to participate, they received 100 euros for participation; to incentivize them to fabricate (supposedly) hard to detect data they could win an additional 100 euros if they belonged to one out of three top fabricators. The contents of the interview were transcribed for further research on qualitatively assessing how researchers might fabricate experimental data.

Data analysis

Results

Discussion

General discussion

`ddfab` package

All the methods used in this paper are also implemented in the open source R package `ddfab`.

However, these results indicate that application of these methods should be informed and as a screening method.

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

Study fallout

While conducting Study 2 reported in this paper, there was considerable criticism from several parties. We

Session info

```
sessionInfo()
```

```
## R version 3.3.3 (2017-03-06)
## Platform: x86_64-redhat-linux-gnu (64-bit)
## Running under: Fedora 25 (Workstation Edition)
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] stringr_1.1.0     plyr_1.8.4        reshape2_1.4.2
##  [4] dplyr_0.5.0       data.table_1.10.0 lsr_0.5
##  [7] effects_3.1-2     car_2.0-19        httr_1.2.1
## [10] xtable_1.7-1      gridExtra_2.2.1   ggplot2_2.2.1
## [13] latex2exp_0.4.0   foreign_0.8-67    pROC_1.8
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.11     highr_0.6        nloptr_1.0.4     tools_3.3.3
##  [5] digest_0.6.8     lme4_1.1-12      evaluate_0.10    tibble_1.2
##  [9] gtable_0.2.0     nlme_3.1-131     lattice_0.20-34  Matrix_1.2-8
## [13] DBI_0.5-1        yaml_2.1.14      knitr_1.15.1     rprojroot_1.1
## [17] grid_3.3.3       nnet_7.3-12      R6_2.2.0         rmarkdown_1.3
## [21] minqa_1.2.4      magrittr_1.5     backports_1.0.4  scales_0.4.1
## [25] htmltools_0.3.5  MASS_7.3-45      splines_3.3.3    assertthat_0.1
## [29] colorspace_1.3-2 labeling_0.3     stringi_1.1.2    lazyeval_0.2.0
## [33] munsell_0.4.3
```

References

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

Altman, D. G., & Bland, J. M. (1995). Statistics notes: Absence of evidence is not evidence of absence. *BMJ*, *311*(7003), 485–485. http://doi.org/10.1136/bmj.311.7003.485

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

Baker, M. (2016). Stat-checking software stirs up psychology. *Nature*, *540*(7631), 151–152. http://doi.org/10.1038/540151a

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

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

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

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

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

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

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

Carlisle, J. B. (2017). Data fabrication and other reasons for non-random sampling in 5087 randomised, controlled trials in anaesthetic and general medical journals. *Anaesthesia*. http://doi.org/10.1111/anae.13938

Carlisle, J. B., Dexter, F., Pandit, J. J., Shafer, S. L., & Yentis, S. M. (2015). Calculating the probability of random sampling for continuous variables in submitted or published randomised controlled trials. *Anaesthesia*, *70*(7), 848–858. http://doi.org/10.1111/anae.13126

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

Deckert, J., Myagkov, M., & Ordeshook, P. C. (2010). The irrelevance of benford's law for detecting fraud in elections. Retrieved from http://wayback.archive.org/web/20170619150556/https://pdfs.semanticscholar.org/ae6b/9811d93caeda15ab8ad7060ee474cc186860.pdf

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

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

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

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

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

Medicine, *130*(12), 995. http://doi.org/10.7326/0003-4819-130-12-199906150-00008

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

Hein, J., Zobrist, R., Konrad, C., & Schuepfer, G. (2012). Scientific fraud in 20 falsified anesthesia papers. *Der Anaesthesist*, *61*(6), 543–549. http://doi.org/10.1007/s00101-012-2029-x

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

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

Hoekstra, R., Finch, S., Kiers, H. A. L., & Johnson, A. (2006). Probability as certainty: Dichotomous thinking and the misuse of p values. *Psychonomic Bulletin and Review*, *13*(6), 1033–1037. http://doi.org/10.3758/bf03213921

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

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

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

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Bernstein, M. J., … Nosek, B. A. (2014). Investigating variation in replicability. *Social Psychology*, *45*(3), 142–152. http://doi.org/10.1027/1864-9335/a000178

Koppers, L., Wormer, H., & Ickstadt, K. (2016). Towards a systematic screening tool for quality assurance and semiautomatic fraud detection for images in the life sciences. *Science and Engineering Ethics*. http://doi.org/10.1007/s11948-016-9841-7

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

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

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

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

O'Brien, S. P., Danny Chan, Leung, F., Ko, E. J., Kwak, J. S., Gwon, T., … Bouter, L. (2016). Proceedings of the 4th world conference on research integrity. *Research Integrity and Peer Review*, *1*(S1). http://doi.org/10.1186/s41073-016-0012-9

Parker, A., & Hamblen, J. (1989). Computer algorithms for plagiarism detection. *IEEE Transactions on*

Education, *32*(2), 94–99. http://doi.org/10.1109/13.28038

Post-publication criticism is crucial, but should be constructive. (2016). *Nature*, *540*(7631), 7–8. http://doi.org/10.1038/540007b

Sijtsma, K. (2015). Playing with dataOr how to discourage questionable research practices and stimulate researchers to do things right. *Psychometrika*, *81*(1), 1–15. http://doi.org/10.1007/s11336-015-9446-0

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

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

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

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

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

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. http://doi.org/10.1126/science.185.4157.1124