Can statistical methods reliably detect fraudulent data? Examining the utility of *p*-value analyses, extreme effect sizes, GRIM, and GRIMMER

Gabriel Crone

A Thesis Submitted to the Faculty of Graduate Studies

in Partial Fulfillment of the Requirements for the Degree of Master of Arts

Graduate Program in Psychology

York University

Toronto, Ontario

May 2025

Abstract

Data fraud occurs when one creates fake data (i.e., fabrication) or alters real data (i.e., falsification), often to support a desired research hypothesis. It is detrimental to science and occurs frequently, making it a pressing concern. Fortunately, there exist several statistical tools to detect it. Extant research, however, is largely inconsistent regarding which tools work well, and no research examines how well they differentiate fraudulent articles (containing fake data) from legitimate controls. The present thesis investigated how well four popular methods to detect data fraud differentiated retracted psychology articles from legitimate controls. I included the method of extreme effect sizes, *p*-value analysis, GRIM, and GRIMMER. Extreme effect sizes performed quite well: standardized effect sizes for retracted articles were noticeably larger than controls. The other methods performed at chance levels or worse. I contend that the method of extreme effect sizes could provide valuable information during investigations of potentially fraudulent studies.

Acknowledgments

A special thanks to my supervisor, Dr. Christopher Green, for his support, humor, and encouragement throughout the thesis process. I am so grateful to have had such a wise, experienced advisor to help lend a hand, foster my academic curiosities, and push me to be the best researcher I can be. Also, a special thanks to my committee member, Dr. Raymond Mar, for his helpful and kind comments. His advice and encouragement were wonderful, and I am grateful to have had such an excellent committee member. Lastly, a thanks go out to my family, who have been—and continue to be—my rock. Thank you so much, dad, for your unwavering encouragement, mom for your kind heart (*et pour toujours m'écouté* [and for always listening to me]), Michael for always being kind to me, and thank you Nate for being the best twin brother anyone could ask for.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Research misconduct is detrimental to science and data fraud is no exception. Data fraud occurs when one creates false data (i.e., fabrication) or modifies existing data (i.e., falsification), often so that they confirm a desired research hypothesis (Steneck, 2006). In any form, data fraud has serious consequences for both scientists and the public.

Data fraud harms scientists because it degrades the quality of the research literature. When a study is based on fraudulent data, it presents spurious support for its claims. And when enough of these studies appear, they clutter and distort the research literature. Even when such studies are retracted, they are difficult to fully expunge and often "linger" in the form of post-retraction citations. This creates a facade in which effects appear to have more support than they really do. This problem is prevalent within the biomedical sciences (Hsiao & Schneider, 2021) and psychology (Fernàndez et al., 2019).

Beyond research, data fraud affects the public because it sows seeds of mistrust toward science. When new cases of data fraud emerge, they tend to get heavily publicized. As the public becomes aware of these cases, some may misconstrue them as "evidence" that science cannot be trusted. Worse, people may weaponize these cases to promote anti-scientific attitudes, beliefs, and behaviors. Since data fraud harms both scientists and the public, it is important to consider how frequently it occurs and ways to prevent it.

Data Fraud: A prevalent and persistent problem

Data fraud is often remembered through striking examples. One such example was the case of Diederik Stapel, a social psychologist who fabricated data for 55 co-authored studies and 10 PhD theses (Levelt committee et al., 2012). At the time, the magnitude of such fraud in psychology was unheard of, sending shockwaves through the psychological (and broader

scientific) research community (for a review of the case, see Bhattacharjee, 2013). Since then, there have been relatively few additional cases of similar magnitude—in fact, no case has surpassed Stapel's in terms of extremity. This might give a false sense of security, making one believe that they need not be concerned with data fraud. However, this is not the case.

Data fraud occurs frequently enough to be a concern. *Retraction Watch*, for example, is an organization that keeps a meticulous database of retracted research papers (Centre for Scientific Integrity, 2018; to access the data, see Hendricks et al., 2023). Currently, their database contains 34,347 retracted papers. Of these, roughly a quarter (8881; 25.9%) were retracted due to "Concerns/Issues About Data", and 3.53% (1223) were retracted due to "Fabricated/Falsified Data" (i.e., data fraud). Of note, retractions also appear to be occurring more frequently over time: in recent years, they have become especially common (see Figure 1 for the frequency of retracted articles over time).[1]

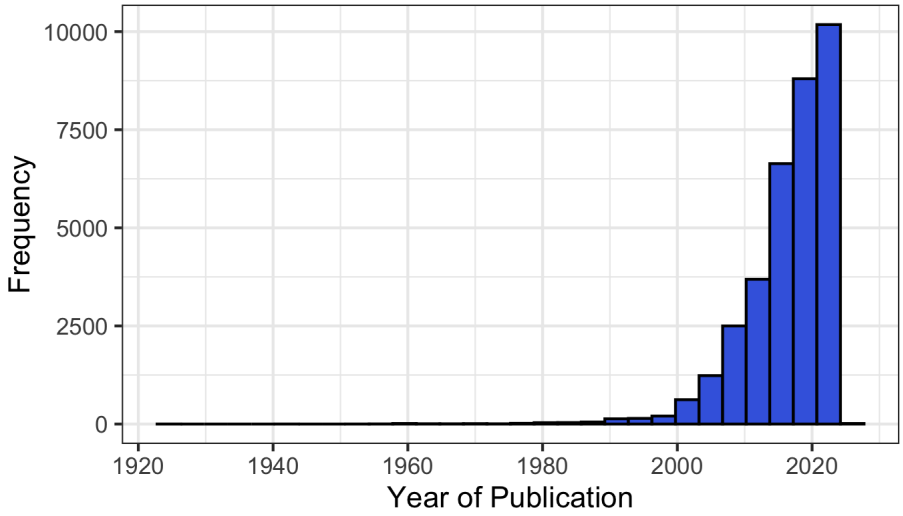
Growing research suggests that data fraud is prevalent. Within psychology, Stricker and Günther (2019) analyzed a sample of 10,000 journal articles from PSYCinfo,[2] finding that 0.82 per 10,000 (or ~0.01%) were retracted due to scientific misconduct. Although this percentage is small, given the very large number of publications in psychology, it potentially implicates thousands of studies. Xie and colleagues (2021) meta-analyzed papers that examined data fabrication and falsification. They estimated that 1.9% of journal articles used fabricated data, and 3.3% used falsified data. Based on other reports, they estimated that 12.4% of researchers have observed others fabricating data, and 10.3% have observed others falsifying data. Lastly, Fanelli (2009) also meta-analyzed studies to determine if researchers, or others they knew, had

[1] This increase may be due to more detection of newer research studies rather than higher incidences of fraud.
[2] PSYCinfo is a popular abstract database in psychology.

committed research misconduct (including data fraud). They found that 1.97% of scientists admitted that they participated in misconduct, and 14.7% admitted that they knew someone else who had participated in misconduct. In summary, although certainly a minority of cases, data fraud is prevalent enough to remain a serious issue.

Figure 1

Histogram for the Year of Publication of Retracted Research Articles



Note. Data were collected by Retraction Watch (Centre for Scientific Integrity, 2018; for access, see Hendricks et al., 2023).

Statistical Tools to Detect Data Fraud: A Brief Overview

Given the harms of data fraud and its prevalence, researchers have developed statistical tools to detect it. There are two classes of such tools: those that examine raw data, and those that examine reported statistical results (for a review, see Hartgerink et al., 2019). Each class will be examined in turn.

Detecting Data Fraud- Raw Data Tools

Two techniques exist to analyze raw data: the Newcomb-Benford Law (NBL) and the method of multivariate associations.

The Newcomb Benford Law (NBL) is a mathematical law—first-discovered by Newcomb (1881) and later re-discovered by Benford (1938)—that defines the expected distribution of digits in random, ratio-level data. One might expect that within random data, the probability of acquiring the leading digit, F_a, is uniform (equally likely for the digits 1–9). However, the NBL states that the probability, F_a, of acquiring the leading digit, a, is

$$F_a = log_{10}\left(\frac{a+1}{a}\right)$$

(Benford, 1938, p. 554). Using the law, one can compute the set of leading integer digits from 1–9, as shown in Figure 2. In short, smaller leading digits (e.g., 1, 2, 3) tend to occur more frequently than larger ones (e.g., 7, 8, 9).

Continuous, random data normally align well to the NBL (Benford, 1938). When such data do not align, however, it might indicate that the data were fabricated.[3] Indeed, one can evaluate whether their data aligns with the NBL (e.g., with *z*-tests or chi-square tests). By leveraging this fact, a multitude of studies have used the NBL to detect fraud within several applied realms, such as finance, banking, and economics.[4]

Beyond these applied uses, the NBL has also been used to test the authenticity of study data in the fields of medicine (Hüllemann et al., 2017), economics (Tödter, 2009), accounting (Horton et al., 2020), and biology (Eckhartt & Ruxton, 2023). How did these studies use the

[3] There are other reasons this occurs, such as if the data themselves are not suitable to be examined with the NBL.

[4] There have been a vast array of studies that have used the NBL in an applied context. The Benford Online Bibliography (see https://www.benfordonline.net/; Berger et al., 2024) contains an extensive list of extant citations on the NBL. As of writing (in March, 2025), it contains 2336 unique citations.

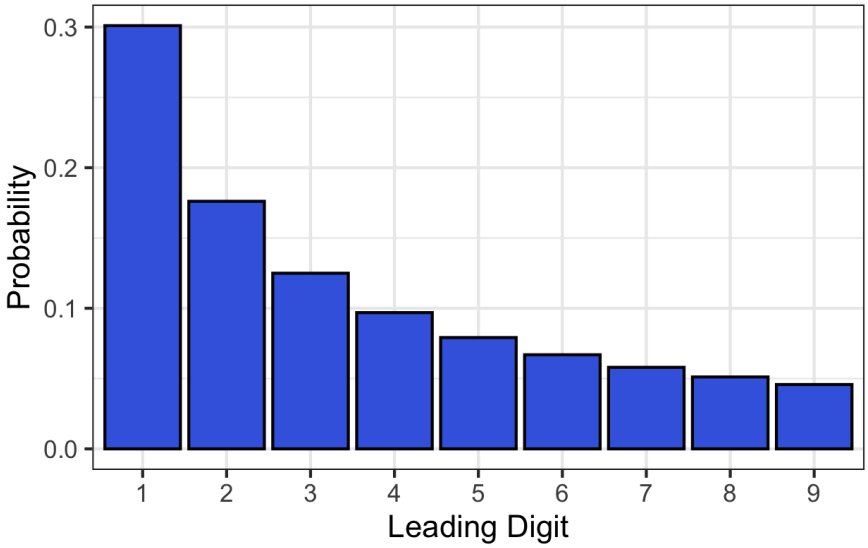
NBL to accomplish this? They did so by collecting a sample of articles with known fraudulent data, gathering a sample of legitimate comparable articles (i.e., the control articles), and examining whether the study's data conforms to the NBL. With this technique, the NBL has been used to reliably differentiate fabricated from non-fabricated studies (Eckhartt & Ruxton, 2023; Horton et al., 2020; Hüllemann et al., 2017).

Notably, the NBL necessitates strict assumptions that are seldom met in psychology: data must be true ratio-level (Berger & Hill, 2011); range from 1–100,000 (Fewster, 2009); not be overly rounded (Nagrini, 2015); not be normally, uniformly, or exponentially distributed (Berger & Hill, 2011); and must have at least 250 observations for the test to be adequately powered (Joenssen, 2014).

The second way of identifying potentially fraudulent data is the method of multivariate associations (Hartgerink et al., 2019). Within any dataset, there exist variables that tend to be correlated to some level of magnitude with one another (e.g., depression and anxiety). The logic is as follows: If individuals fabricate/falsify data, they might miss (or fail to replicate) the underlying associations between variables in the data. To employ this method for a given study, one would need to do a review of similar studies (with similar variables and ideally with a similar overall design), record the reported correlation between a variable of choice, and compile them. Then, a statistical test (e.g., a *t*-test) would be run to determine whether the observed correlation in the suspected study is statistically (i.e., statistically significantly) different from the observed correlations (Hartgerink et al., 2019). Of note, multivariate association tests require raw data, which presents a minor challenge because open data is difficult to acquire (Hardwicke et al., 2022; Houtkoop et al., 2018).

Figure 2

Distribution of Digit Probabilities Derived from the Newcomb-Benford Law



Detecting Data Fraud- Summary Data Tools

The second class of tools analyze summary data from results sections. These tools include variance analysis, *p*-value analysis, the method of extreme effect sizes, the GRIM test, and the GRIMMER test.

Variance analysis determines if a study's set of reported standard deviations (SDs) are too similar relative to simulated data (Hartgerink et al., 2019). Specifically, one would first compute a theoretical distribution of standardized variances (based on the "null" assumption that the data were collected at random), then perform a bootstrap procedure, computing the variance each time (i.e., compute the standardized "dispersion" of variances; see Simonsohn, 2013 for details). Once the bootstrapping is complete, one would have a full bootstrapped distribution of variances. If the target variance deviates from this distribution, then the study is considered anomalous. This method was primarily used by Simonsohn (2013) to demonstrate that the

reported standard deviations from two psychology researchers (Dirk Smeesters and Lawrence Sanna) were abnormally alike, indicating their studies were anomalous.

A second tool that uses summary data to detect fraud are *p*-value analyses. Such analyses determine whether the distribution of reported *p*-values in a study (with *p* ranging from 0 to .05) is anomalous (Hartgerink et al., 2019). An anomalous distribution is one that deviates from expected distribution of *p*-values. If an effect of interest exists, then the distribution of *p*-values within a given study is positively skewed. If no effect exists, the distribution of *p*-values is uniform (Simonsohn et al., 2014). However, if a *p*-value distribution is negatively skewed, it is evidence of either intense *p*-hacking (i.e., when a researcher exploits degrees of freedom to obtain a statistically significant result when none would otherwise exist; Simonsohn et al., 2014) or potential data fabrication (Ulrich & Miller, 2015).

P-value analyses test whether *p*-value distributions deviate from what one would expect under the null hypothesis of a uniform distribution. That is, it detects if a collection of *p*-values is negatively skewed, which may indicate data fabrication. Fisher (1925) originally created a meta-analytic method that aggregated *p*-values across several empirical studies and determined if their total distribution was positive (which would suggest evidence in favor of some studied effect). Hartgerink and colleagues (2019) modified Fisher's (1925) approach, creating a new statistic—the "Reversed Fisher Method" (p. 5)—that aggregates *p*-values and determines if their distribution is negatively skewed. Generally, *p*-value analyses require that tests have properly specified directionality (one- or two-tailed; Hargerink et al., 2019).

The third approach is the method of extreme effect sizes. Its logic is simple: a study is considered anomalous if its reported effect sizes are larger than those typically reported within a given research literature (Hartgerink et al., 2019). Typically, reported effect sizes are collected

from a set of studies highly similar to the study of interest. These "typical effects" comprise a distribution of their own. The effect size from the study of interest is then compared to this empirical distribution, usually via an independent samples *t*-test. If the effect size is abnormally large, it would be statistically significantly different from the mean of the empirical effect size distribution, indicating potential data fraud.

Lastly, there exist the GRIM (Brown & Heathers, 2017) and GRIMMER (Anaya, 2016) tests.[5] The granularity-related inconsistency of means (GRIM) test checks if the reported means within a study are mathematically possible, given the reported sample size and the range of the (ordinal) scale used. Given a particular sample size and ordinal scale range, there exists a limited set of mathematically plausible means. If the reported mean is not within the set of mathematically possible means (within reasonable rounding errors), it fails the GRIM test. Failure of the GRIM test may indicate that the researchers made a genuine reporting error, or that data were fabricated or falsified. This test requires that the sample size be sufficiently small to ensure a proper level of granularity (in particular, it is invalid if the per-cell sample size is larger than 100; Brown & Heathers, 2017).

A similar test to GRIM is the granularity-related inconsistency of means mapped to error repeats (GRIMMER) test. This evaluates whether the reported standard deviations (SDs), and mean/SD pairs, are mathematically possible. Its underlying logic is similar to the GRIM test, except that it runs three simultaneous tests: one on the mean (a GRIM test), one on the SD, and one on the consistency between the mean and SD (Anaya, 2016). Of note, the GRIMMER test is invalid for very small ($N < 5$) or large ($N > 99$) samples (Anaya, 2016).[6]

[5] As a note of caution, Anaya's (2016) work is a preprint and has not been peer-reviewed.

[6] Anaya (2016) mentions that if the SD < 1, the GRIMMER test can be used with sample sizes up to 200.

Statistical Misreporting

Lastly, while not a tool to detect data fraud, one related in concept to GRIM and GRIMMER is software that checks statistical misreporting. Statistical misreporting occurs when a researcher's reported *p*-values are inconsistent with the reported degrees of freedom and test statistics. Initially, it is easy to assume that such an error is rare—copying output from a screen to a word processor is a simple task—but this error occurs frequently (e.g., Bakker & Wicherts, 2011; Green et al., 2018; Nuijten et al., 2015). Although previous research checked reported statistics manually, there now exists software to automate the process (*Statcheck*; Nuijten & Epskamp, 2024).[7]

Misreported statistics are often due to general sloppiness and lack of attention paid when writing a manuscript rather than fabricating data. Yet, it could hint toward more vigilance or caution taken on behalf of a paper's authors.

How Effective are the Different Methods of Data Fabrication?

Little research has been done on the performance of the various methods, since most research focuses on one method at a time, and most only examine the NBL. There is, however, one notable exception to this.

Hartgerink and colleagues (2019) ran two empirical studies in which they investigated the efficacy of the previously discussed methods to detect data fabrication (except for GRIM and GRIMMER). They recruited psychological researchers and asked them to fabricate data based upon designs from real datasets (collected from the Many Labs 3 project; Ebersole et al., 2016). For the first study, they asked participants to fabricate summary statistics, and for the second, they asked participants to fabricate individual-level data. Within each study, they ran various

[7] *statcheck* is discussed in further depth in the Method section.

tests on both the fabricated data/results, the real data/results, and compared the accuracy of the various tools across them.

The methods that most reliably distinguished between real and fabricated data were variance analysis, the method of extreme effect sizes, and multivariate associations. In contrast, the methods that performed the worst, detecting anomalies at near chance levels, were *p*-value analysis (i.e., the "Reversed Fisher Method") and the NBL. (Remember that the GRIM and GRIMMER tests were not investigated.)

Other research is mixed regarding the utility of data fraud detection methods. Early studies suggested that the NBL was inaccurate when it came to detecting human-generated, random digits (e.g., Hsü, 1948). However, the NBL was more effective at detecting human-generated values with a substantive meaning, such as with fabricated regression coefficients (Diekmann, 2007). As discussed above, the NBL has also been used to reliably differentiate fabricated from genuine data (see the Raw Data Tools section).

The only other major application of these methods occurred when Brown and Heathers (2016) reviewed the reported means within 260 articles, finding that half contained inconsistent means. The GRIM test itself was able to, in 9 cases, indicate genuine errors. However, in terms of GRIM as a tool to detect data fabrication, no research exists. For example, no study has run GRIM (or GRIMMER) tests on articles with fabricated data to tell whether they correctly reported their means and SDs. Beyond the studies discussed, there is a genuine lack of research on the other statistical tools (e.g., *p*-value analysis, method of multivariate associations, method of extreme effect sizes).

Limitations of Existing Research

The nascent research is limited in three central ways. First, few studies have focused on these techniques and their effectiveness in detecting fraudulent data. Only one study has compared the different methods to one another in an empirical way (Hartgerink et al., 2019), and it did not include two important tools, the GRIM and GRIMMER tests. Other studies have exclusively focused on the NBL's eligibility as a tool to detect fraudulent data. Aside from these, no other studies exist, meaning that there is a need to examine these tools in more depth.

Second, no research to date has compared the various detection methods using real research articles. Hartgerink and colleagues (2019) compared the methods to one another, but did so by having participants fabricate data (and results). Their design, however, was highly contrived, and would likely generalize poorly to real cases of data fraud. Other research has tested if the NBL could differentiate between fabricated from control articles (e.g., Eckhartt & Ruxton, 2023; Horton et al., 2020; Hüllemann et al., 2017). This work had the advantage of testing methods on genuine fabricated research articles but only examined the NBL. Thus, research has yet to adopt this ecologically valid method with other data fraud detection techniques.

Third, information is lacking regarding whether retracted articles tend to contain additional inconsistencies above and beyond typical research studies (e.g., mean, SD, or mean-SD inconsistencies, as detectable by GRIM and GRIMMER tests). For example, if retracted articles tended to contain substantially more inconsistencies, then an article with such inconsistencies might be more likely to contain fabricated/falsified data.

Present Research

As discussed previously, data fraud is a serious issue, yet statistical tools to detect it tend to be under-researched. In lieu of this, the purpose of the present study is to examine the effectiveness of traditionally understudied tools to detect data fraud. This study has two central aims:

1) To test if several methods to detect data fraud—namely GRIM, GRIMMER, *p*-value analysis, and the method of extreme effect sizes—can reliably differentiate a sample of retracted psychology articles from a "control" sample of similar articles.[8]
2) To determine if retracted psychology articles (due to data fabrication/falsification) tend to contain a high degree of mathematically implausible summary statistics (i.e., inconsistent means and inconsistent SDs) and incorrectly reported inferential statistics (misreported *p*-values) relative to control articles.

Chapter 2: Method

Data Processing

The primary data consisted of all retracted psychology articles whose reason for retraction was data fabrication or falsification. These articles were located by first consulting with the Retraction Watch database (Centre for Scientific Integrity, 2018) and conducting an internet search for the manuscripts. The database was filtered such that the results only contained empirical articles from psychology journals whose reason for retraction was "Falsification/Fabrication of Data".

[8] The other methods were not included because at present, they cannot be run on a large amount of studies. For example, variance analysis would involve running over 420 individual Monte Carlo simulation studies, one per article, and the NBL would likely be inappropriate to run with only means and SDs.

For each retracted article, a small sample of "comparable" (i.e., control) research articles were located. To do so, I recorded the type of design used (i.e., experimental, observational), the variables of interest (extracted from the keywords of the study), measures used (e.g., "Remote Association Test"), the journal name, and the publication year. Then, using the Scholars Portal database, I ran a library search. The search was always restricted such that resulting articles were in the same year and journal as the retracted paper. Then, three search terms were added, two for basic design details, and a third for a keyword for the original study. Each search term was separated by Boolean *AND* operators. For example, if a study was experimental, used priming, and focused on emotions, the search query would be: "Experiment" AND "Priming" AND "Emotions". If an initial search returned more than 16 results, additional search terms were added until the number of results was less than 16.[9] When results were less than 16, each source was opened, and abstracts were read to see if the methods used aligned well with the original retracted source. If methods aligned well, they were included in the final sample, and if not, they were discarded. If more than 10 sources remained after this final filtering step, only the first 10 sources were recorded. As well, if a control article appeared more than once (e.g., appeared for retracted article A then again for Retracted article B), then it only counted toward the first retracted article (article A).

Once the total sample of articles was collected, I extracted relevant reported statistical data from each article. In particular, I extracted the following information: (1) all mean/SD pairs that reflected a mean difference and a substantive research hypothesis, (2) any means that were

[9] Early on when running initial searches, I found that when more than 15 sources were produced, many abstracts tended to be too distinct from the original study. Thus, I used 16 as the cutoff value to indicate that more search terms should be added.

GRIM testable, and (3) any mean/SD pairs that were GRIMMER testable.[10] Any data reported in an article was usable, but means/SDs for (1) were only counted if they related to a particular mean difference test that attempted to address a particular research question. In addition, for each mean/SD pair, a standardized mean difference (i.e., Cohen's *d* value) was computed, so effect sizes were on a roughly comparable metric.[11]

In addition, I inputted all of the HTML versions of the articles into the *Statcheck* R package (Nuijten & Epskamp, 2024), which automatically extracted all of the inferential statistical information. *Statcheck* works by sifting through text data and extracting any statistical tests it locates (based on standardized ways of reporting test statistics); from these, it re-computes *p*-values (based on the reported test statistics and degrees of freedom) to check if the authors correctly reported them.

This resulted in two datasets: one with means and SDs, and another with *p*-values and test statistics. Each dataset was analyzed separately.

Data Analysis

Once the summary statistical information was available and organized, the data was imported into R (R Core Team, 2024) for analysis. In R, all means and SDs that were eligible for GRIM and GRIMMER tests had said tests run on them. For *p*-value analyses, the "Reversed Fisher method", as implemented in the *ddfab* package (Hartgerink et al., 2019), was run on each study's *p*-values to determine if the distributions were highly positively skewed, and therefore of

[10] Please note that ordinal scales were required to run GRIM/GRIMMER tests (Anaya, 2016), so many were excluded due to not fitting this criterion. As well, sample sizes were seldom reported for each mean/SD pair, so they were approximated based on the experimental design, always rounding down to the nearest integer. For example, if 50 participants were divided among 4 groups, the sample size per group was approximated to be 12 (because floor(50/4) = floor(12.5) = 12).

[11] The computation was done using the formula for Cohen's *d* values, with SD_{pooled} computed assuming equal sample sizes across groups. This assumption may not have been met, but was used to simplify the data collection step. This also served to work around the fact that authors rarely reported sample size information at the level of individual bivariate statistical tests.

concern. Lastly, the method of extreme effect sizes was run on the fabricated articles by comparing the median of Cohen's *d* values across all studies and comparing it to the median Cohen's *d* value from all control studies. As well, I ran separate *t*-tests to compare each article's means to the set of means from its corresponding control articles. The *tidyverse* suite of packages was used to assist with data analysis (Wickham et al., 2019).

Chapter 3: Results

All related data and R code is accessible via the following open-science framework directory: https://osf.io/rwvqu/.

Sample of Articles

The Retraction Watch Database (2018) contained 82 psychology journal articles that were retracted due to "Falsification/Fabrication of Data". Of these, I was able to locate 77 full-text copies. Seven were excluded because they focused on topics unrelated to psychology (e.g., finance, economics), which left 70 total retracted articles in the data.[12]

Figure 3 depicts the distributions of the year these 70 articles were originally published (top-left), and the breakdown of the sub-disciplines of the journals they were published in (bottom-left). The figure suggests that most of the retracted articles were originally published between 2005 and 2015, and that most came from journals focusing on social and personality psychology (30.4%), social psychology alone (23.2%), social and experimental psychology (10.1%), or general psychology (10.1%). Thus, most subsequent results would primarily reflect the sub-disciplines of social, personality, and experimental psychology.[13]

---

[12] Only one article came from a non-psychology journal: *Meal Size, Not Body Size, Explains Errors in Estimating the Calorie Content of Meals* by Brian Wansinck (2001) was published in the *Annals of Internal Medicine*.

[13] This high proportion is not too surprising, given that Diederik Stapel, who has many retracted papers in psychology, primarily published within journals related to these sub-disciplines.

Based on these 70 retracted papers, I was able to locate a total of 366 control articles (see Method section for details). For each retracted article, I found, on average, 5.55 control articles (*Mdn* = 5.00). Though I was unable to locate any controls for three retracted articles, I decided to include them because I wanted the final sample of retracted articles to be as large as possible.

Figure 3 shows a breakdown of the year and disciplines covered by control and retracted articles. As can be seen, both the year and sub-disciplines were quite similar across these samples, thus they were considered comparable.

_p_-Value (Mis)reporting

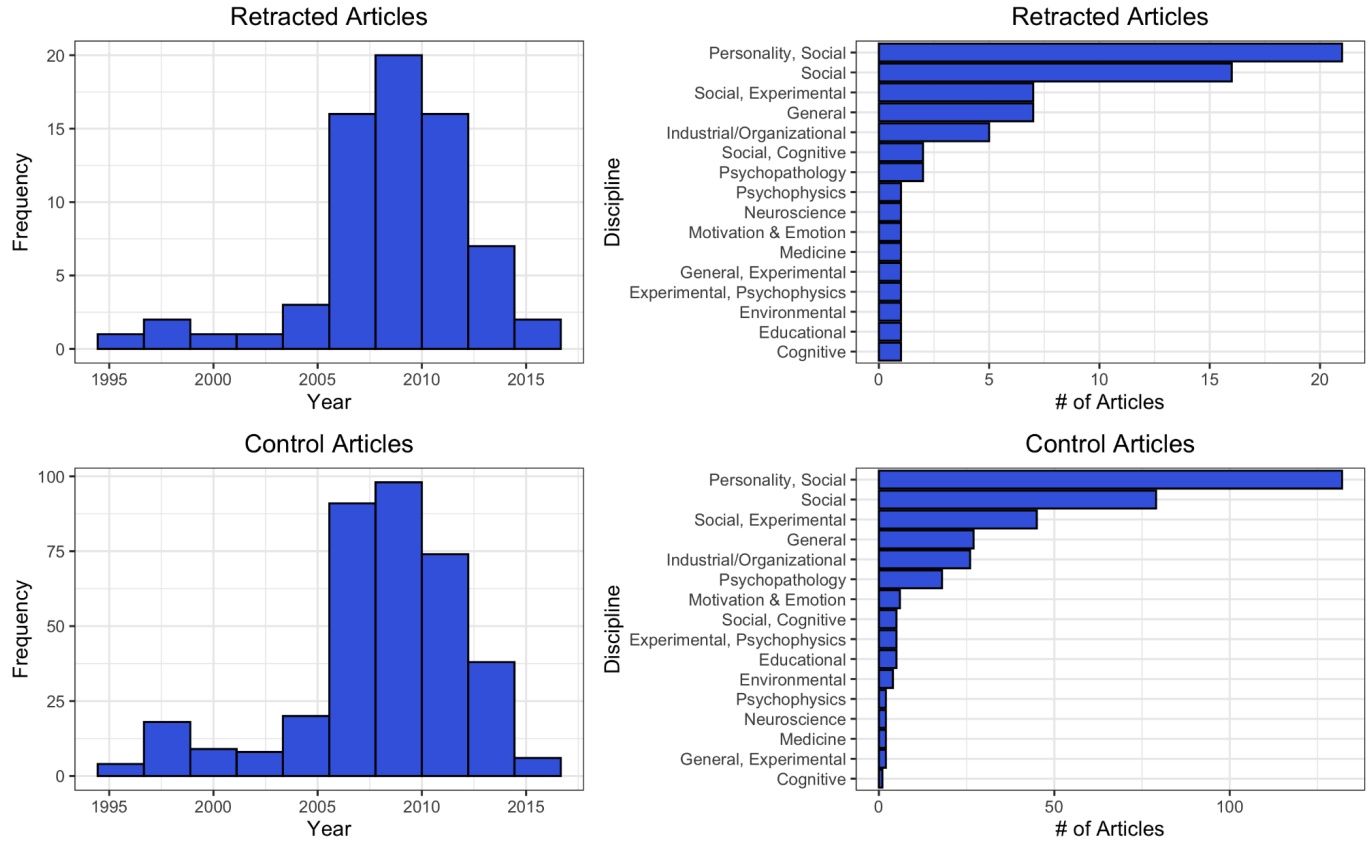
After running *Statcheck* on all eligible articles, the software was able to detect at least one statistical test within 50% of retracted articles (n = 34) and 55.2% of control articles (n = 195). In all, it detected a total of 1039 tests in the retracted sample and 5017 tests in the control sample. For each test, the software computed the correct p-value based on the test statistic and degrees of freedom, then determined if the originally reported p-value contained a *minor reporting error* (i.e., an error that would not affect the decision to reject the null hypothesis, e.g., $p = .001 \rightarrow p = .006$) or a *major reporting error* (i.e., an error that would affect the decision to reject the null hypothesis, e.g., $p = .04 \rightarrow p = .06$).

Surprisingly, *Statcheck* detected *fewer* statistical misreporting errors within the retracted sample compared to the control sample: 5.87% of statistical tests from retracted articles had a minor reporting error, whereas 9.27% of tests from control articles had such an error. Similarly, 0.87% of statistical tests from retracted articles contained a major reporting error, whereas 1.77% of tests from control articles had such an error.

At the article level, within the sample of articles *Statcheck* could detect, the same pattern held. A higher percentage of control articles had at least one minor reporting error (66.2% vs. 55.8%), and a higher percentage had at least one major reporting error (25.1% vs. 17.6%).

Figure 3

Distributions of Year and Discipline within Retracted and Control Articles



I also examined the number of reported *p*-values that were statistically significant or non-significant to see whether retracted articles tended to report non-significant *p*-values less often than control articles. This was not the case: statistical tests from control and retracted articles

contained a roughly equal proportion of non-significant *p*-values (retracted: 26.7%; control: 26.1%).

p-Value Analysis

Using the *p*-values from the *Statcheck* analysis, *p*-value analyses (i.e., "Reversed Fisher" tests; Hartgerink et al., 2019) were run at the level of individual articles (i.e., on each article's vector of re-computed *p*-values). If the test returned a statistically significant result, it would indicate an anomalous distribution of *p*-values for that article.

In all, none of the retracted articles contained anomalous *p*-value distributions, whereas 3.09% of control articles contained anomalous *p*-value distributions, suggesting that the retracted articles tended to not have detectable anomalous distributions of *p*-values.

Sample of Cohen's *d* Values, Means and SDs

From the sample of 70 retracted and 366 control articles, I manually coded all (eligible) means and SDs that were used as pairwise tests or linear contrasts between two groups. In all, I recorded 3 pieces of information: (1) pairs of means and SDs used to contrast groups, (2) directly reported Cohen's *d* values, and (3) any GRIM/ER testable means and SDs. For (1), I computed the Cohen's *d* values and re-incorporated them into the data.

Through the coding process, there were 215 articles with (computable or reported) cohen's *d* values ($n_{\text{Retracted}} = 47$; $n_{\text{Control}} = 168$), and 103 articles with GRIM-testable means or GRIMMER-testable mean/SD pairs ($n_{\text{Retracted}} = 39$; $n_{\text{Control}} = 64$). All of the other articles did not have such data available so were not used for this analysis.

From the eligible articles, there were a total of 2355 Cohen's *d* values ($n_{\text{Retracted}} = 567$; $n_{\text{Control}} = 1788$), 1427 GRIM-testable means ($n_{\text{Retracted}} = 831$; $n_{\text{Control}} = 596$), and 1167 GRIMMER-testable mean/SD pairs ($n_{\text{Retracted}} = 709$; $n_{\text{Control}} = 458$).

Method of Extreme Effect Sizes

First, I transformed all values of Cohen's *d* by taking their absolute values to ensure they were more directly comparable.[14] Figure 4 depicts a boxplot comparing the distribution of Cohen's *d* values between retracted and control articles. There were many outlying cases (i.e., those above $|d| = 2$), so medians will be used to describe central tendency instead of means. In the aggregate, the median $|d|$ was higher within the retracted sample of articles ($Mdn = 0.95$, $SD = 1.16$) than those within the control articles ($Mdn = 0.57$, $SD = 1.27$). To determine if this difference was statistically significant, I needed to account for the nesting structure of the data (since control articles were nested within respected retracted articles). Thus, I ran a random-intercept, multilevel model. Using control articles as the reference category, the model predicted a fixed mean intercept of 0.71 (95% CI: [0.38, 0.48]) and a fixed slope term of 0.35 (95% CI: [0.24, 0.46]). In other words, the model predicted that the median $|d|$ value was 0.71 within control articles, and 0.71 + 0.35 = 1.06 within retracted articles. Thus, these results confirm that effect sizes are more extreme within retracted articles than control articles.

However, the method of extreme effect sizes calls for running a statistical test to compare effect sizes (such as Cohen's *d*) within a single retracted article and its controls, not within a large aggregate sample (as I have done). Thus, I ran independent samples *t*-tests for all *d*-values within each retracted article and its respective control articles.

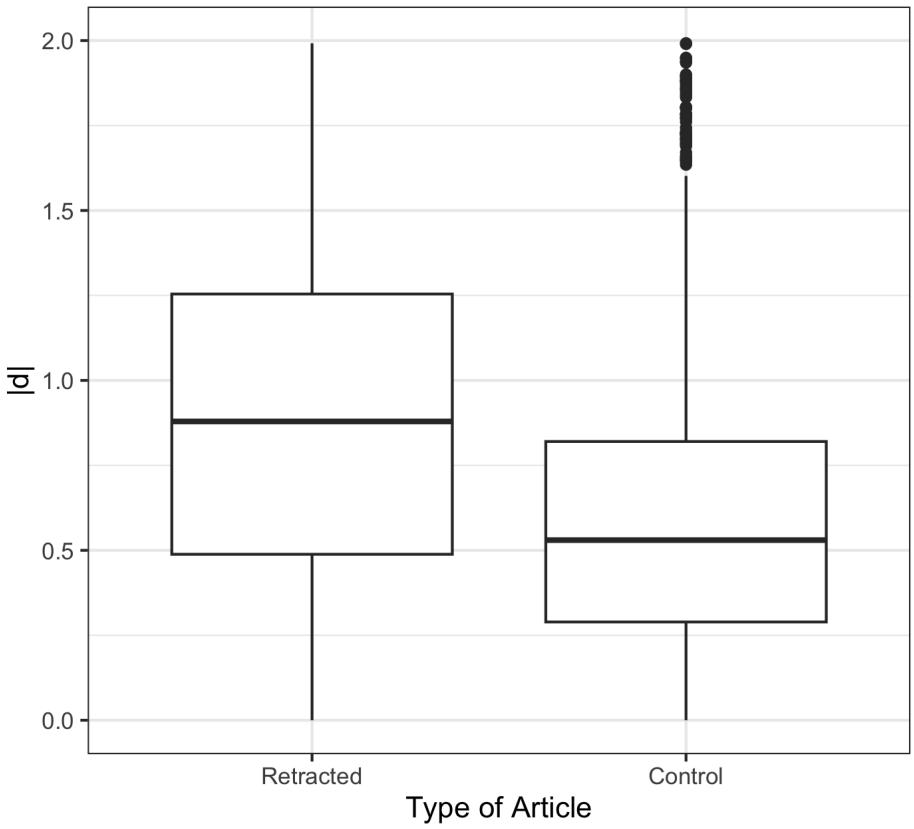
Using this approach, I was able to run 36 individual 1-tailed independent samples *t*-tests. (Many retracted articles lacked ample *d* values to run a *t*-test). Roughly half of the tests (19 or 52.8%) returned a statistically significant result, whereas the other approximate half (17 or 47.2%) did not.

---

[14] Because there is no standardized way of deciding which group to compare against which in a bivariate statistical test, the choice of groups is arbitrary, and so too is the sign of the reported *d* value.

Figure 4

Boxplot of |Cohen's d| Values within Retracted vs. Control Papers



Note. Observations above $|d| = 2.0$ were removed for easier viewability.

GRIM and GRIMMER Tests

The vast majority of means and SDs passed the GRIM & GRIMMER tests regardless of whether they came from retracted or control articles. Across all 1427 GRIM-testable means, 98.1% passed the GRIM test. This percentage did not change much across the retracted sample (98.0%) and control sample (98.5%). A similar result occurred with the GRIMMER-testable mean/SD pairs: Across all 1167 mean/SD pairs, 97.9% passed GRIMMER (Retracted: 97.6%; Control: 98.5%). Thus, it seems that retracted articles have a similarly low percentage of mathematically impossible means and SDs compared to control articles.

Chapter 4: Discussion

The purpose of this study was to determine if extant tools to detect data fabrication could reliably differentiate a sample of psychology articles containing fabricated/falsified data from similar controls. A secondary goal was to see if retracted articles (due to data fraud) contained abnormally different rates of misreported *p*-values, as well as misreported means/SDs.

To address the central goal, I tested both *p*-value analyses and the method of extreme effect sizes, finding that the latter technique far outperformed the former. The *p*-value analyses (conducted via Hartgerink et al.'s [2019] "Reversed Fisher" approach) failed to detect any abnormalities in the distribution of reported *p*-values within any retracted paper, yet it detected an abnormality in 3% of control papers. This result is consistent with Hartgerink et al.'s (2019) experimental study, which found that *p*-value analyses performed quite poorly. Given the results found there and at present, *p*-value analyses likely have little usefulness in practice.

In contrast, the method of extreme effect sizes seem more promising. Although it was not possible to run *t*-tests for all retracted papers due to missing data, this method still detected an abnormally large distribution of effects more than half of the time. Moreover, in the aggregate, there was a clear difference in the $|d|$ values such that retracted articles had a higher median $|d|$ ($Mdn_{|d|} = 0.95$) than control articles ($Mdn_{|d|} = 0.57$). This finding echoes previous research suggesting that effect sizes tend to be inflated among retracted articles compared to controls (e.g., Akhtar-Danesh & Dehghan-Kooshkghazi, 2003; Hartgerink et al., 2019).

Curiously, the median $|d|$ values in both the retracted and control samples were noticeably larger than those found in psychology. For example, Bosco and colleagues (2015) examined 147,328 correlations, finding that the median $|r|$ was .16 (equivalent to $|d| = 0.32$), with a lower bound of $|r| = 0.09$ and an upper bound of $|r| = 0.26$ (equivalent to $|d| = 0.18$ and $|d| = 0.54$,

respectively).[15] The acquired median |*d*| values therefore exceeded the highest bound of the empirically estimated values, which could indicate that the effects found at present were abnormally large across both fabricated and control samples (but especially among the retracted sample). Similarly, Gignac and Szodorai (2016) estimated the |*r*| values typically observed to inform cutoffs for interpreting effects. They classified |*r*| = 0.10 (i.e., |*d*| = 0.20) as small, |*r*| = 0.20 (i.e., |*d*| = 0.41) as typical, and |*r*| = 0.30 (i.e., |*d*| = 0.63) as large. The median |*d*| values found at present would be considered large even for the control articles, and excessively large for the retracted papers. In all, this provides further support of how truly anomalous the effect sizes within the retracted sample were.

To address the second goal, I first ran all retracted and control articles through *Statcheck* (Nuijten & Epskamp, 2024) to extract reported statistics and determine if any were misreported. Curiously, *Statcheck* found fewer reporting errors—both major and minor—within the sample of retracted articles than the control articles. This suggests that data fraudsters might tend to be hypervigilant when fabricating/falsifying their data, taking care to report their results accurately to avoid arousing suspicion. Alternatively, it might reflect the idea that fabricators tend to create customized (i.e., "designer") *p*-values that correspond with their desired hypotheses no matter what the raw data suggest. In either case, reported *p*-values with little to no statistical misreporting might act as a subtle hint toward potential fabrication.[16]

I also ran GRIM (Brown & Heathers, 2017) and GRIMMER (Anaya, 2016) tests on all (relevant) extracted means and SDs to determine how many were incorrectly reported. Most eligible means and SDs passed the GRIM and GRIMMER tests (> 95% of the time), regardless

[15] *d* values were computed using the *d_to_r()* function within the *effectsize* R package (Ben-Shachar et al., 2020).
[16] Of course, if the authors explicitly used *Statcheck* to check their own test statistics before publishing their paper, then using the misreporting rate would be misleading.

of whether they came from a retracted or control article. This indicates that the misreporting of means and SDs is rather uncommon and not a sign of data fraud. A caveat of this result, however, is that the sample size information required for these analyses were not reported in-text, so they needed to be inferred based on the overall sample size and study design. It is therefore possible that the inferred values were inaccurate, rendering this test inconclusive. In any case, this method has limited utility in real investigations, since it is difficult to run (due to both the scarce availability of eligible means/SDs, and consistent lack of reporting of mean/SD-specific sample sizes).

How useful are these tools in practice?

A useful tool would be one that is not only simple to use but also performs well at differentiating retracted articles (due to data fraud) from control articles. Based on the present findings, how useful are these tools at detecting fabricated/falsified data? Table 1 summarizes the presently reviewed tools based on two variables: ease of conducting the test, and whether it was accurate in differentiating retracted papers from controls.

Table 1

Ease of Use and Accuracy of Investigated Tools to Detect Data Fraud

Tool	Easy to conduct?	Reasonably accurate?
Method of extreme effect sizes	✓	✓
p-value analyses	✓	
GRIM		
GRIMMER		

The method of extreme effect sizes clearly emerges as the most useful tool. Not only was it the simplest to conduct, but it most clearly differentiated retracted articles from control articles. The implementation of the method, however, is heavily reliant on Null Hypothesis Significance Testing (NHST), which is problematic (e.g., because statistically significant *p*-values do not provide definitive evidence regarding the existence of an effect, and .05 is an arbitrary threshold for significance; for a review, see Nickerson, 2000). In the current study, only around half of retracted articles were "flagged" as having abnormally large effects, even though in this context the vast majority should have been flagged as potentially fraudulent. This method could be improved by focusing less on NHST and more on descriptive measures of the effect sizes reported within the target article relative to those from the control articles, or in implementing Bayesian methods to produce more meaningful probability estimates.[17]

Overall, the *p*-value analyses exhibited little usefulness. Although they were easy to conduct (with the help of *Statcheck*), they were very inaccurate, flagging only 3% of control articles and no retracted papers. To be fair, the implementation could have been limited due to using all tests flagged by *Statcheck* instead of manually recording those testing substantive research hypotheses. As well, the inputted *p*-values came from a variety of statistical tests. Ideally, in practice, *p*-values should come from identical forms of statistical tests with the same directionality (e.g., *p*-values from only one-way *t*-tests, only two-way *t*-tests, only ANOVAs, etc…) rather than the mix of tests used at present.[18] If additional *p*-values had been carefully

[17] NHST, while flawed, does provide some basic information. It suggests whether a particular data anomaly (or one more extreme) is sufficiently unlikely to occur by chance alone (assuming the null hypothesis of no anomaly existing is true). However, NHST results should be considered as merely a supplement of other more key descriptive and inferential results, such as effect sizes, confidence intervals, and Bayesian analyses, and should never be the sole focus of any test run.

[18] The designers of *p*-value analyses argue that they should not be used when directionality differs across tests because it results in far less accurate results (Hartgerink et al., 2019).

chosen and manually extracted, the results may have been different. However, such a process would not have been feasible at present given the large number of studies examined. Further, in practice, such a test would be rather time-consuming and difficult to conduct. This, in addition to previous findings of weak performance (Hartgerink et al., 2019), render the method of *p*-value analyses with little to no practical use.

Much like *p*-value analyses, the GRIM and GRIMMER tests were not useful in determining if an article used fabricated/falsified data. In particular, the vast majority of means and SDs passed both tests, regardless of their source. This is not to say that the GRIM and GRIMMER tests are not useful for their originally intended purpose: detecting inconsistently reported descriptive statistics with the goal of correcting them. Indeed, Brown and Heathers (2017) ran GRIM tests on 71 empirical articles, finding that around half (31) contained at least one inconsistent mean and more than 20% (16) contained more than one inconsistent mean. This finding resulted in many issued corrections, improving the accuracy of several high-impact psychology journal articles. It is possible, too, that the present implementation of GRIM and GRIMMER involved too many inaccurate estimations of sample sizes. Since sample sizes were missing in so many articles, I computed them based on experimental samples and context. However, it is possible that they were inaccurate, thus resulting in more false negatives. This could explain why the present study uncovered far fewer errors based on this test compared to the previous study (i.e., only 10.68% of articles (11) had at least 1 GRIM error, and 4.85% (5) had more than 1 GRIM error). Perhaps if implemented more rigorously (i.e., by only including articles that explicitly reported the sample size per experimental group), the GRIM and GRIMMER tests would become more reliable indicators. Still, they were rather difficult to implement in practice, given that many articles did not report sufficient information to run

accurate GRIM calculations, and in how much detailed reading of articles is required to capture the relevant information.

Limitations and Extensions

There were some limitations in the present study that might affect the generalizability of its findings. The first limitation pertains to the veracity with which statistical information was extracted from each article. Due to time restrictions, I was unable to manually extract statistical tests from each article, instead automatically extracting them with *Statcheck*. Thus, it is possible that *Statcheck* included many *p*-values in the final dataset that should not have been included, such as those from manipulation checks or pertaining to statistical modeling assumptions, which might explain the poor performance of these analyses. As well, though I did as careful a job as I could extracting relevant values, I was the sole coder, so it is possible that some recording errors occurred. Future work would benefit from having a second (or third) coder to verify the accuracy of the statistics. Future research could also benefit from a way to more easily—ideally, automatically—record means, SDs, and Cohen's *d* values from articles, perhaps by implementing it through a software much like *Statcheck*. Although such software might have limitations (e.g., missing means and SDs not written in a standardized way or recorded within a table), it could more quickly sift through large amounts of literature and detect relevant effect sizes. Furthermore, *r* and *d* statistics tend to be reported in a standardized way, reducing errors for these targets. Although software may help, it is no solution for the GRIM and GRIMMER tests, given the information required. To avoid missing data appropriate for these tests, a streamlined method of extracting data may involve running the PDF of each article through a large language model (e.g., ChatGPT version 4.0) trained for this purpose. Validating this method against human coders would be essential, and such an approach would pave the way to a

far faster and more reliable way of determining if GRIM and GRIMMER can be used to flag retracted articles.

Another limitation was the nature of the sample used. As mentioned in the Method section, the sample of retracted articles was highly homogeneous, composed primarily of articles within the fields of social, personality, and experimental psychology. As well, most of these articles only spanned between 2005–2015. Thus, the results may not generalize well to articles published more recently, or to articles from different psychological sub-disciplines (e.g., educational psychology, neuropsychology, industrial-organizational psychology) or research fields. Future research might benefit from including other disciplines within the sample of retracted articles, or by searching for additional eligible retracted articles from other research databases beyond Retraction Watch's.

The final limitation pertains to the way sample articles were collected. During the process of screening potentially eligible studies, I made sure to select those that were found in the same year, within the same journal, and employing similar methods. Unfortunately, there were cases in which no articles could be located, so the search criteria were made more liberal. Although controlling for year and journal appears to be a reasonable strategy, it might have qualified articles that were not closely related to the effects of interest. Future research may benefit from a rigorous search strategy involving a careful look at the methods employed in each comparison article to ensure comparability, perhaps by being more flexible in terms of year or journal.

Conclusion

Although data fabrication is a relatively rare phenomenon in psychology, it still occurs frequently enough to be of concern. Fortunately, there exist several statistical tools to detect potential data fabrication, but there has been a dearth of research on these methods. The present

study explored which statistical tools could reliably differentiate retracted articles (due to fraudulent data) from controls. Although most tools were either inaccurate (*p*-value analyses) or difficult to implement (GRIM and GRIMMER), the method of extreme effect sizes appears to be the most promising. Retracted articles were notably similar to controls, except that retracted articles reported statistical tests with far fewer errors. Future research is needed to verify the veracity of the present claims. However, the present findings still function as a useful first step in preventing data fraud, to ultimately improve the psychological literature for the better.

References

Akhar-Danesh, N., & Dehghan-Kooshkghazi, M. (2003). How does correlation structure differ between real and fabricated data-sets? *BMC Medical Research Methodology*, *3*.

Anaya, J. (2016). The GRIMMER test: A method for testing the validity of reported measures of variability. *PeerJ Preprints*, *4*, e2400v1.

Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavioural Research*, *46*, 666–678. https://doi.org/10.3758/s13428-011-0089-5

Ben-Shachar, M., Lüdecke, D., & Makowski, D. (2020). effectsize: Estimation of effect size indices and standardized parameters. *Journal of Open Source Software*, *5*(56), 2815. https://doi.org/10.21105/joss.02815

Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society*, *78*(4), 551–572.

Berger, A., Hill, T., & Rogers, E. (2009–2024). *Benford Online Bibliography*. https://www.benfordonline.net

Bhattacharjee, Y. (2013, April 26). *The Mind of a Con Man*. The New York Times Magazine. https://www.nytimes.com/2013/04/28/magazine/diederik-stapels-audacious-academic-fraud.html

Brown, N., & Heathers, J. (2017). The GRIM Test: A Simple Technique Detects Numerous Anomalies in the Reporting of Results in Psychology. *Social Psychological and Personality Science*, *8*(4), 363–369. https://doi.org/10.1177/1948550616673876

Bosco, F. A., Aguinis, H., Singh, K., & Pierce, C. A. (2015). Correlational effect size benchmarks. *Journal of Applied Psychology*, *100*(2), 431–449. http://dx.doi.org/10.1037/a0038047

Centre for Scientific Integrity. (2018). *The Retraction Watch Database*. [Data set]. ISSN: 2692-4579. http://retractiondatabase.org/

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., Baranski, E., Bernstein, M. J., Bonfiglio, D. B. V., Boucher, L., Brown, E. R., Budiman, N. I., Cairo, A. H., Capaldi, C. A., Chartier, C. R., Chung, J. M., Cicero, D. C., Coleman, J. A., Conway, J. G., ... Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology, 67*, 68–82. https://doi.org/10.1016/j.jesp.2015.10.012

Eckhartt, G. M., & Ruxton, G. D. (2023). Investigating and preventing scientific misconduct using Benford's Law. *Research Integrity and Peer Review*, *8*(1), 1–10. https://doi.org/10.1186/s41073-022-00126-w

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

Fernàndez, L. M., Hardwicke, T. E., & Vadillo, M. A. (2019). Retracted papers clinging on to life: An observational study of post-retraction citations in psychology. *PsyArXiv*. https://osf.io/preprints/psyarxiv/cszpy

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. (11th edition). Oliver Boyd.

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, *102*, 74–78. http://dx.doi.org/10.1016/j.paid.2016.06.069

Green, C. D., Abbas, S., Belliveau, A., Beribisky, N., Davidson, I. J., DiGiovanni, J., Heidari, C., Martin, S. M., Oosenbrug, E., & Wainewright, L. M. (2018a). Statcheck in Canada: What proportion of CPA journal articles contained errors in the reporting of *p*-values? *Canadian Psychology*, *59*(3), 203–210. https://doi.org/10.1037/cap0000139

Hardwicke, T. E., Thibault, R. T., Kosie, J. E., Wallach, J. D., Kidwell, M. C., & Ioannidis, J. P. A. (2022). Estimating the prevalence of transparency and reproducibility-related research practices in psychology (2014–2017). *Perspectives on Psychological Science*, *17*(1), 239–251. https://doi.org/10.1177/1745691620979806

Hartgerink, C. H. J., Voelkel, J. G., Wicherts, J. M., & van Assen, ALM. (2019). Detection of data fabrication using statistical tools. *PsyArxiv*. https://osf.io/preprints/psyarxiv/jkws4

Hendricks, G., Lammey, R., Ofiesh, L., Bilder, G., & Pentz, E. (2023, September 12). *News: Crossref and Retraction Watch*. Crossref.

Horton, J., Kumar, D. K., & Wood, A. (2020). Detecting academic fraud using Benford law: The case of Professor James Hunton. *Research Policy*, *49*, 1–19. .https://doi.org/10.1016/j.respol.2020.104084

Houtkoop, B. L., Chambers, C., Macleod, M., Bishop, D. V. M., Nichols, T. E., & Wagenmakers, E. (2018). Data sharing in psychology: A survey on barriers and preconditions. *Advances in Methods and Practices in Psychological Science*, 70–85. https://doi.org/10.1177/2515245917751886

Hsiao T., Schneider J. (2021). Continued use of retracted papers: Temporal trends in citations and (lack of) awareness of retractions shown in citation contexts in biomedicine. *Quantitative Science Studies*, *2*(4), 1144–1169. https://doi.org/10.1162/qss_a_00155

Hsü, E. H. (1948). An experimental study on "mental numbers" and a new application. *Journal of General Psychology*, *38*, 57–67.

Hüllemann, S., Schüpfer, G., & Mauch, J. (2017). Application of Benford's law: A valuable tool for detecting scientific papers with fabricated data? *Qualitätssicherung und Medizinökonomie*, *66*, 795–802. https://doi.org/10.1007/s00101-017-0333-1

Levelt Committee, Noort Committee, & Drenth Committee. (2012, November 28). Flawed science: The fraudulent research practices of social psychologist Diederik Stapel (Report). https://www.tilburguniversity.edu/sites/default/files/download/Final%20report%20Flawed%20Science_2.pdf

Newcomb, S. (1881). Note on the frequency of use of the different digits in natural numbers. *American Journal of Mathematics*, *4*(1), 39–40.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*(2), 241–301. https://doi.org/I0.1037/1082-989X.S.2.241

Nuijten, M. B., & Epskamp, S. (2024, February 16). *Package 'statcheck'*. https://cran.r-project.org/web/packages/statcheck/statcheck.pdf

Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2015). The prevalence of statistical reporting errors in psychology (1985–2013).

Behaviour Research Methods, *48*, 1205–1226. https://doi.org/10.3758/s13428-015-0664-2

Simonsohn, U. (2013). Just post it: The lesson from two cases of fabricated data detected by statistics alone. *Psychological Science*, *24*(10), 1875–1888. https://doi.org/10.1177/0956797613480366

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-Curve: A key to the file drawer. Journal of Experimental Psychology: General, 143(2), 534–547. https://doi.org/10.1037/a0033242

Steneck, N. H. (2006). Fostering integrity in research: *Definitions, current knowledge, and future directions. Science and Engineering Ethics*, *12*(1), 53–74. https://doi.org/10.1007/pl00022268

Stricker, J., & Günther, A. (2019). Scientific misconduct in psychology: A systematic review of prevalence estimates and new empirical data. *Zeitschrift für Psychologie*, *227*(1), 53–63. https://doi.org/10.1027/2151-2604/a000356

Tödter, K. (2009). Benford's law as an indicator of fraud in economics. *German Economic Review*, *10*(3), 339–351. https://doi.org/10.1111/j.1468-0475.2009.00475.x

Ulrich, R., & Miller, J. (2015). p-hacking by post hoc selection with multiple opportunities: detectability by skewness test?: Comment on Simonsohn, Nelson, and Simmons (2014). *Journal of Experimental Psychology: General*, *144*(6), 1137–1145. https://doi.org/10.1037/xge0000086

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pederson, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, … Yutani, H. (2019). Welcome

to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Xie, Y., Wang, K., & Kong, Y. (2021). Prevalence of research misconduct and questionable research practices: A systematic review and meta-analysis. *Science and Engineering Ethics*, *27*, article 41, 1–28. https://doi.org/10.1007/s11948-021-00314-9