**5 Study 2 - detecting fabricated raw data**

In Study 2 we tested the performance of statistical methods to detect fabrication of raw data that is available at the level of individual research participants. Our procedure is comparable to that used in Study 1: We again asked actual researchers to fabricate data that they thought would go undetected. However, instead of summary statistics, in Study 2 we asked participating researchers to fabricate lower level data (i.e., raw data) and included a face-to-face interview in which we interviewed participants on how they fabricated their data (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, the current 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 the classic Stroop experiment (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 across this set 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 our 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

Figure 7: Example of a filled out template spreadsheet used in the fabrication process for Study 2. Respondents fabricated data in the yellow cells and green cells, which were used to compute the results of the hypothesis test of the condition effect. If the fabricated data confirmed the hypotheses, a checkmark appeared. This template is available at https://osf.io/2qrbs.

raw data and analysis script from ML3 (https://osf.io/qs8tp/), we computed the mean (M) and standard deviation (SD) of response latencies for each participant in both within-subjects conditions of congruent trials and incongruent trials (i.e., two M-SD combinations for each participant). This format was also the basis for the template spreadsheet that we requested fabricators to use to supply the fabricated data (see also Figure 7 or https://osf.io/2qrbs/). We calculated the Stroop effect as a t-test of the difference between the congruent and incongruent conditions (H0 : μX ̄1−X ̄2 = 0).

We collected twenty-eight fabricated data sets on the Stroop task in a two-stage sampling procedure. First, we invited 80 Dutch and Flemish psychology researchers who published a peer-reviewed paper on the Stroop task between 2005-2015 as available in the Thomson Reuters’ Web of Science database. We selected Dutch and Flemish researchers to allow for face-to-face interviews on how the data were fabricated. We chose the period 2005-2015 to prevent a decrease in the probability that the corresponding author would still be reachable via the given corresponding e-mail address. The database was searched on October 10, 2016 and 80 unique e-mails were retrieved from 90 publications. Two of these 80 researchers we contacted actually ended up participating in our study. Subsequently, we implemented a second, unplanned sampling stage where we collected e-mails from all PhD-candidates, teachers, and professors of psychology related departments at Dutch universities. This resulted in 1,659 additional unique e-mails

20

that we subsequently invited to participate in this study. Due to a malfunction in Qualtrics’ quotum sampling, we oversampled, resulting in 28 participating researchers instead of the originally intended 20 participating researchers. The second sampling scheme was not part of the original ethics review, but was considered crucial to obtain sufficiently large sample.

Each participating researcher received instructions on the data fabrication task via Qualtrics and was allowed to fabricate data until the face-to-face interview took place. In other words, each participant could take the time they wanted or needed to fabricate the data as extensively as they liked. Each participant received downloadable instructions (original: https://osf.io/7qhy8/) and the template spreadsheet via Qualtrics (see Figure 7; https://osf.io/2qrbs/). The interview was scheduled via Qualtrics with JGV, who blinded the rest of the research team from the identifying information of each participant and the date of the interview. All interviews took place between January 31 and March 3, 2017. To incentivize researchers to participate, they received 100 euros for participation; to incentivize them to fabricate (supposedly) hard to detect data they could win an additional 100 euros if they belonged to one out of three top fabricators. Participants were not informed about how we planned to detect data fabrication. JGV transcribed the contents of the interview, CHJH blind-reviewed these transcripts to remove any potentially personally identifiable information (these transcripts are freely available for anyone to use at https://doi.org/10.5281/zenodo.832490).

**Data analysis**

To detect data fabrication in raw data using statistical tools, we performed a total of sixteen analyses (preregistration: <https://osf.io/ecxvn/)> for each of the 21 genuine datasets and 28 fabricated datasets. These sixteen analyses consisted of four NBL digit analyses, four terminal digit analyses, two variance analyses, four multivariate association analyses (deviated from preregistration in that we used a parametric approach instead of the planned non-parametric approach to increase precisions), a combination test of these methods, and effect sizes at the summary statistics level (the latter test replicated Study 1 and was not preregistered). Each participating researcher fabricating data and each

contribution lab in the Many Labs study offered one data set.

To apply digit analyses to each data set, we first separated the Ms and SDs per within-subjects condition and conducted χ2-tests for each of the four statistics.10 Specficially, in each data set, we conducted digit analyses across (fabricated or actual) participants on the digits of (i) the mean response latencies in the congruent condition, (ii) the mean response latencies in the incongruent condition, (iii) the standard deviation of the response latencies in the congruent condition, and (iv) the standard deviation of the response latencies in the incongruent condition. For the NBL, we used the first (or leading) digit, whereas for the terminal digit analyses we tested the same sets but on the final digit. The sample sizes in the genuine data sets lay between X and Y, whereas the sample sizes in the fabricated datasets were fixed at 25.

For the variance analyses, we analyzed the standard deviations of the response latencies separated for the within-subjects conditions. That is, we analyzed the 25 standard deviations of the response latencies in the congruent condition for excessive consistency separately from the 25 standard deviations of the incongruent condition. We conducted this analysis for each genuine- or fabricated dataset, using the maxz − minz operationalization (not preregistered, but based on results from Study 1 indicating it is more robust to violations of the assumption of equal variances).

For the multivariate association analyses in the data sets, we analyzed four correlations between 25 pairs of fabricated statistics (both Ms and SDs) and compared this correlation to the corresponding correlations for genuine data. More specifically, across the fabricated/genuine participants, we computed a correlation between (i) the means of congruent- and incongruent conditions, (ii) standard deviations of both conditions, (iii) means and standard deviations within the congruent condition, and

(iv) means and standard deviations within the incongruent condition. We compared these four correlations per data set to the corresponding correlation for the genuine data after computing a random-effects estimate of the observed (Fisher transformed) correlations from the Many Labs 3 data. The estimated effect distribution served as the parametric model for each of those four relations under investigation (N ∼ (μ, τ )). Using the estimated parametric distribution, we computed two-tailed p-values for each fabricated- and genuine dataset.

We also combined the terminal digit analyses, the variance analyses, and the analyses based on multivariate associations using the Fisher method. More specifically, we included the p-values of 10 statistical tests; four terminal digit analyses, two variance analyses, and four analyses of the multivariate associations. We



21

excluded the NBL digit analyses because we a priori expected that psychological measures (e.g., response times) are rarely true ratio scales with sufficient range to show the NBL properties in the first digit, hence that this type of analysis would not be productive in detecting data fabrication in these types of data

(preregistration: doi.org/10.3897/rio.2.e8860).

Study 1 showed that effect sizes are a potentially valuable tool to detect data fabrication, which we exploratively replicate in Study 2. This was not preregistered because we had not yet determined results of Study 1 before designing Study 2. Based on the genuine- and fabricated data sets, we computed effect sizes for the Stroop effect based on the effect computation from the Many Labs 3 scripts (https://osf.io/qs8tp/). Using a t-test of the difference between the congruent and incongruent conditions (H0 : μ = 0) across genuine/fabricated participants, we computed the t-value and its constituent effect size as a correlation using (Hartgerink et al., 2017)

􏰄F×df1 +1 df2

where df1 = 1, F = t2, and df2 is the degrees of freedom of the t-test. We can simplify the effect size calculation to

􏰆􏰅 t2 r=􏰅 df2

􏰄t2 +1 df2

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

**Results**

**Digit analyses**

Figure 8 shows the aggregated first digit distributions of the genuine- and fabricated data side-by-side with the expected first digit distributions according to the NBL. In the first row the first digit distributions of the means are presented, for both the congruent condition (left column) and incongruent condition (right column). The first row indicates that the first digit distributions of mean response times fail to adhere to the NBL for both the genuine data and the fabricated data. The first digit distributions of the standard deviations (second row) better adhere to the NBL than the means at first glance, but still deviate substantially from what would be expected according to the NBL for both the genuine and fabricated data. These aggregate results already suggest that using the NBL to test for data fabrication is definitely not appropriate for means and probably also not appropriate for standard deviations.

The AUROC results indeed corroborate that the Newcomb-Benford Law performs close to chance level in classifying genuine- and fabricated data. More specifically, for the congruent standard deviations, using the results of the NBL test are on par with chance classification (AUROC = 0.553, 95% CI [0.389-0.717]). Values from other measures showcase that the fabricated data are actually *more* in line with the NBL than the genuine data. Consequently, the genuine data and fabricated data are often wrongly classified. This is reflected by the AUROC values that are significantly smaller than .5; congruent means, AUROC = 0.039, 95% CI [0-0.087]; incongruent means, AUROC = 0.024, 95% CI [0-0.059]; incongruent standard deviations, AUROC = 0.156, 95% CI [0.045-0.268].

Figure 9 shows the aggregated terminal digit distributions of the genuine- and fabricated data side-by-side with the expected terminal digit distributions. The first row depicts the terminal digit distributions of the means, for both the congruent (left column) and incongruent (right column) conditions. The first row shows that the terminal digit distributions of the genuine- and fabricated mean response times are approximately uniform with only minor differences between the genuine- and fabricated data. The terminal digit distributions of the standard deviations (second row) show slightly more deviation from uniformly distributed digits, but still approximate the expected distribution of terminal digits reasonably

22

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

well. Based on these aggregate digit distributions, it seems like the classification based on the terminal digit analyses will be poorly equipped to differentiate between genuine- and fabricated data .11

The AUROC results indeed show that terminal digit analyses perform close to chance level in distinguishing genuine- and fabricated data. More specifically, for the incongruent standard deviations, AUROC = 0.511, 95% CI [0.343-0.679]; congruent means, AUROC = 0.383, 95% CI [0.222- 0.543]; incongruent means, AUROC = 0.387, 95% CI [0.226-0.548]; congruent standard deviations, AUROC = 0.401, 95% CI [0.241-0.562]. The terminal digit analysis classified at most 2 of the 28 fabricated datasets as being fabricated (and 2 of the 21 genuine data as being fabricated).

**Variance analysis**

Results indicate that the fabricated- and genuine data can be perfectly separated based on results from the variance analyses. More specifically, the AUROC of both the variance analyses for the congruent standard deviations and the incongruent standard deviations is AUROC = 1 (confidence intervals cannot be reliably computed in this case). We note that these results are likely to be sample specific and do not mean to imply that this method will always be able to separate the genuine- from fabricated data perfectly. However, they also indicate that given the number of standard deviations participants had to fabricate (k = 25), it was difficult for participants to make them look similar to those found in the genuine data.

Upon closer inspection of the individual level results of the variance analyses per data set (see Appendix A), all p-values are statistically significant if compared to traditional α levels (i.e., .05; maximum 0.006 across both the genuine- and the fabricated data)

Figure 10: Density distributions of the multivariate relations (first two rows) and the effect sizes (final row), split for the genuine and fabricated data.

**Multivariate associations**

We expected that fabricated multivariate associations would be different from genuine multivariate associations. Using the parametric test of multivariate associations, results indicate classification is fair to good in the current sample. Figure 10 shows the density distributions of the various multivariate associations (rows 1-2), which already indicates the genuine data are less dispersed and more normally distributed when compared to the fabricated multivariate associations. Using the parametric estimates of the associations to test the various sets of multivariate relations between the (in)congruent means and standard deviations, AUROC values range from 0.549 through 0.842. More specifically, the AUROC for the various sets of relations (going clockwise with the first four figures in Figure 10) are AUROC = 0.818, 95% CI [0.689-0.947] for M-SD in the congruent condition, AUROC = 0.833, 95% CI [0.705-0.962] for M-SD in the incongruent condition, AUROC = 0.714, 95% CI [0.568-0.861] for M-M across conditions, AUROC = 0.549, 95% CI [0.379-0.72] for SD-SD across conditions. Overall, it seems that comparing multivariate associations to known genuine ones is a good way to detect (potential) data fabrication.

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

As preregistered, we combined both variance analyses, the terminal digit analyses, and the tests of the multivariate associations with the Fisher method (14 results in total). Results of the combined analysis perform excellent at classifying fabricated- and genuine data in this sample. More specifically, the results for the combination method indicate AUROC = 0.959 (95% CI [0.912-1]). This combination method is affected by the effectiveness of the individual methods involved; given that the performance of the multivariate associations and variance analyses ranged from sufficient to excellent, it makes sense that this combination method also performs quite well. However, the maximum p-value of the combination of these tests for either the genuine- or fabricated data is 0.003 (see also Appendix A). This indicates that all datasets would be classified as fabricated if we did not compare the results from the genuine- and

25

fabricated data, but instead used a prespecified α level (e.g., .05).

**Effect sizes**

Figure 11 (final row) shows the density distributions of the fabricated- and genuine Stroop effect sizes, which is an excellent classifier of fabricated/genuine data in this sample. More specifically, the classification performance for detecting fabricated data in this sample is AUROC = 0.981, 95% CI [0.954-1] (the 95% CI is truncated at 1). Upon closer inspection of the effect sizes, we note that only three (of 28) fabricated effect sizes fall within the range of genuine effect sizes (see Appendix A for all genuine- and fabricated effects). As such, this is a particularly good result within this sample (we did not preregister this analysis).

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

Using Random Number Generators (RNGs) in the raw data fabrication procedure did not seem to have a substantial effect on how genuine the fabricated results appeared. We explored this in our data (i.e., not preregistered) and Table 6 presents the AUROC values split on participating researchers who said they used (k = 19) or did not use RNGs (k = 9) to fabricate data (based on manual coding of the interview transcripts). Noteworthy from our exploration is that the effect size distribution seems approximately similar for both data fabricated with and without RNGs (Figure 11). Given these minor and inconsistent changes to the density distributions, we do not regard RNGs as having substantial effects

on the effectiveness of statistical methods to detect data fabrication in this sample.

Table 6: AUROC values with 95 percent confidence intervals for each test, when split for those with Random Number Generators

(RNGs) and those without.

**Discussion**

Our second study investigated how well statistical methods that use individual-level (raw) data can distinguish genuine data from fabricated. To this end, we replicated the procedure from Study 1 and asked researchers to fabricate data for individual participants for the classic Stroop task. We also collected (arguably) genuine data from the labs involved in the Many Labs study, which included the classic Stroop task. As such, we had both genuine- and fabricated data sets on the same effect.

Using these data sets we attempted to classify genuine- and fabricated individual level data using digit analyses, variance analyses, multivariate associations, and effect sizes. Results of preregistered analyses indicate that digit analyses of raw data performed at chance level, variance analyses of raw data performed excellent, and analyses of multivariate relations between variables in the raw data performed fairly to excellent. Moreover ,and not preregistered, the summary statistic effect size appeared to strike a surprisingly good balance between efficacy and parsimony for classifying fabricated- from genuine raw data (only superseded in performance by the more complex variance analyses). It is somewhat ironic that the summary statistic of the effect performs so well in classifying the genuine- from fabricated data. This replicates the finding from Study 1 that effect sizes are a valuable piece of information to discern genuine- from fabricated data. Fabricators’ use of Random Number Generators (RNGs) did not appear to have a consistent relation with classification performance with raw data.

Our results confirmed our prediction that leading digit analyses (i.e., NBL) are not fruitful in detecting fabricated response times. The Newcomb-Benford Law is frequently observed in various natural phenomena (e.g., population numbers) but Figure 6 (clearly) indicates this is not the case for summary statistics of response times. Response times are untruncated ratio measures in theory that technically satisfy the NBL’s requirements, but in practice response time measures are truncated severely (e.g., nobody can respond within <50 milliseconds and few take longer than 2000 milliseconds). If the NBL is being considered for applications to detect (potential) misconduct, there need to indications that the data generation process is in line with the requirements of the NBL, but we consider that this is hardly the

case for experimental studies in the social sciences.

Going against our predictions, participants fabricated raw data that was almost indistinguishable from the genuine raw data when looking at terminal digit analyses. Given the theoretical framework we use, wherein humans are expected to be poor at fabricating stochastic processes that underlie data collection procedures, we expected that our participants would be unable to fabricate more uniformly distributed terminal digits than terminal digit distributions arising from genuine data. Our sample indicates this is not the case. Moreover, given that these stochastic processes are expected to be better included when data is fabricated with RNGs, it was a surprise that this did not affect classification performance. This raises questions with respect to whether the framework of human’s lack of intuitive understanding of probabilities manifests itself in fabricated raw data, and if so, under which conditions.

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

We originally planned to extend Study 2 with a qualitative exploration of the fabrication process. We transcribed all 28 interviews, but due to time constraints we did not get around to conducting the qualitative analyses. We note that all transcripts are available online (without reuse restrictions; https://doi.org/10.5281/zenodo.832490) and that the initial work can be found online as well. We invite anyone with an interest to look at these documents and further build on our work.

**6 General discussion**

We presented the first two empirical studies on detecting individual sets of fabricated data, where the fabricated data pertained to existing experiments and detection occurred purely by using statistical

28

methods. By comparing results from genuine- and fabricated data across summary statistics and raw data, it seems like classification based on statistically significant effect sizes strikes the best balance between parsimony, effectiveness, and usability. Variance analyses on the other hand are a well performing option that is somewhat more complex in its application. The digit analyses based on the Newcomb-Benford law and the terminal digit principle did not perform well. We bundled our functions for the variance- and digit analyses and the (reversed) Fisher method in the ddfab (short for detecting data fabrication) package for R, which is available through GitHub (https://github.com/chartgerink/ddfab) for application in further research and development.

We designed the current studies to have sufficient information to detect data fabrication within a given set of data, but not necessarily to generalize our results to a larger population. As such, the sample sizes of the presented studies and the type of effect we chose as the empirical context necessarily restrict the drawing of more general inferences. Further research should consider whether these results also apply to other types of data or effects. Nevertheless, our studies have highlighted that variance- and effect size analysis and multivariate associations are methods that look promising to detect problematic data. Our descriptive results with confidence intervals may be regarded as an initial step in understanding the effectiveness of these methods to detect data fabrication. Next, we highlight some of the difficulties that remain.

All presented results throughout the two studies pertain to relative comparisons between genuine- and fabricated data. Hence, all statements about the performance of classification depends on the availability of unbiased genuine data to compare to and cannot readily be done by using generic decision criteria such as alpha-levels. As we saw for example in the variance analyses for Study 2, there was excellent relative classification, but absolute classification as many researchers are used to by comparing p < α remained problematic. More specifically, we would have classified all datasets as fabricated if we had used the traditional hypothesis testing approach. So we agree with the call to always include a control sample when applying these tools to studies that look suspicious ( Simonsohn, 2013). This is also the reason we refrain from formulating general decision rules for the methods presented in this paper. This might also have implications for general applications of statistical methods to detect potentially problematic data, such as the recent application by Carlisle (2017). Carlisle (2017) used the same method applied in the Fujii case to approximately 5000 clinical trials without any further validation of the methods. Our results suggest that in practice aberrant effects are best detected in relative fashion, for example in a meta-analysis (corroborating our own anecdotal experience), or to look for excessively large effect sizes (e.g., r > .95) as an initial screening of a set of effects (especially when that effect size is larger than the reliability of the product of the measures involved). Using absolute classification (i.e., p < α) can be problematic, considering that many of the methods we tested (e.g., variance analyses, digit analyses) are not specific enough, flagging both genuine- and fabricated data as problematic.

Because we included the Many Labs data (Ebersole et al., 2016; Klein et al., 2014) we had (arguably) unbiased estimates of the effects under investigation, which is key for relative comparisons. If we had used the peer-reviewed literature on the anchoring effect (Study 1) or the Stroop effect (Study 2), we would likely have found inflated effect size estimates of the anchoring- or Stroop effects due to publication bias and related biases caused by researcher degrees of freedom. These inflated effect size estimates could have resulted in worsened classification of genuine- and fabricated data because publication bias results in inflated effect sizes (M. B. Nuijten et al., 2015) and our studies indicate fabricating data has a similar effect. That publication bias and fabricating data might have similar effects in turn conflates the detection of fabricated data. Collecting an unbiased genuine effect distribution thus requires careful attention; when arguably genuine effects are collected from a literature ridden with publication bias and related biases, detection of data fabrication may be undermined. We recommend retrieving unbiased effect size distributions for an effect from large-scale replication projects, such as Registered Replication Reports (e.g., Cheung et al., 2016) and building systemic efforts to reduce publication bias

(see also Hartgerink & Zelst, 2018).

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

29

Highly relevant to the application of these kinds of methods in screening for problems in the published literature (e.g., Bik et al., 2016; Carlisle, 2017) or during peer review is that the diagnostic value of any instrument is dependent on the base rate of afflicted cases (here: fabricated data). In our study design, we built in a high prevalence of data fabrication, which directly affects the positive predictive value of these statistical methods. The positive predictive value is the chance of getting a true positive when a positive result is found. More specifically, Study 1 by design has a prevalence of 52% of data fabrication and Study 2 has a prevalence of 57%. This strongly affects the positive predictive value (PPV) of these methods if they would be applied in a more general setting. After all, even if we could classify all fabricated data correctly and falsely regard genuine data as fabricated in 5% of the cases, then with a prevalence of 2% (Fanelli, 2009) the positive predictive value would only be 29%. This is a best-case scenario that would cause approximately 1 out of 3 cases of ‘detected data fabrication’ to be false. Hence, we do not recommend attempting to detect data fabrication on statistical methods alone.

We do advise to use some of the more successful statistical methods as screening tools in review processes and as additional tools in formal misconduct investigations where prevalence is supposedly higher than in the general population of research results. We note that this should only happen in combination with evidence from other sources than statistical methods. As we mentioned before, excessively large effect sizes might be used as a screening approach for further manual or in-depth investigation, but we warn against the potential for confirmation bias that results from these earlier tests might create. As such, if any of these statistical tools are used, we recommend to solely use them to screen for indications of potential data anomalies, which are subsequently further inspected by a blinded researcher to prevent confirmation bias and using a rigorous protocol that involves due care and due process.

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

Another ethical issue is the dual use of these kinds of statistical methods to detect data fabrication. Dual use is the ethical issue where the development of knowledge can be used for both good- and evil purposes, hence, whether we should want to morally conduct this research. A traditional example is the research into biological agents that might be used for chemical warfare. Here, a data fabricator might use our research to test their fabricated data until it goes undetected based on these methods. There is no inherent way to control whether malicious actors do this and one might argue that this is sufficient reason to shy away from conducting this kind of research to begin with. However, we argue that the potential ethical uses of these methods are substantial (improved detection of fabricated data in practice) and outweigh the potential unethical uses of these methods (undermining detection by some wrongdoers). Secrecy in this respect would actually enhance the ability of malicious actors to remain undetected, because when they find a way to exploit the system fewer people can investigate suspicions they might have. Hence, we regard the ethical issue of dual use to ultimately weigh in favor of doing the research, although we recognize that this might start a competition in undermining detection of problematic data.

Some of our participants in Study 2 indicated using the Many Labs (or other open) data to fabricate their own dataset. During the interviews, some participants indicated that they thought this would make it more difficult to detect their data as fabricated. We did not investigate evidence for this claim specifically (this could be avenue for further research) but we note that our detection in Study 2 performed well despite some participants using genuine data. Moreover, we note that open data might actually facilitate the detection of fabricated data for two reasons. First, open data from preregistered projects improves the unbiased estimation of effect sizes and multivariate associations, where the peer-reviewed literature inflates estimated effect sizes

30

due to publication bias and often lacks the required information to compute these multivariate associations. As we mentioned before, having these unbiased effect size estimates seem key to detecting issues. Second, if data are fabricated based on existing data, it is more likely to be detected if it is based on open data than when based on closed data. For example, in the LaCour case data were fabricated based on open data (@ **???**; **???**). Researchers detected that this data had been fabricated because it seemed to be a(n almost) linear transformation of variables in an open dataset (Broockman, Kalla, & Aronow, 2015). As such, we see no concrete evidence to support the claim that open data could lead to worsened detection of fabricated data, but we also recognize that this does not exclude it as an option. We see the effect of open data on detection of data fabrication as a fruitful avenue for further research.

All in all, we see a need for unbiased effect size estimates to provide meaningful comparisons of genuine- and potentially fabricated data, but even when those are available the (potentially) low positive predictive value of widespread detection of data fabrication is going extremely difficult. Hence, we recommend meta-research to focus on more effective systemic reforms to counter the causes of data fabrication. One major cause is likely to be the incentive system that rewards bean-counts of outputs and does not put them in the context of a larger collective scientific effort where validity counts. Our two research studies aim the detection of a problem, but addressing of the underlying causes that give rise to data fabrication might ultimately be more sustainable and effective. Nonetheless, we also recognize that there will always be dishonesty involved for some researchers, and we recommend that research engage in more penetration testing of how those with dishonesty can fool a system.

31