Detecting Academic Fraud Using Benford Law: The Case of Professor James Hunton

Joanne Horton*
University of Warwick
Joanne.Horton@wbs.ac.uk

Dhanya Krishna Kumar
University of Warwick
D.Krishna-Kumar@warwick.ac.uk

Anthony Wood
University of Exeter
A.P.Wood@exeter.ac.uk

Forthcoming: Research Policy

We thank Dan Segal, Yuval Millo, Peter Pope, Gilad Livne, Fani Kalogirou, Julie Barrow, Kevin McMeeking and Facundo Mercado, workshop participants at the University of Exeter, University of Warwick, University of Gothenburg, and the AARG Scholars conference at Queens Mary, University of London for their helpful and constructive comments.

*Corresponding author, Department of Accounting, University of Warwick, Coventry, CV4 7AL.

Detecting Academic Fraud Using Benford Law:
The Case of Professor James Hunton.

ABSTRACT

We investigate whether Benford's Law can be used to differentiate retracted academic papers that have employed fraudulent/manipulated data from other academic papers that have not been retracted. We use the case of Professor James Hunton who had 37 of his articles retracted because there were grave concerns that they contained mis-stated or fabricated datasets. We construct several Benford conformity measures, based on first significant digits contained in the articles, to determine whether Hunton's retracted papers differ significantly from a control group of non-retracted articles by competing authors. Our results clearly indicate that Hunton's retracted papers significantly deviate from Benford Law, relative to the control group of papers. In additional analysis we also find these results are generalisable to other authors with retracted papers. Our findings suggest that potentially both co-authors and journals could consider implementing a data analytical tool which employs Benford Law to highlight potential 'red flag' papers, with a view to decreasing the risk of fraudulent activity and thereby enhancing the credibility of academic papers and journals.

Keywords: Academic Fraud, Hunton, Benford Law, Academic Integrity

JEL Classification: M49

Competing Interests: Authors have no competing interest to declare.

Data Availability: Data are available from the public sources cited in the text.

1. INTRODUCTION

The credibility of science depends on the integrity of scientists. When fraud occurs, it is highly damaging to the reputation of the entire scientific community. Academic dishonesty is perceived to be increasing within the community (Clarke, 2006; Collberg and Kobourov, 2005; Enders and Hoover, 2006; Honig and Bedi, 2012; Hubbard and Vetter ,1996; Lacetera and Zirulia, 2011; Karabag and Berggren, 2012; Grienensen and Zhang, 2012) although it is unclear whether this is because there is more public scrutiny of such misconducts or an actual increase in questionable research practices (QRP). Certainly, the number and frequency of retracted articles is failing to subside (Foo and Tan, 2014). Surveys of academics in various disciplines consistently reveal concerns of unethical behaviour by colleagues (Fanelli, 2009; John et al., 2012). For example, both List et al. (2001) and Necker (2014) find self-confessed fraud rates amongst academic economists of 4% and Bailey et al. (2001) find a similar rate amongst accounting academics. These self-confessed fraud rates are however likely to be higher in reality, given truthful self-admission of this type of fraud by respondents is likely to be small (John et al., 2012). Bedeian et al. (2010) find academics in a sample of 104 management departments within US Business Schools had either 'observed or heard of 27% of colleagues employing data falsification, and Bailey et al. (2001) find that accounting academics believe that on average 21% of articles in the top 30 accounting journals are 'tainted'. Evidence also suggests that replicability is low with researchers only able to replicate studies in a fraction of cases (see McCullough et al., 2006, 2008; Anderson et al., 2008; McCullough, 2009; Begley and Ellis, 2012; Bergh et al., 2017).

The potential benefits from such academic misconduct is high, whilst the likelihood of such behaviour being detected is extremely low and the level of misconduct currently observed is potentially just the 'tip of the iceberg' (Marcovitch, 2007; Lacetera and Zirulia, 2011; Necker, 2014). There are many benefits to academics of undertaking questionable research

practices in order to reduce external pressures caused by: (a) the need to publish in high ranking journals for career progression e.g. tenure, promotion and pay; (Crain and Carruth, 1992; Loeb and Merino, 2000; Graber and Walde, 2008; Almer et al., 2013; Almer et al., 2015; Glover et al., 2006; Glover et al., 2012; Necker, 2014); (b) the inability to detect fraud because by their very nature fraudulent data are designed to elude all the self-correcting process of science (Reich, 2009); (c) the escalating performance expectations (Glover et al., 2012; Craig et al., 2014; McNay, 2016); (d) the demand for complex research as well as increased competition for research funding (Kock, 1999; Bedeian et al., 2010; Graber and Walde, 2008); and (e) the desire by individuals for academic prestige (Almer et al., 2015). However, there are currently no significant deterrents (Cox et al., 2018) because: (a) the use of fraudulent/misrepresented research data is unobservable (Lacetera and Zirulia, 2011); (b) there are no processes to scrutinize the authenticity of such data (Lacetera and Zirulia, 2011; Cox et al., 2018); and (c) there is no real demand by journals to publish replication studies nor implement alternative forms of checking (Delwald et al., 1986; Hamermash, 2007; French, 2012; Cox et al., 2018; Martin and Clarke, 2017).

Although the number of retractions due to academic fraud is on the increase, most are attributable to a small number of authors (Dickins and Schneider, 2016). For example, 5 researchers on the Retraction Watch Top 30 League Table account for 41% of the retractions, with Yoshitaka Fujii being the biggest offender with a total of 183 retractions to his name. Bailey et al. (2001) highlighted accounting research is not immune to this type of fraud, and this was proved spectacularly by Professor Hunton (hereafter Hunton) when 37 of his papers were retracted due to significant concerns that his data was fabricated/misrepresented. One could argue these are a few ‘bad apples’ and we should not be overly alarmed. Even if this is true, unfortunately the actions of a few can have huge ramifications for individual academic’s careers (Hussinger and Pellens, 2019; Mongeon and Larivière, 2016; Jin et al., 2013) and the

academic community as a whole. For example, as noted by Reich (2009), Professor Schön (with 32 retractions to his name due to fabricated data) misled other scientists, consequently at least a dozen laboratories wasted time and money chasing "rainbows" with millions of dollars' worth of US Government Research commissioned to follow up on his fraudulent claims. What we do know is the extent of fraudulent data is not known as it is rarely easy to detect (Marcovitch, 2007). Therefore, any process that can be used to uncover potentially fabricated data should be of benefit to, and welcomed by, the academic community.

In this paper, we aim to investigate whether an analytical process utilising Benford's Law (hereafter BL) can be applied to research output as a screening mechanism, permitting others to assess the authenticity of the research data. Specifically, we investigate whether the digits contained within Hunton's retracted papers significantly deviate from the expected theoretical distributions of BL compared to the digits contained in a set of non-retracted papers. If an analytical tool, based on BL, can identify these retracted papers from a set of non-retracted papers then this potentially will enable authors to assess the integrity of co-authors' data and journal Editors to potentially identify fabrication long before the article is accepted for publication.

BL is the law of natural numbers and was established on a curious observation - certain digits appear more frequently than others in naturally occurring data sets. Specifically, the probability that a number begins with the number one is 30% while the probability that the number begins with the number 9 is only 5%. This distribution of digits, like the normal distribution, is an empirically observable phenomenon (Hill, 1995).

Prior research finds data in many contexts conform to BL, from speed of light and gravitational force (Knuth, 1969; Burke and Kincanon, 1991) to stock prices and returns (Ley and Varian, 1994; Ley, 1996). However, if data has had some human intervention it is unlikely to comply with the Benford distribution (Hill, 1998) because individuals are not particularly

good at replicating known data-generating processes (Camerer, 2003). Specifically, fraudulent data is likely to have more evenly distributed leading digits than BL requires. Consequently, BL is often utilised to help identify fraudulent activity. Tax inspectors and auditors in many countries apply BL to identify fraud and other forms of data manipulation in financial reports and tax returns (Nigrini, 1996; Nigrini, 1999; Durtschi et al., 2004; Nigrini and Miller, 2009; Nigrini, 2012).

In relation to academic fraud, Varian (1972) was the first to suggest using BL as a diagnostic tool for screening model output and forecasts in social science research. It wasn't until recently that this was explicitly examined. Both Diekmann (2007, 2002) and Günnel and Tödter (2009) find the aggregate distribution of digits from regression coefficients and standard errors are very close to the Benford distribution. However, both Diekmann and Jann (2010) and Bauer and Gross (2011) express doubts concerning the discriminatory power of BL to correctly identify academic fraud. To-date only Carlisle (2012) has evaluated data contained in retracted papers, not utilizing BL, but by comparing the variables reported by Fujii with distributions expected by chance. Fujii's data was inconsistent with the expected distributions 85% of the time.

Diekmann and Jann (2010) argue that to ascertain the validity of the Benford test as a fraud detection tool one needs to demonstrate that un-fabricated/non-manipulated data is in accordance with the distribution posed by BL while the manipulated/fabricated data follows a different distribution. We answer this call by examining whether the level of conformity to Benford Law can be used to differentiate between academic papers that have been retracted due to concerns with the data and those papers that have not been retracted and are assumed to be credible and reliable. We use the case of Professor James Hunton who had 37 of his papers retracted because there were concerns that his data had been manipulated/fabricated.

We match these retracted papers to a group of non-retracted papers (which are assumed to be credible) published by competing authors based on the journal, publication year and methodology. Using the first significant digit (leftmost non-zero digit) reported, we examine whether Hunton's retracted articles deviate significantly from the Benford distribution relative to the matched control group of articles. We employ 2 measures of Benford conformity used in the prior literature and calculate the conformity measures for each type of analysis reported in the articles: a) the descriptive statistics and b) regression outputs. We find under all measures, for all types of analysis and after controlling for article characteristics, Hunton's retracted papers consistently and significantly deviate from the Benford distribution relative to the control group.

We also investigate Hunton's 18 non-retracted papers, which have been called in to question (Bentley University, 2014; Healy, 2012). We find mixed results. Under some specifications Hunton's non-retracted articles are not significantly different from the control group whilst under other specifications they are. These findings suggest non-retracted papers may potentially require further investigations to determine the validity of the data.

To mitigate the possibility that our results are driven by other sources of bias, other than Hunton's manipulation, we create an alternative control group which contains non-retracted articles published by his co-authors during our period of interest. It was made clear following the outcome of Bentley University's investigation that none of his co-authors on the retracted papers were implicated in any way. This research design is like a difference-in-difference specification where we observe for the same set of authors, the conformity to BL for papers co-authored with Hunton and those papers not co-authored with Hunton. We find our results are robust to this new specification.

We also examine the use of a prediction model to determine a base-line of conformity for each individual article. We find that our model identifies 70% (24 papers) of Hunton's

retracted papers which significantly deviate from the conformity measure our model would predict if there was no manipulation/fabrication - but with similar article characteristics. Similarly, we find 61% (11 papers) of Hunton's non-retracted papers also significantly deviate from our predictive conformity measure. To test the sensitivity of BL we also utilise a Monte Carlo simulation approach and find (in untabulated results) 47% of first digits within each control sample article needed to be randomly manipulated in order to achieve the average conformity score of Hunton's retracted papers. While only 11% would need to be randomly changed to be significantly different from the control sample.

To test the generalizability of our findings to other retraction cases we examine four additional authors in other academic fields who had a number of their papers retracted due to concerns that their data had been fabricated/manipulated. Specifically, we examine the retracted publications of Professor Stapel, Professor Walumbwa, Professor Lichtenthaler and Professor Sato. We find under all measures, after controlling for article characteristics, the retracted papers consistently and significantly deviate from the Benford distribution relative to the control group, for all analysis except the regression output.

Our overall findings suggest the application of BL, based on the numbers published within a research paper, can potentially highlight abnormalities and thereby raise a 'red flag' of suspicion of fraudulent academic activity. We therefore contribute to the literature by providing the first evidence that using an analytical process utilising BL has the potential to discriminate between academic articles that have been retracted due to concerns of fabricated/manipulated data from those that have not been retracted and are assumed to be unfabricated.

The use of BL to screen for data anomalies is not without risk for fraudsters. The development and use of text-mining platforms have increased the risk of plagiarism (Honig and Bedi, 2012). We argue that the use of BL, in conjunction with other data analytical tools,

could similarly increase the risk of academic fraud. Although we acknowledge the possibility of some unintended consequences, for example that 'milder' QRPs may increase as a result (Gall and Maniadis, 2019).

BL has its limitations and statements of specificity cannot be drawn directly from this study for a number of reasons. First, we cannot be certain the data presented in the research papers in the control group are 'truthful' and 'correct' and 'untrue' and 'fraudulent' for the retracted papers. It may be the case that the non-retracted papers do actually contain fraudulent/manipulated data that is yet to be noticed, and conversely that the retracted papers do not contain fraudulent/manipulated data. Certainly, in the latter case the authors investigated denied such behaviour[1]. However, we can take some comfort from our simulation results which suggest that the conformity measures for fabricated data are significantly different from those derived from the non-fabricated data.

Second, there is a lack of a golden criteria to judge deviation or compliance with BL since there is currently no clear mathematically-derived critical value. Moreover, the thresholds of acceptability or conformity vary depending on sample size and the nature of the sample population (Banks, 2000).

Third, not all populations conform to a Benford distribution (Nigrini and Mittermaier, 1997) and the effectiveness of BL declines as the level of contaminated entries drops (Bauer and Gross, 2011). All these limitations result in a risk of both false positives and false negatives (Diekmann and Jann, 2010).

Lastly other questionable research practices (QRPs), such as p-hacking, cannot be identified using BL as they do not involve fraudulent behaviour. A number of procedures have been suggested to identify such practices as well as ways of reducing QRP incentives (e.g.

[1] Although the behaviour of Prof Hunton during the University's investigation would suggest he had something to hide.

publishing no result papers). Fortunately, unlike fraud, QRPs appear to have relatively less impact on the scientific community (Head et al., 2015).

With these limitations in mind, we therefore advocate that its application could be used for screening articles for fraud i.e. identifying 'red flags', whereby any abnormalities observed (i.e. significant deviations relative to a set of comparable non-retracted papers) would raise suspicion and thus provide the basis for a deeper discussion with the authors. For example, by requesting the original dataset and code or an explanation as to why the data may not be expected to conform to such a distribution. This process will not prevent fraudulent behaviour, and to the best of our knowledge nor do any of the current (or suggested) policy recommendations (Hussinger and Pellens, 2019; Necker, 2014; Cox et al., 2018). Utilising BL should, however, potentially make the task more difficult and riskier for the fraudster (like a burglar alarm for a criminal). The fraudster would have to engineer both the statistical results and closer compliance with Benford Law - a much more difficult task, especially when you factor in additional analysis required by co-authors and reviewers. As noted by Bailey (2015) the topic of research misconduct is of great importance to academia, so any additional processes reducing this risk should be welcomed. Our paper answers this call by identifying BL as potentially an additional risk-reducing mechanism.

The remainder of the paper is organised as follows: Section 2 documents the case of Professor Hunton. Section 3 describes BL, its origins and prior application both within practice and academic literature. Section 4 describes the data and methodology used within this study, Section 5 presents the findings and Section 6 presents some additional analysis, while Section 7 concludes and discusses possible applications.

2. THE CASE OF JAMES HUNTON

The *Accounting Review* journal received correspondence from a reader, highlighting a concern regarding the unrealistically high sample of 150 US-based offices used in Hunton and Gold

(2010). Following an enquiry, the journal reported a retraction notice in November 2012 and uncovered the first evidence of misstatement and fraud by Professor James Hunton, former award-winning accounting professor at Bentley University.[2]

This initial retraction led to an 18-month investigation and an avalanche of further retractions, the most recent being issued in April 2016. To-date there have been 37 retraction notices issued on research papers published by Hunton, and he is currently ranked #12 on the Retraction Watch Leader board.[3] Hunton resigned from his professorship at Bentley University one month after the initial retraction (Healy, 2012) and has not made any public comment regarding the allegations. The investigation into Hunton's alleged academic fraud by his then-employer Bentley University determined that, acting alone, Hunton falsified data in 2 papers, and attempted to destroy further evidence pertaining to the case.[4]

The report concluded that the *whole body of Dr Hunton's extensive research while a faculty member at Bentley University* [approx..50 papers] *must now be considered suspect*. On 25 June 2015, the American Accounting Association (AAA) publicly retracted 25 articles, and one section of an article, from its journal collection. The retractions were based upon the evidence provided by Bentley University (2014) and the "*co-authors' inability to provide data or other information supporting the existence of primary data, or to confirm that their studies were conducted as described in the published articles".* Akin to the findings in the Bentley University investigation summary*, "the Association review team found no evidence that Dr. Hunton's co-authors were aware of or complicit in Dr. Hunton's actions*". (AAA, 2015).

Retraction notices suggest the validity and legitimacy of the data which underpins each of the retractions cannot be established. In many cases Hunton (through his counsel) contended

[2] Hunton was bestowed the "Scholar of the Year" award by Bentley University in 2006, and Accounting Horizons Best Paper Award, also in 2006.
[3] A list of the top 30 academic authors by number of retractions can be found here: http://retractionwatch.com/the-retraction-watch-leaderboard/
[4] After Hunton's resignation, Bentley discovered that his office had been completely cleaned out of all physical files and that his laptop had been wiped clean; despite having been cautioned on numerous occasions to retain all documents relevant to the investigation (Bentley University, 2014).

the data had been genuine, despite not being able to provide requested information to support this validity. It was his co-authors, complying with the investigations who decided that their papers should be retracted, as Hunton had provided the data for each of the studies and legitimacy could not be confirmed.

The response by the AAA was to establish a Publication Ethics Task Force in 2013 which was tasked with developing standards around plagiarism and fabrication of data. The outcome of this review was to increase the onus on researchers to minimize misconduct. Specifically, the requirement for author(s) to provide positive assurances and accept joint responsibility for the integrity of the data employed in the manuscripts (AAA, 2014, p1). Some journals also responded by requiring additional data assurances - *Journal of Accounting Research* (JAR, 2014; 2016; 2018)[5].

3. BENFORD LAW (BL) AND PRIOR RESEARCH

BL relates to the theoretical mathematical distribution of the leading digits contained within a wide variety of data sets. Its origins can be traced to the astronomer Newcomb (1881) who observed that the first pages in logarithmic tables would consistently wear out faster than the last ones. He noted that "*the first significant figure is oftener 1 than any other digit, and the frequency diminishes up to 9*", determining the occurrence probability of the leading first digit, *d*, as being:

$$P_d = log_b(d+1) - log_b(d) = log_b\left(\frac{d+1}{d}\right) \quad (1)$$

Where *P* is the probability of the occurrence of first digit *d*, and *b* is the logarithmic base. For example, the probability of the first digit of number *n* in a decimal system being a one would be $\log_{10}(2) = 0.301$. The expected probability of occurrence for leading digits 1 through 9, results in the theoretical distribution which today is referred to as BL.

[5] https://research.chicagobooth.edu/-/media/research/arc/docs/journal/updated-data-policy-for-jar.pdf

Leading first digit, d	1	2	3	4	5	6	7	8	9
Occurrence probability, P_d	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046

Benford (1938) was the first to popularise and provide empirical evidence finding 'real-life' data which is surprisingly close to the theoretical distribution defined by equation (1). Since Newcomb's earlier work had been overlooked, the frequencies became enshrined as BL.

Pinkham (1961) demonstrates BL is independent from scale, and Boyle (1994) finds data following Benford's distribution will continue to do so after being repeatedly multiplied, divided or raised to integer power. A formal explanation and derivation of BL can be found in Hill (1995), highlighting that the Benford distribution, like the normal distribution, is an empirically observable phenomenon.

Researchers have detailed conformity with BL across numerous areas, from innocuous field studies to complicated scientific expressions. Examples of conformity to BL include: common physical constants such as speed of light and gravitational force (Knuth, 1969; Burke and Kincanon, 1991); numbers on the World Wide Web (Leibon, 2008); internet traffic (Arshadi and Jahangir ,2014); survey and response data (Diekmann, 2002; Schäfer et al., 2005; Schräpler and Wagner, 2005); ebay bids (Giles, 2007); stock prices and returns (Ley and Varian, 1994; Ley, 1996; Pietronero et al., 2001); financial variables and statements (Clippe and Ausloos, 2012; Nigrini, 2012; Amiram et al., 2015). Not all populations conform to a Benford distribution, for example, numbers influenced by human thought (Nigrini and Mittermaier, 1997), ranking systems that constrain choices, and other naturally constrained measures e.g. systolic blood pressures, cholesterol values etc. (Al-Marzouki et al., 2005).

Additionally, when data has had some human intervention it is unlikely to conform to the Benford distribution (Hsü, 1948; Kubovy, 1977; Hill, 1998) principally because individuals are not particularly good at replicating known data-generating processes (Camerer, 2003). As

Durtschi et al. (2004) find that when an individual adds or deletes expense claims, they will deviate from BL. While Watrin et al. (2008) find that individuals taking part in experimental studies do not adhere to BL when they are instructed to concoct the numbers. This inability of individuals to produce random numbers that conform to a Benford distribution has been utilised by many to detect fraudulent activity.

Prior research addresses the use of analytical procedures, including BL, employed by auditors to uncover account manipulations and fraudulent behaviours (Bierstaker et al., 2006). It is claimed digit analysis is one of the most cost-effective methods to help identify groups of data which have high probabilities of being manipulated (Carslaw, 1988; Berton, 1995; Nigrini, 1996; Wright and Ashton, 1989; Quick and Wolz, 2003; Hales et al., 2009; Bhattacharya et al., 2011; Bierstaker et al., 2006).

Nigrini (1996, 1997, and 1999) established BL within forensic accounting as a tool for the detection of tax evasion and other fraudulent activities. Tax authorities in many countries apply BL to check for anomalies. Nigrini and Miller (2009) also provide a guide to auditors on how to use BL to detect errors in transactional data. Relatedly, several innovative algorithms have also proven the effectiveness of BL as a fraud detection tool (Busta and Weinberg, 1998; Huang et al., 2008; Bhattacharya et al., 2011). BL has also been applied to detect fabricated interview responses created by the interviewers rather than legitimate responses from interviewees (Schäfer et al., 2005; Schräpler and Wagner, 2005).

Both Diekmann (2007, 2002) and Günnel and Tödter (2009) test whether the regression coefficients and other statistics reported in *non-retracted* academic articles follow BL. Diekmann (2007) examined the first two digits of unstandardized coefficients reported in tables published in two volumes of the *American Journal of Sociology*. Diekmann (2007) finds the published coefficients follow BL, whilst experimental regression data which he provided to his students and asked them to fabricate did not conform. Overall, Diekmann (2007) concludes

that the digits of published coefficients closely approximate BL and therefore potentially may be used to detect fraud in empirical sciences. Günnel and Tödter (2009) also finds that BL applies in aggregate to regression coefficients and standard errors reported in economic research.

Recently, both Diekmann and Jann (2010) and Bauer and Gross (2011) question the efficacy of BL to identify manipulated/fabricated research output. Diekmann and Jann (2010) argue that in order to ascertain the validity of the Benford test as a fraud detection tool one needs to demonstrate that un-fabricated/non-manipulated data is in accordance with the distribution posed by BL, while the manipulated/fabricated data follows a different distribution. While Bauer and Gross (2011), argue that although regression coefficients follow BL, its application can only discover fraud if the number of cases of forgery is significantly large, which they believe is not the case.

Based on these prior findings and concerns we empirically test whether BL can identify retracted articles (due to fraud) from a group of non-retracted articles based on regression and statistical output reported in the papers. Although replication is considered the prime strategy against scientific misconduct it is seldom performed and there is a strong likelihood that such a process would not have identified Hunton's fabricated dataset anyway. Certainly, replication will fail to detect fraud if fraudulent data is submitted along with the code as the results will still replicate those published. Therefore, a system that can signal issues without the initial need for the original data should be of value to co-authors and journal editors. This may overcome the concerns of many authors who are hesitant to provide original data, arguing a proprietary nature and provision may infringe upon their competitive advantage (McCullough and Vinod, 2003; Gill and Meier, 2000). As Tödter (2015) argues, BL could be used to increase the effectiveness of replication studies.

3.1. When would we expect genuine research data to closely follow BL and not follow BL?

BL is applicable to either primary or secondary sources of data. If, however, the data is constrained then BL may not apply. For instance, if a survey or questionnaire constrains the responses of the participants to a range between 1 to 5 (e.g. Likert Scale) then the reported descriptives relating to the mean, median and quartile values of the data will never exceed 5. Under these conditions the probability of the first digit being between 6 and 9 is zero contrary to the Benford Law frequency. However, the reported standard deviation, correlation and regression results of these constrained responses will still follow BL as mathematically the first lead digit can be any number between 1 and 9 (Shiffler and Harsha, 1980). Moreover, data may not conform to BL if the number of digits is below 22.[6] Also, as noted in the Introduction, data will not follow BL if there has been some human intervention, this includes the use of any data generating processes (DGP). Specifically, if the underlying data (Xs and Ys) are not naturally occurring but the product of a DGP they will not follow BL nor will any subsequent output that has been generated from the DGP dataset.

Winsorizing, censoring, ranking or truncating the variables does not prevent conformity to BL, given the resulting data is considered a sample that represents the population. If, the descriptive and analytical results have a log-normal distribution, then it may follow the BL pattern even if the data is manipulated. So, in summary, any analysis of data will follow BL if the data has a geometric tendency and there is no pre-defined structure (e.g. DGP) as to how the numbers should occur.

4. DATA AND METHODOLOGY

4.1. Sample

The sample consists of 55 papers sole or co-authored by Hunton during the period 1996-2012. 37 of these papers have been retracted and his remaining 18 papers are non-retracted, although

[6] Prior literature suggests that anything below 80 digits potentially may not follow BL (Morrow, 2014). We note that 32 papers from our sample had less than 80 digits. However, we find our results are not sensitive to the exclusion of these papers.

have been called into question by Healy (2012) and Bentley University (2014). We match these papers to a control group of papers. The selection process for the control group is based on the following criteria: for each Hunton paper, we obtain all non-Hunton/non-retracted papers that were published in the same journal, in the same year, and using the same methodology.[7] We categorise the methodological approach using 4 classifications: database, survey, questionnaire, or experimental. Our initial sample consists of 248 control group papers. From these we exclude 2 papers as they do not meet the minimum threshold of 22 digits, thus our final sample consists of 246 control papers and 55 Hunton papers and is detailed in Table 1.

<Table 1>

Table 2 lists the 37 retracted papers.

<Table 2>

We review our sample of papers to identify any that due to their construction would not follow BL. We identify 68 articles where the data is constrained and therefore we exclude the reported mean, median and quartile values from their descriptive analysis.

4.2. Measuring Level of Conformity to Benford's Law (BL)

A common measure of conformity to BL is the mean absolute deviation (*MAD*) score (Nigrini, 2012). The *MAD* score is defined as the mean of the absolute value of the difference between the frequency of each first digit within the sample, and the frequency as determined by BL.

$$MAD = \frac{\sum_{i=1}^{K} |AF - EF|}{K} \quad (2)$$

Where *AF* is the actual frequency of the leading digit observed, *EF* is the expected frequency as determined by BL [equation (1)], and *K* is the number of leading digit bins (equal to 9 for the first leading digit). We employ the Johnson and Weggenmann (2013) modified version of this MAD score which is applicable to our smaller dataset (Nigrini, 2012). Under the modified

[7] In 4 instances, we find no control papers of the same methodology within the same journal-year. In these cases, we select all control papers from the next subsequent year where data is available.

model *AF* is the difference between the actual frequency and the BL frequency and *EF* is the mean of the differences between the actual frequency and BL frequency. The *MAD* score does not have a mathematically-derived critical value to determine the level of significant deviation from BL although several papers offer a range of critical values. For example, Slepkov, et al. (2015) suggest values above 0.015, whilst Banks (2000) suggests over 0.020 signals the population should be scrutinized. However, as noted by Banks (2000) thresholds of acceptability or conformity vary depending on sample size and/or the nature of the sample population. We calculate all conformity measures using the first significant digit (leftmost non-zero digit) reported, following the findings of Günnel and Tödter (2009). The scale invariance of the *MAD* score makes it particularly useful when examining large pools of first digits (Amiram et al., 2015) and more importantly when comparing *MAD* scores across articles since the pool of digits in each article can differ significantly.

Our second measure of conformity is the Kolmogorov-Smirnov (*FSD-Score*) statistic which is defined as the maximum cumulative deviation from the theoretical distribution of BL, for leading digits 1-9 and is calculated as follows:

$$FSD\text{-}Score=Max(|AF_1 - EF_1|,|(AF_1+AF_2)-(EF_1+EF_2)|,..,|(AF_1+AF_2+..+AF_9)-(EF_1+EF_2+..+EF_9)|) \quad (3)$$

The advantage that *FSD-Score* has over *MAD* is that we can empirically test conformity to BL using critical values of the Kolmogorov-Smirnov test statistic. At 1% significance, the critical value is $1.63/\sqrt{P}$ assuming $P>35$, where P is the total number (*Pool*) of first digits used in the calculation of *FSD-Score*. When *FSD-Score* is greater than the critical value then we can infer that the distribution does not follow BL. We therefore create an indicator variable *KS1* which is equal to one if *FSD-Score* is greater than the critical value and zero otherwise. The criticism of this measure is that it is highly sensitive to the pool of digits and tends to over reject. For no rejection, the statistic requires near perfect adherence to BL (Nigrini, 2012) as the pool increases. Consistent with Amiram et al. (2015) we focus primarily on the *MAD* score results

and report *KS1* results for completeness.

4.3. Modelling level of conformity between Hunton's retracted articles and the Control Group articles

To determine whether the retracted papers of Hunton significantly deviate from the expected theoretical distributions of BL compared to a set of non-retracted papers we estimate the following cross-sectional regression. The variable of interest is the indicator variable *Hunton_R* which equals one if the article is retracted and authored by Hunton and zero otherwise:

$$Conformity\ Measure = \alpha + \beta_1 Hunton\_R + \beta_2 Pool + \beta_3 Num\_Authors + \beta_4 Experimental + \beta_5 Survey + \beta_6 Questionanaire + \beta_7 Linear + Year\ F.E. + Journal\ F.E. \quad (4)$$

The conformity measure is either *MAD*, *FSD-Score* or *KS1*. We employ ordinary least squares regression for dependent variables *MAD* and *FSD-Score*; and a logit regression model for the dependent variable *KS1*. All dependent variables increase the higher the levels of the non-conformity measures. We expect our variable of interest, *Hunton_R* to deviate more from BL relative to the control group and predict β_1 will be positive and significantly different from zero. Equation (4) includes several control variables to capture individual article characteristics. The first controls for the number of first digits used in the calculations (*Pool)* and proxies for paper complexity. The second controls for the number of authors attributable to each paper (*Num_Authors)*. More co-authors could infer a greater level of cross-checking although, as Foo and Tan (2014) find, fraudulent researchers tend to work in larger teams. We also control for the methodology employed in each article *Experimental*, *Survey*, *Questionnaire* and *Database*. One could argue papers which are database-driven will be more likely to show signs of conformity to BL - the data is readily available and easier to replicate thus decreasing the risk of fraud compared to a proprietary survey. Lastly, we control for whether the analysis is conducted employing a linear model (*Linear*), as it may be the case that a non-linear process

could potentially bias the output and reduce the likelihood of non-conformity to Benford, even though the raw data conforms. We also include journal fixed effects (*Journal F.E.*) as journals have different review processes which may be more or less stringent. Lastly, we also include year fixed effects (*Year F.E.)*.

To examine Hunton's remaining 18 articles that have not been retracted but called into question, we replace the variable *Hunton_R* with *Hunton_NR* which takes the value of one if the article is non-retracted and authored by Hunton, zero otherwise. It is unclear whether these articles are more likely to deviate from BL compared to the control group and therefore we do not make any predictions regarding the sign of the coefficient. Certainly, given the papers have not been retracted this would suggest that they are free from fraud or misstatement. However, Bentley University concluded that all Hunton's work should be considered suspect, and therefore there is a possibility that a few or all 18 papers contain some fraudulent data.

5. RESULTS

5.1. Univariate Analysis

We report descriptive statistics in Table 3, Panels A and B. Panel A reports the distributions of the conformity measures (*MAD*, *FSD-Score* and *KS1*) for both Hunton's retracted papers and for the control group. For each article conformity scores are calculated for all digits reported, digits just in the descriptive analysis (which also includes the digits from correlation matrix) and digits just in the regression analysis. Preliminary investigation of the differences between Hunton's retracted papers (columns 1-5) and the control group (columns 6-10) reported in Panel A, reveal that for all conformity measures Hunton's retracted papers deviate more from BL relative to the control group (columns 11 and 12). Specifically, the average (median) *MAD* score for all digits contained in Hunton's retracted papers is 0.0221 (0.0209) which is significantly higher than the control group average (median) score of 0.0148 (0.0148). The *MAD* scores from the descriptive and regression analysis increase with an average (median) of

0.0286 (0.0241) and an average (median) of 0.0306 (0.0258) respectively. Both are significantly higher than the control group with a mean (median) *MAD* score of 0.0212 (0.0191) for the descriptive analysis and a mean (median) 0.0221 (0.0194) score for the regression analysis. *FSD-Score* and *KS1* provide a similar picture.

<Table 3>

For Hunton's non-retracted articles, reported in Panel B, the picture is not as clear as above. The average (median) *MAD* score for all digits is 0.0200 (0.0180) which is only marginally different from the control group (columns 6 and 8). The mean *MAD* score relating to the descriptive digits is significantly higher than the control group's mean, but not for the regression digits. The *FSD-Scores* provide a similar picture. The mean *KS1* for both the all digit group and regression digit group indicates approximately 44% (8 articles) of Hunton's non-retracted articles are above the critical value of *KS*. Both groups are significantly different from the control group.

In Figure 1 we plot over time the average *MAD* score values, for all digits, for both Hunton's retracted and non-retracted articles and those from the combined control groups. The control group *MAD* score remains stable over the entire timeframe; while, more noticeable is the volatile nature of Hunton's retracted papers. The retracted group of papers, *Hunton_R*, shows a dramatic spike in 1999 and 4 other peaks are evident; in total 51% (19 papers) of Hunton's retracted papers exceed the 0.02 threshold suggested by Banks (2000), compared to the control sample with only 19% (29 papers) exceeding the threshold. The group of non-retracted papers, *Hunton_NR*, although showing a large peak in 2004, are close to 0.02; in total 39% (7 papers) of Hunton's non-retracted papers exceed 0.02. Overall, we can see that Hunton's retracted papers consistently demonstrate large and inconsistent *MAD* scores relative to the control group.

<Figure 1>

With respect to our control variables (Table 3, Panel A) we find the average (median) number of first digits (*Ln_Pool*) in Hunton's retracted articles is 5.65 (5.58), with an average (median) *Num_Authors* of 2.65 (3). The majority, 65%, of Hunton's retracted papers used experimental methodology, while 22% were survey-based. Compared to the control sample these control variables are not statistically different except in the case of *Num_Authors* which for the control group is marginally lower with an average (median) of 2.32 (2).

We report correlations for the conformity measures and the control variables for all digits in Panel C. Below the diagonal we report correlations for Hunton's retracted analysis and above the diagonal Hunton's non-retracted analysis. Both above and below the diagonal all the conformity measures are positive and significantly correlated to one another. For the control variables, *Database* is negatively and significantly correlated to *MAD* and *FSD-Score* while *Num_Authors* is positively and significantly correlated with all conformity measures (below diagonal).

5.2. Regression Analysis

Table 4 presents the estimates of equation (4) where the dependent variable is either *MAD* score (columns 1 to 3), or *FSD-Score* (columns 4 to 6) or *KS1* (columns 7 to 9). Columns (1), (4), and (7) contain the analysis for all digits, columns (2), (5), and (8) report the analysis for the digits contained in descriptive and columns (3), (6), and (9) for the digits contained in regression output. All the results reported are based on continuous variables that have not been winsorized. In untabulated results we find winsorizing at the top and bottom 1% does not affect our results. In addition, the reported heteroskedasticity-robust standard errors are not clustered, as it is unclear what potential bias in the estimates we would observe (Petersen, 2009). However, we do investigate alternative clustering specifications - by methodology, by journal - but find our results do not significantly change, suggesting no significant bias in our reported estimates.

The coefficients on the indicator variable *Hunton_R* under all specifications (columns 1 to 9) is positive and statistically significant, except column (7). Column (1) indicates Hunton retracted papers *MAD* score is on average 0.006 higher than those in the control group ($p<0.01$). This coefficient implies, given the average *MAD* score for the whole sample is 0.0162, Hunton's retracted papers are 37% larger than the control group. Similarly, the *MAD* score for the descriptive digits (column 2) and regression digits (column 3) for Hunton's retracted articles is on average 35% and 38% higher than the control group respectively. The *FSD-Score* results provide a similar picture. Column (4) indicates Hunton's retracted papers' *FSD-Score* is on average 0.023 ($p<0.01$) or approximately 35% higher than those in the control group. Similarly, the descriptive digits (column 5) and regression digits (column 6) for Hunton's retracted articles the *FSD-Score* is on average 24% and 36% higher than the control group respectively, both significant at the 10% or better significance level. Columns 7 to 9 present the estimates from the logit model, where the dependent variable is *KS1*. We find, in columns (8) and (9), the coefficient on *Hunton_R* is positive and significant at the 1% and 5% level significance respectively. Column (8) reports a coefficient of 1.286 indicating digits contained in the descriptive output of Hunton's retracted papers are nearly 4 times the odds of not conforming with BL compared to the control group ($p<0.01$). We find similar odds for the regression output (1.290; $p<0.05$). In column (7) we find the coefficient is positive but not statistically different from zero.

<Insert Table 4>

Among the control variables in column (1) the coefficients *Survey* is positive and statistically significant at the 10%. While *Ln_Pool* is negative and statistically significant at the 1% level. The remaining control variables are not significantly different from zero.

Table 5 reports the results of Hunton's non-retracted articles. Overall, we find a mixed picture consistent with the prior univariate analysis. When measuring conformity using the

MAD score, we find Hunton's non-retracted papers for the descriptive and regression digits are significantly different from the control group, indicating that on average Hunton's non-retracted articles deviate more than the control group. However, when measuring conformity using the *FSD-Score* only those digits contained in the all digits and regression analysis are found to be statistically different from the control group. Similarly, the results from *KS1* analysis indicate that Hunton's non-retracted papers are not significantly different from the control group[8]. These results indicate, consistent with Healy (2012) that Hunton's non-retracted papers potentially need further investigation. In unreported results we aggregate the sample of retracted and non-retracted Hunton articles. The results continue to indicate a clear and consistent deviation from BL of Hunton's retracted articles relative to the control group, while providing a mixed picture for his non-retracted articles.

<Insert Table 5>

Overall, these results provide the first evidence that BL can discriminate between papers that have been retracted due to fraud concerns, relative to the control group of non-retracted articles where there is no evidence of any fraudulent behaviour by the authors. If this latter assumption is incorrect, and some of the non-retracted articles do contain some level of fraud/misrepresentation, then this would bias our results against finding significant differences between the conformity measures of Hunton's retracted articles and those of the control group.

6. ADDITIONAL ANALYSIS

6.1. Alternative Control Group

To investigate the sensitivity of our results to different control group specifications we construct an alternative control group containing only those articles published by Hunton's co-authors during 1996 to 2012. Such a control group mitigates the possibility that our results are

[8] Due to the small sample size, several multicollinearity issues prevent us from adequately running the logit model for the pool of regression analyses numbers.

driven by other sources of bias, rather than Hunton's manipulation/fabrication. Bentley's initial investigations found no evidence that Hunton's co-authors were aware of or complicit in Hunton's misconduct (Bentley University, 2014). To-date no co-authors have been implicated in any of the fraudulent behaviour.

Creating a control group of all of Hunton's co-authors provides a research design that is like a difference-in-difference specification where we observe for the same set of authors, the conformity to BL for papers co-authored with Hunton and those papers not co-authored with Hunton.

Table 6 presents the results of equation (4) using this alternative control group. We find our results are robust to this new specification. Consistent with the prior results we find under all specifications (columns 1-9) the coefficient on *Hunton_R* is positive and statistically significant. The coefficient on *Hunton_R* in column 7 is 1.460 ($p<0.01$) indicating Hunton's retracted articles have 4 times the odds of significantly deviating from BL relative to his co-author's articles. We also examine the two partial retractions from Seybert (2010) and Bhojraj & Libby (2005). Specific sections from these articles were retracted by the authors because they relied on data supplied by Hunton which could not be verified. We find removing the retracted sections reduces both papers' *MAD* scores by approximately 33% and 25% respectively and result in both *MAD* scores below 0.02 threshold.

<Insert Table 6>

For Hunton's non-retracted articles, we find our results are consistent, except in one case. Specifically, the coefficients on *Hunton_NR* when the dependent variable is *MAD* is no longer significantly different from the alternative control group for the descriptive digits (column 2). Similarly, for the *KS1* analysis we find again consistent results except in one case. The coefficient on *Hunton_NR* is now marginally statistically significant for the all digits specification (column 7).

Overall the results provide a similar picture, the retracted articles are more likely to deviate from BL relative to the alternative control group, whereas the non-retracted papers provide a mixed picture - under certain specifications they show lower deviation and in other specifications a degree of greater deviation, relative to the control group.

6.2. Predicted Conformity *MAD* Scores ($\widehat{MAD}$).

One of the limitations of the *MAD* score, as noted earlier, is that it does not have a mathematically-derived critical value to determine the level of significant deviation from BL. As Banks (2000) notes, the threshold of acceptability or conformity may vary depending on sample size and/or the nature of the sample population. However, we are interested to determine whether, on a paper-per-paper basis, a paper's *MAD* score is significantly larger than one would expect from a comparable set of papers. We therefore consider the utilization of a version of equation 4 as a potential *MAD* prediction model.

$$MAD = \alpha + \beta_1 Pool + \beta_2 Num\_Authors + \beta_3 Experimental + \beta_4 Survey + \beta_5 Questionanaire + \beta_6 Linear + Year\ F.E. + Journal\ F.E. \quad (5)$$

We run the above model on the aggregated control group samples (573 papers) which we assume are free from manipulation and fabrication. We then obtain the estimates and apply these to each of Hunton's papers to determine a predicted *MAD* score ($\widehat{MAD}$), which reflects a *MAD* score that we would expect *if* the paper was not retracted. The abnormal *MAD* score (*AB_MAD*) for each of Hunton's papers is the difference between its predicted $\widehat{MAD}$ and its actual *MAD* score, so a negative *AB_MAD* (*MAD* > $\widehat{MAD}$) potentially captures the impact of the Hunton's manipulation. Table 7 reports our findings. Panel A reports the descriptives for the abnormal *MAD* scores. For both Hunton's retracted and non-retracted samples, the average *AB_MAD* scores (for each analysis) are negative and significantly different from the control group at 5% level or better. Panel B splits the samples based on the sign of the *AB_MAD*. We find 27 of Hunton's retracted papers (73%) had a *MAD* score (for all digits) above $\widehat{MAD}$, and

the magnitude of the *AB_MAD* is significantly different from the control group at the 1% level. Whereas Hunton's remaining 10 retracted papers whose *MAD* score is below the threshold predicted $\widehat{MAD}$ the *AB_MAD* is not significantly different from the control group. We find similar results for the descriptive and regression digits with 65% and 60% of Hunton's retracted papers' actual *MAD* scores being greater than $\widehat{MAD}$ respectively. Again, the magnitude of the *AB_MADs* when *MAD* is greater than $\widehat{MAD}$ are significantly different from the control group at the 1% level but not significantly different when *MAD* is less than $\widehat{MAD}$.

<Insert Table 7>

Overall, our model identifies approx. 73% of Hunton's retracted papers that are suspicious as the *MAD* score is significantly higher than one would predict from a comparable set of non-retracted papers. Given the possibility of both false positive and negatives noted earlier we are cautious to suggest that we have identified correctly all of Hunton's dishonest papers.

6.3. Investigating Hunton's Co-Authors

To provide further evidence that our findings indicate BL can differentiate between retracted papers due to fraudulent/manipulated data and non-retracted papers we also investigate other individual authors - Hunton's co-authors. Specifically, we re-run all regression analyses for each individual co-author whose papers were not co-authored with Hunton during 1996-2012. We find no individual author's conformity measures are significantly different to the conformity measures of either of the control groups. For example, Table 8 provides the results for two co-authors of Hunton's - Professor Libby and Professor Wier, using the *MAD* conformity measure. Professor Libby co-authored with Hunton 7 times and published 10 papers during the period without Hunton, while Professor Wier co-authored with Hunton 6 times and published 11 papers during the period without Hunton. We find neither Professor Libby's nor Professor Wier's published articles deviate significantly from BL relative to both

control groups. In un-tabulated results, we find consistent results using the alternative conformity measures. These results provide strong evidence that our findings are not an artefact of investigating one author relative to a collection of authors. Replication of our models using different individual authors provides consistent and clear evidence - these authors' papers are not significantly different from other non-retracted papers, unlike Hunton's retracted papers.

<Insert Table 8>

6.4. Association between MAD scores and Citations

We also examine whether there is any association between the *MAD* scores and the number of citations each paper receives before the first retraction notice in 2012 by the *Accounting Review*. The number of citations are hand collected from *Google Scholar* and consistent with Meyer et al. (2018) we control for several paper characteristics, including the subject matter of the article (e.g. auditing, tax, financial accounting etc.). Table 9 reports the results. We find Hunton's papers, retracted or otherwise, do not differ significantly in terms of citations from those of the control group. We also find that the number of citations is not significantly related to the *MAD* score. This suggests, certainly within the accounting literature, citations per se may not be a good indicator of paper quality.

<Insert Table 9>

6.5. Generalizability of our findings to other academic fields.

To determine the generalizability of our findings we examine additional authors with retracted papers. Specifically, Professor Stapel (Professor of Social Psychology); Professor Walumbwa (Professor of Management); Professor Lichtenthaler (Professor of Management); and Professor Sato (A Bone Specialist). The combined total of retractions for these authors is 116 of which 92 were retracted due to concerns with the data. Upon examination only 45[9] of these 92 retracted papers have data analysis or sufficient analysis to construct our conformity

[9] Specifically, 29 for Professor Sato; 8 for Professor Walumbwa; 5 for Professor Lichtenthaler; and 3 for Professor Stapel.

measures.[10] We match these papers to a set of control papers consistent with our prior methodology, resulting in a sample of 166 control papers.[11] We report the descriptive statistics in Table 10, Panel A. Preliminary investigation of the differences between the retracted papers (columns 1-5) and the control group (columns 6-10), reveal the *MAD* conformity measures of the retracted papers differ significantly to the control group (columns 11 and 12) at the 10% level of significance or better using a one-tailed t-test where we expect the score to be greater for the retracted group of papers. The *KS1* and *FSD-Score* analysis indicates no significant differences.

Table 10, Panel B presents the estimates of equation (4) when we replace *Hunton_R* with the indicator variable *Retracted* which equals one if the article is retracted and zero otherwise. The dependent variable is either *MAD* score (columns 1 to 3), or *FSD-Score* (columns 4 to 6) or *KS1* (columns 7 to 9). Columns (1), (4), and (7) contain the analysis for all digits, columns (2), (5), and (8) report the analysis for the digits in descriptive output and columns (3), (6), and (9) for the digits contained in regression output. The coefficients on the indicator variable *Retracted* for the all digit and descriptive analysis, with the exception of the *FSD-Score* (columns 1 to 9) is positive and statistically significant. No significant differences are observed for the regression analysis for any of the conformity measures. Column (1) indicates the retracted papers' *MAD* score is on average 0.002 higher than those in the control group and statistically significant at 10% level. This coefficient implies, given the average *MAD* score for the whole sample is 0.0161, that the retracted papers' *MAD* scores are 12% larger than the control group. Similarly, the *MAD* score for the descriptive output (column 2) for the retracted articles is on average 27% higher than the control group and is statistically

[10] For example, many of these retracted papers reported only graphs, figures, or pictures rather than tables of analysis. We also examined Professor Jan Hendrik Schön's 26 retractions (out of 33) which were due to concerns with the data. However, in all 26 retracted papers no numerical analysis was reported.

[11] A number of the papers used a Likert Scale ranging from either 1 to 5 or 1 to 7; consequently, we removed the mean, median etc. from the descriptive output.

significant at 1% level. Both the *FSD-Score* and *KS1* results provide a similar picture (columns 4 to 9).

<Insert Table 10>

These results therefore suggest that our earlier findings are generalizable to other authors, although with more moderate levels of significance. Interestingly the regression output provides no clear differences between the retracted and control groups (unlike the Hunton results above). This latter finding may in part be due to a) the regression outputs reported in this sample of retracted papers generally report fewer variables and hence there are fewer digits to analyse, and b) very few of the retracted papers (20 out of our sample of 45) reported such analysis. Compare this with Hunton's papers which are much richer in terms of the digits and analysis reported. This potentially indicates a further limitation of applying BL to specific types of analysis.

6.6. Simulations

We undertake a number of simulations to examine alternative modus operandi of fraudulent behaviour by authors: a) employing a data generating process (DGP) to achieve a desired result, b) manually changing the output to provide the required level of significance and c) manually changing the underlying data set to achieve a desired output.

6.6.1. Employing DGP

As noted earlier, output from data generated via a DGP will not follow Benford Law. However, many argue that this is exactly the process some researchers have used to generate fabricated datasets. For example, Professor Schön's response to the fraud investigators was that he started from the conclusion he wanted and then assembled the dataset to show it (Reich, 2009). Similarly, Professor Hunton was also accused of using a similar modus operandi (Seadle, 2016). To provide evidence that the output generated from a DGP will not follow BL we obtain two naturally occurring datasets (X and Y) that are theoretically expected to be

highly correlated (Income and Share Price). We also create six datasets (3 X's and 3 Y's) using three different DGPs criteria to produce a pre-defined level of correlation between X and Y (beta normally distributed around 3, beta normally distributed, X and Y normally distributed with a correlation of 0.6). Using a bootstrap regression, utilizing two different bootstrap sample sizes 100 and 500, and repeating the regression 500 and 1000 times we obtain the output from each individual bootstrap regression. Table 11 reports the MAD scores for the outputs. We find irrespective of the DGP criteria and the number of bootstrap iterations, the MAD scores for the outputs (total, descriptives and regressions), significantly deviate relative to the output from the naturally occurring X and Y datasets. For example, repeating 1000 times when n=100, the DGP MAD for the reported regression output is approximately between 115% to 318% higher than regression output derived from a naturally occurring dataset.

<Insert Table 11>

6.6.2. Manually Adjusting Output

The second possible modus operandi is simply editing the regression output to a desired significance level. For example, replacing the "correct" / "truthful" betas with one that equals the standard error multiplied by a number in order to give the illusion that a specific significant level has been achieved (e.g. Beta = S.E.*1.96). To investigate whether BL can help identify this type of fraud we analyse the MAD scores of a regression output pre and post manipulation of the beta. To obtain the correct betas we use the same naturally occurring dataset used in 6.6.1 above. Using a similar bootstrap regression with three different bootstrap sample sizes (100, 500 and 600) we repeat a simple OLS regression 600 times. We replace 5%, 7% and 10% of the true betas with a beta that equals the standard error times 1.96 (beta = S.E * 1.96) thus resulting in a t-statistic of 1.96 (p-value=0.05)[12]. Table 12 reports the MAD scores for the unmanipulated and manipulated regression output. We find, for all the regression outputs, the

[12] We would like to thank one of our anonymous reviewers for this helpful suggestion.

manipulated MAD scores are between 1.24 and 1.65 times higher than the unmanipulated MAD scores.

<Insert Table 12>

6.6.3. Controlled Manipulation of Underlying Data

The third modus operandi, which we believe is probably the most common type of fraud, is to manipulate a percentage of the underlying dataset. Thus, in order to understand the MAD score and its ability to detect this type of manipulation we artificially create a typical set of numbers commonly found within empirical academic publications and subject the resulting output to a number of sensitivity tests. We first create an artificial sample of one dependent, and five independent variables – each containing 1,000 observations and importantly this underlying dataset is constructed in such a way that each variable conforms to BL. Using these six variables we then produce a set of descriptive statistics typically found in empirical research.[13] We then run a simple OLS regression on the sample and collect the output for coefficients[14]. This process is repeated 1,000 times so that we have the statistical output for 1,000 randomly-generated samples which closely conform to BL.

To derive a base-level *MAD* score we randomly choose *n* numbers from the statistical output ranging from *n*=10 to *n*=100,000 in order that we might gauge the effect of sample size on the *MAD* score under the base-case scenario. We do this via three groupings of output type; *Descriptive Statistics*, *Regression Output*, and a final group where type does not matter (*All Numbers*). Table 13 column 2 "*Base*" presents the results of this initial analysis. Each cell within the table represents the average *MAD* score for 100 iterations of the selection process to control for possible outliers. We observe a steady increase in the *MAD* score as the number of observations diminishes, and exponentially rises once the number of observations is below 100.

[13] Namely; sample size, mean, standard deviation, minimum, 25th quartile, median, 75th quartile, maximum, and correlation matrix

[14] Namely, t-statistics, p-values, adjusted R^2, and standard errors

The next stage of the analysis investigates the effect of controlled manipulation of a percentage of the underlying dataset (Xs and Y) and the extent that this has on the *MAD* score - relative to our base-case scenario. Specifically, we manipulate our underlying datasets (Xs and Y) by randomly assigning first digits ranging from 1 to 9, for differing decile levels ranging from 10% (100 numbers manipulated) to 100% (1000 numbers manipulated). Then in a similar process to obtaining the base-level figures, we randomly choose *n* numbers from the statistical output ranging from *n*=10 to *n*=100,000 although we also require that the first digit ranges between 1 and 9 and are uniformly distributed. The results of these manipulations are also shown in Table 13 where each cell represents the average *MAD* score for 100 iterations of the manipulative process. Shaded areas indicate where the level of manipulation has a *MAD* score significantly greater than the base-level at the 5% significance level.

<Insert Table 13>

We observe that in all three number groupings, the level of manipulation required in order to produce a significantly higher *MAD* score is inversely related to the number of digits used in its creation. For example, in Panel C we observe that for larger samples we only require a relatively small manipulation level of 10% in order to create a significantly larger *MAD* score - compared to smaller samples below *n*=50 – and to our smallest sample of *n*=10 where the level of manipulation required to be significantly greater is 80%. We also note that the level of manipulation required to become significantly greater varies between the groups. Descriptive statistics require much larger levels of manipulation than their regression counterparts – largely as a result of their base-levels being higher on average, but we also suggest that the type of descriptive chosen, along with the nature of the underlying sample data will also play a key role in determining the *MAD* score.

Given BL requires the data does not follow any pre-defined structure (e.g a DGP), by the very nature of this statistical exercise we have had to manually construct a dataset that

follows BL. We therefore do not present these results as hard statistical evidence of what a *MAD* score should be under various alternate conditions, but a mere example of how manipulation of data can change the *MAD* score from its base-state given one set of particular underlying circumstances.

7. CONCLUSION AND DISCUSSION

Overall our results provide empirical evidence that the level of conformity to Benford Law can be used to differentiate between academic papers that have been retracted due to concerns with the data and those papers that have not been retracted and are assumed to be credible and reliable. Therefore, potentially one could use the application of BL as a screening process to highlight potential 'red flags'.

Since fraudulent data is designed to elude all the self-correcting processes available to co-authors and reviewers, we recommend co-authors, not involved in the analysis/collection of data, should consider, *inter alia*, employing BL to screen their co-authors' data and outputs. This is particularly important if there is a high degree of division of labour between them. As Walsh et al. (2019) argues this division of labour increases the vulnerability of the project to pathologies, such as fraud and other QRPs. Specifically, Walsh et al. (2019) finds the likelihood that a paper will be retracted is positively associated with the degree of division of labour between co-authors, consistent with organisational theory of scientific pathologies. One way to mitigate this vulnerability is for co-authors to coordinate and verify the integrity of the research findings (Walsh et al, 2019). Our results provide one possible mechanism to do this. Author(s) could compare the conformity measure of their co-authors' analysis to a set of comparable published papers (for example papers they cite in their paper). If their results suggest a significantly higher deviation in the conformity measure relative to the control papers then this would provide a basis for a deeper discussion and a request for more detail from the co-author. We would however suggest authors ask their co-authors for as much analysis as

possible as our findings suggest BL is more discerning the richer the dataset, e.g. the simulation results and findings with regard to the generalizability results.

There are significant benefits of applying BL. First, the authors do not need to be experts in the field of their co-authors. They do not need to understand all the technical processes undertaken by their co-authors to determine the conformity measure of their output (except with respect to the scenarios under which conformity is unlikely to occur, noted in section 3.1). Second, many journals now require, prior to acceptance, that authors accept joint responsibility for the integrity of the data employed in their manuscripts, a lack of expertise in relation to the co-author's analysis is less likely to be a plausible excuse (unlike in prior retraction cases). Third, there are significant costs to authors whose papers are retracted as a result of a co-author's fraudulent behaviour, especially for more eminent authors (Azoulay et al., 2017). Even innocent prior collaborators suffer from 'stigmatization by association' and incur the cost of mistrust (Hussinger and Pellens 2019).

Although co-authors should be the first line of defence as they have access to technical details and analysis (Beasley et al., 2002), no system of checking is perfect. Therefore, journal editors may also wish to consider implementing a 'state of the art' detection software that would include not only plagiarism detection but also fraud/manipulation detection analytics which would include BL, in conjunction with other data tools such as those used by Carlisle (2012) and Bergh et al. (2017).[15] This software could be implemented as a screening process and if abnormalities are observed then authors would be asked to provide additional information with respect to the data along with codes or checklists detailing their various methodological choices (Gall and Maniadis, 2019) before acceptance or review of the manuscript. However, as with any detection system there may be unintended consequences,

[15] We would like to thank one of our anonymous reviewers for this insightful comment.

for example Gall and Maniadis (2019) suggests making scientific fraud prohibitive may potentially make minor QRPs more attractive.

As we acknowledged earlier, the use of BL in this context is not without its limitations and its specificity remains to be firmly determined. Therefore, we can only use the application of BL as a screening process to highlight potential 'red flags'. It should instigate the start of the conversation, not end it.

We hope future work however, based on these findings, will investigate the possibility of applying Artificial Intelligence to increase the specificity of BL and provide a golden criteria with which to judge deviation or compliance. For example, specific algorithms could potentially be generated to measure conformity on a paper-per-paper basis, taking into account the characteristics of each paper and author teams, thereby minimising the likelihood of Type I and Type II errors. For information the *MAD* score for this manuscript is 0.006932.

REFERENCES

Almer, E. D., M. Bertolini, and J. L. Higgs. 2013. A model of individual accounting faculty salaries. *Issues in Accounting Education* 28 (3): 411–433.

Almer, E. D., A. A. Baldwin, L. A. Jones-Farmer, M. Lightbody, and L. E. Single. 2015. Tenure track opt-outs: Leakages from the academic pipeline. In Advances in Accounting Education: Teaching and Curriculum Innovations, edited by T. J. Rupert and B. B. Kern, 1–36. Bingley, U.K.: Emerald.

AAA (American Accounting Association). 2015. Available at http://aaahq.org/Sections-Regions/Accounting-Information-Systems/Journals-Publications.

AAA (American Accounting Association) Publication Ethics Policy. 2014. Available at: http://commons.aaahq.org/groups/2371c0896a/summary.

Amiram, D., Z. Bozanic, and E. Rouen. 2015. Financial statement errors: Evidence from the distributional properties of financial statement numbers. *Review of Accounting Studies* 29(4): 1540-1593.

Anderson, R., W. Greene, B.D. McCullough, and H.D. Vinod. 2008. The role of data/code archives in the future of economic research. *Journal of Economic Methodology* 15(1): 99-119.

Al-Marzouki, S., S. Evans, T. Marshall, and I. Roberts. 2005. Are these data real? Statistical methods for the detection of data fabrication in clinical trials. *British Medical Journal* 331(7511): 267-270.

Arshadi, L., and A.H. Jahangir. 2014. Benford's law behaviour of Internet traffic. *Journal of Network and Computer Applications* 40: 194-205.

Azoulay, P., A. Bonatti, and J.L. Krieger. 2017. The career effects of scandal: Evidence from scientific retractions. *Research Policy* 40: 1552-1569.

Banks, D. 2000. Get M.A.D with numbers! Moving Benford's law from art to science. *Fraud Magazine*, September/October.

Bailey, C., J.R. Hasselback, and J.N. Karcher. 2001. Research misconduct in accounting literature: A survey of the most prolific researchers' actions and beliefs. *Abacus* 37(1): 26-54.

Bailey, C. 2015. Psychopathy, academic accountants' attitudes toward unethical research practices, and publication success. *The Accounting Review* 90(4): 1307-1332.

Bauer, J., and J. Gross. 2011. Difficulties detecting fraud? The use of Benford's law on regression tables. *Jahrbücher für Nationalökonomie und Statistik* 231(5/6): 733-748.

Beasley, M.R., Datta, S., Kogelnik, H., Kroemer, H., and Monroe, D. 2002. Report of the Investigation Committee on the Possibility of Scientific Misconduct in the Work of Hendrik Schön and Coauthors. Lucent Technologies.

Bedeian, A. G., Taylor, S. G., and A. N. Miller. 2010. Management science on the credibility bubble: Cardinal sins and various misdemeanours. *Academy of Management Learning & Education 9*: 715-725.

Begley, C.G., and L.M. Ellis. 2012. Drug development: Raise standards for preclinical cancer research. *Nature* 483(7391): 531–533.

Bergh, D.D., Sharp, B.M. and Li, M., 2017. Tests for identifying "Red flags" in empirical findings: demonstration and recommendations for authors, reviewers, and editors. *Academy of Management Learning & Education* 16: 110–124.

Berton, L. 1995. He's got their number: Scholar uses math to foil financial fraud. *Wall Street Journal*, July.

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

Bentley University. 2014. *Report of Judith A. Malone, Bentley University Ethics Officer, Concerning Dr. James E. Hunton*. Downloaded from https://www.bentley.edu/files/Hunton%20report%20July21.pdf

Bhattacharya, S., D. Xu, and K. Kumar. 2011. An ANN-based auditor decision support system using Benford's law. *Decision Support System*s 50: 576-584.

Bhojraj, S., and R. Libby. 2005. Capital market pressure, disclosure frequency-induced earnings/cash flow conflict, and managerial myopia. *The Accounting Review* 80(1): 1-20.

Bierstaker, J., R. Brody, and C. Pacini. 2006. Accountants' perceptions regarding fraud detection and prevention methods. *Managerial Auditing Journal* 21(5): 520-535.

Boyle, J. 1994. An application of fourier series to the most significant digit problem. *The American Mathematical Monthly* 101: 879–886.

Burke, J., and E. Kincanon. 1991. Benford's law and physical constants: The distribution of initial digits. *American Journal of Physics* 59: 952.

Busta, B., and Weinberg, R. 1998. Using Benford's Law and neural networks as a review procedure. *Managerial Auditing Journal* 13(6): 356-366.

Camerer, C.F. 2003. Behavioural studies of strategic thinking in games. *Trends in Cognitive Sciences* 7(5): 225-231.

Carslaw, C. 1988. Anomalies in income numbers: Evidence of goal oriented behavior. *The Accounting Review* 63(2): 321-327.

Carlisle, J.B. 2012. The analysis of 168 randomised controlled trials to test data integrity. *Anaesthesia* 67(5): 521-537.

Clarke, R. 2006. Plagiarism by academics: More complex than it seems. *Journal of the Association for Information Systems* 7: 91-121.

Clippe, P., and M. Ausloos. 2012. Benford's law and Theil transform of financial data. *Physica A: Statistical Mechanics and its Applications* 391(24): 6556–6567.

Collberg, C., and S. Kobourov. 2005. Self-Plagiarism: In Computer Science. *Communications of The ACM* 48(4): 88-94.

Cox, A., R. Craig, and D. Tourish. 2018. Retraction statements and research malpractice in economics. *Research Policy* 47: 924-935.

Craig, R., J. Amernic, and D. Tourish. 2014. Perverse audit culture and the modern public university. *Financial Accountability & Management* 30(1): 1-24.

Crain, J.L., and P.J. Carruth. 1992. Academic accounting research: Opinions of academicians on recommendations for improving ethical behaviour. *The Accounting Educators' Journal* 4: 27-46.

Delwald, W G; J.G. Thursby and R.G.Anderson. 1986. Replication in empirical economics: The Journal of Money, Credit and Banking Project. *The American Economic Review* 76(4): 587-603.

Desruisseaux, P. 1999. Cheating is reaching epidemic proportions worldwide, researchers say. *The Chronicle of Higher Education* 45: A45

Diekmann, A. 2002. Diagnose von fehlerquellen und methodische qualita ̈t in der sozial-wissenschaftlichen forschung. Manuskript 06/2002, Institut fur Technikfolgenabschätzung (ITA). Wien

Diekmann, A. 2007. Not the first digit! Using Benford's Law to detect fraudulent scientific data. *Journal of Applied Statistics* 34(3): 321-329.

Diekmann, A., and B. Jann. 2010. Benford's Law and fraud detection. Facts and legends. *German Economic Review* 11(3): 397-401.

Dickins, D., and D.K. Schneider. 2016. Academic research in accounting: A Framework for quality reviews. *Current Issues in Auditing* 10(1): A34-A46.

Durtschi, C., W. Hillison, and C. Pacini. 2004. The effective use of Benford's law to assist in the detecting of fraud in accounting data. *Journal of Forensic Accounting* 5: 17-34.

Enders, W., and G. Hoover. 2006. Plagiarism in the economics profession: A survey. *Challenge* 49: 92-107.

Fanelli, D. 2009. How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PloS one* 4(5): e5738.

Foo, J.Y.A., and X.J.A. Tan. 2014. Analysis and implications of retraction period and coauthorship of fraudulent publications. *Accountability in Research* 21(3): 198-210.

French, C., 2012. Precognition studies and the curse of the failed replications. Guardian Newspaper (https://www.theguardian.com/science/2012/mar/15/precognition-studies-curse-failed-replications) Accessed 17 October 2018.

Gall,T., and Z. Maniadis. 2019. Evaluating solutions to the problem of false positives. *Research Policy* 48: 506-515.

Giles, D.E.A. 2007. Benford's Law and naturally occurring prices in certain e-Bay auctions. *Applied Economics Letters* 14:157-161.

Gill, J., and K.J. Meier. 2000. Public administration research and practice: A methodological manifesto. *Journal of Public Administration Research and Theory* 10(1): 157-199.

Glover, S.M., D. F., Prawitt, S. L., Summers, and D.A. Wood. 2012. Publication benchmarking data based on faculty promoted at the top 75 U.S. accounting research institutions. *Issues in Accounting Education* 27(3): 647-670.

Glover, S. M., D. F. Prawitt, and D. A., Wood. 2006. Publication records of faculty promoted at the top 75 accounting research programs. *Issues in Accounting Education* 21(3): 195- 218.

Graber, M., and A.L. Walde. 2008. Publish or perish? The increasing importance of publications for prospective professors in Austria, Germany and Switzerland. *German Economic Review* 9: 457-472.

Grienensen, M.L., and M. Zhang. 2012. A comprehensive survey of retracted articles from the scholarly literature. *PLoS One* 7(10): e44118.

Günnel, S., and K-H. Tödter. 2009. Does Benford's Law hold in economic research and forecasting? *Empirica: Journal of Applied Economics and Economic Policy* 36(3): 273-292.

Hales, D., S. Chakravorty, and V. Sridharan. 2009. Testing Benford's Law for improving supply chain decision-making: a field experiment. *International Journal of Production Economics* 122(2): 606-618.

Hamermesh, D S. 2007. Viewpoint: Replication in economics. *Canadian Journal of Economics* 40(3): 715-733.

Healy, B. 2012. Bentley professor resigns after his research is retracted. *The Boston Globe*, 21st December 2012.

Hill, T.P. 1995. A statistical derivation of the significant digit law. *Statistical Science* 10: 354–363.

Hill, T.P. 1998. The first digital phenomenon: A century-old observation about an unexpected pattern in many numerical tables applies to the stock market, census statistics and accounting data. *American Scientist* 86(4): 358-363.

Honig, B., and A. Bedi. 2012. The Fox in the Hen House: A critical examination of plagiarism among members of the Academy of Management. *Academy of Management Learning & Education* 11(1): 101–123.

Huang, S-M., D. Yen L-W. Yang, and J-S. Hua. 2008. An investigation of Zipf's Law for fraud detection. *Decision Support Systems* 46(1): 70-83.

Hubbard, D., and D. Vetter. 1996. An empirical comparison of published replication research in accounting, economics, finance, management and marketing. *Journal of Business Research* 35: 153-164.

Hunton, J.E., and A. Gold. 2010. A field experiment comparing outcomes of three fraud brainstorming procedures: Nominal groups, round robin, and open discussion (Retracted). *The Accounting Review* 85(3): 911-935.

Hussinger, K., and M. Pellens. 2019. Guilt by association: How scientific misconduct harms prior collaborators, *Research Policy* 48(2): 516-530.

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

John, L K., G. Loewenstein, and D. Prelec. 2012. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science* 23(5): 524-532.

Karabag, S., and C. Berggren. 2012. Retraction, dishonesty and plagiarism: Analysis of a crucial issue for academic publishing, and the inadequate responses from leading journals in economics and management disciplines. *Journal of Applied Economics and Business Research* 2(3): 172-183.

Knuth, D. 1969. *The Art of Computer Programming*, *Vol. 2*, Addison-Wesley, New York: 219-229.

Kock, N. 1999. A case of academic plagiarism. *Communications of the ACM*, 42, pp.96–104.

Kubovy M. 1977. Response availability and the apparent spontaneity of numerical choices. *Journal of Experimental Psychology: Human Perception and Performance* 2: 359-364.

Kubovy, M. 1977. A possible basis for conservatism in signal detection and probabilistic categorization tasks. *Perception & Psychophysics* 22(3): 277-281.

Lacetera, N., and L. Zirulia. 2011. The economics of scientific misconduct. *Journal of Law, Economics, & Organization* 27(3): 568-603.

Leibon, G. 2008. Google Numbers, *chance news*, Downloaded from https://www.dartmouth.edu/~chance/chance_news/for_chance_news/ChanceNews13.03/GregProject.pdf.

Ley, E. 1996. On the peculiar distribution of the US stock indexes' digits. *American Statistician* 50: 311-313.

Ley, E., and H. Varian. 1994. Are there psychological barriers in the Dow-Jones Index? *Applied Financial Economics* 4: 217-224.

List, J A., C.D. Bailey, P.J. Euzent, and T.L. Martin. 2001. Academic economists behaving badly? A survey on three areas of unethical behavior. *Economic Inquiry* 39 (1): 162-70.

Loeb, S. E., and B. D. Merino. 2000. A discussion of accounting academic ethics. *Research on Accounting Ethics* 6: 293-311.

Jin, G.Z, B. Jones, S.F. Lu, and B. Uzzi. 2013. The Reverse Matthew effect: Catastrophe and consequence in Science, NBER Working Paper No. 19489.

Johnson, G., and J. Weggenmann. 2013. Exploratory research applying Benford's Law to selected balances in the financial statements of state governments. *Academy of Accounting and Financial Studies Journal* 17 (3): 31–44.

Marcovitch, Harvey. "Misconduct by researchers and authors." *Gaceta sanitaria* 21.6 (2007): 492-499.

Martin, G.N., and R.M. Clarke. 2017. Are psychology journals anti-replication? A snapshot of editorial practices. *Frontiers in Psychology 8*, 523. http://doi.org/10.3389/fpsyg.2017.00523.

Meyer, M., R.W. Waldkirch, I. Duscher, and A. Just. 2018. Drivers of citations: An analysis of publications in ''top" accounting journals. *Critical Perspectives on Accounting* 51: 24-46.

McCullough, B.D. 2009. Open access economics journals and the market for reproducible economic research. *Economic Analysis and Policy* 39(1): 117-126.

McCullough, B.D., and H.D. Vinod. 2003. Verifying the solution from a nonlinear solver: A case study. *American Economic Review* 93: 873-892.

McCullough, B.D., K.A. McGeary, and T.D. Harrison. 2006. Lessons from the JMCB Archive. *Journal of Money, Credit, and Banking* 38(4): 1093-1107.

McCullough, B.D., K.A. McGeary, and T.D. Harrison. 2008. Do economics journal archives promote replicable research? *Canadian Journal of Economics* 41(4): 1406-1420.

McNay, l. 2016. Imbalancing the academy: the impact of research quality assessment. *Sociologia Italiana - AIS Journal of Sociology* 8(7): 119-150.

Mongeon, P. and V. Larivière. 2016. Costly Collaborations: The Impact of Scientific Fraud on Co-Authors' Careers. *Journal of the Association for Information Science and Technology*, 67(3): 535–542.

Morrow, J. 2014. Benford's Law, families of distributions and a test bias. CEP Discussion Paper No. 1291, August, ISSN 2042-2695.

Necker, S. 2014. Scientific misbehaviour in economics. *Research Policy* 43:1747-1759.

Newcomb, S. 1881. Note on the Frequency of the Different Digits in Natural Numbers. *American Journal of Mathematics* 4(1): 39-40.

Nigrini, M. 1996. A taxpayer compliance application of Benford's law. *Journal of the American Taxation Association* 18: 72-91.

Nigrini, M. 1997. Digital Analysis Tests and Statistics. The Nigrini Institute Inc., Allen, Texas.

Nigrini, M. 1999. Adding value with digital analysis. *The Internal Auditor* 56(1): 21-23.

Nigrini, M. 2012. *Benford's law: Applications for forensic accounting, auditing, and fraud detection*. Hoboken, N.J.: Wiley.

Nigrini, M., and S. Miller. 2009. Data diagnostics using second-order test of Benford's law. *Auditing: A Journal of Practice and Theory* 28(2): 305–324.

Nigrini, M., and L.J. Mittermaier. 1997. The use of Benford's law as an aid in analytical procedures. *Auditing: A Journal of Practice & Theory* 16(2): 52-67.

Petersen, M.A. 2009. Estimating standard errors in finance panel data sets: Comparing approaches. *The Review of Financial Studies* 22(1): 435-480.

Pietronero, L., E. Tosatti, V. Tosatti, and A. Vespignani. 2001. Explaining the uneven distribution of numbers in nature: The Laws of Benford and Zipf. *Physica A: Statistical Mechanics and its Applications* 293(1–2): 297–304.

Pinkham, S. 1961. On the distribution of first significant digits. *The Annals of Mathematical Statistics* 32(4): 1223-1230.

Reich, E. S. (2009). The rise and fall of a physics fraudster. *Physics World*, *22*(05), 24.

Quick, R., and M. Wolz. 2003. Benford's Law in deutschen Rechnungslegungsdaten. *Betriebswirtschaftliche Forschung und Praxis*: 208–224.

Seadle, M. (2016). Quantifying research integrity. Synthesis Lectures on Information Concepts, Retrieval, and Services, 8(5), 1-141.

Seybert, N. 2010. R&D capitalization and reputation-driven real earnings management. *The Accounting Review* 85(2):671-693.

Schäfer, C., K. Schräpler, R. Müller, and G. Wagner. 2005. Automatic identification of faked and fraudulent interviews in survey by two different methods. *Journal of Applied Science Studies* 125: 119-129.

Schräpler, J., and G. Wagner. 2005. Characteristics and impact of faked interviews in surveys. *Allgemeines Statistisches Archive* 89: 79-120.

Shiffler, R E and P.D. Harsha. 1980. Upper and lower bounds for the sample standard deviation. *Teaching Statistics: An International Journal for Teachers* 2(3):84-86.

Slepkov, A.D., K.B. Ironside, and D. DiBattista. 2015. *Benford's Law: Textbook Exercises and Multiple-Choice Testbanks*. PLoS ONE 10(2):e0117972. doi:10.1371/journal.pone.0117972.

Tödter, K. 2015. Benford's Law and Fraud in Economic Research. Chapter 12, *Benford's Law: Theory and Applications*, Editor S. J. Miller. Princeton University Press.

Varian, H. 1972. Benford's law. *American Statistics* 23: 65-66.

Walsh, J P, Y-N. Lee and L. Tang. 2019. Pathogenic organization in science: Division of labour and retractions. *Research Policy* 48: 444-461.

Watrin, C., R. Struffert, and R. Ullmann. 2008. Benford's law: an instrument for selecting tax audit targets. *Review of Managerial Science* 2(3): 219-237.

Wright, A., and R. Ashton. 1989. Identifying audit adjustments with attention-directing procedures. *The Accounting Review* 64: 710-728.

Figure 1. The average values of *MAD* for each of the sample groups plotted over time.

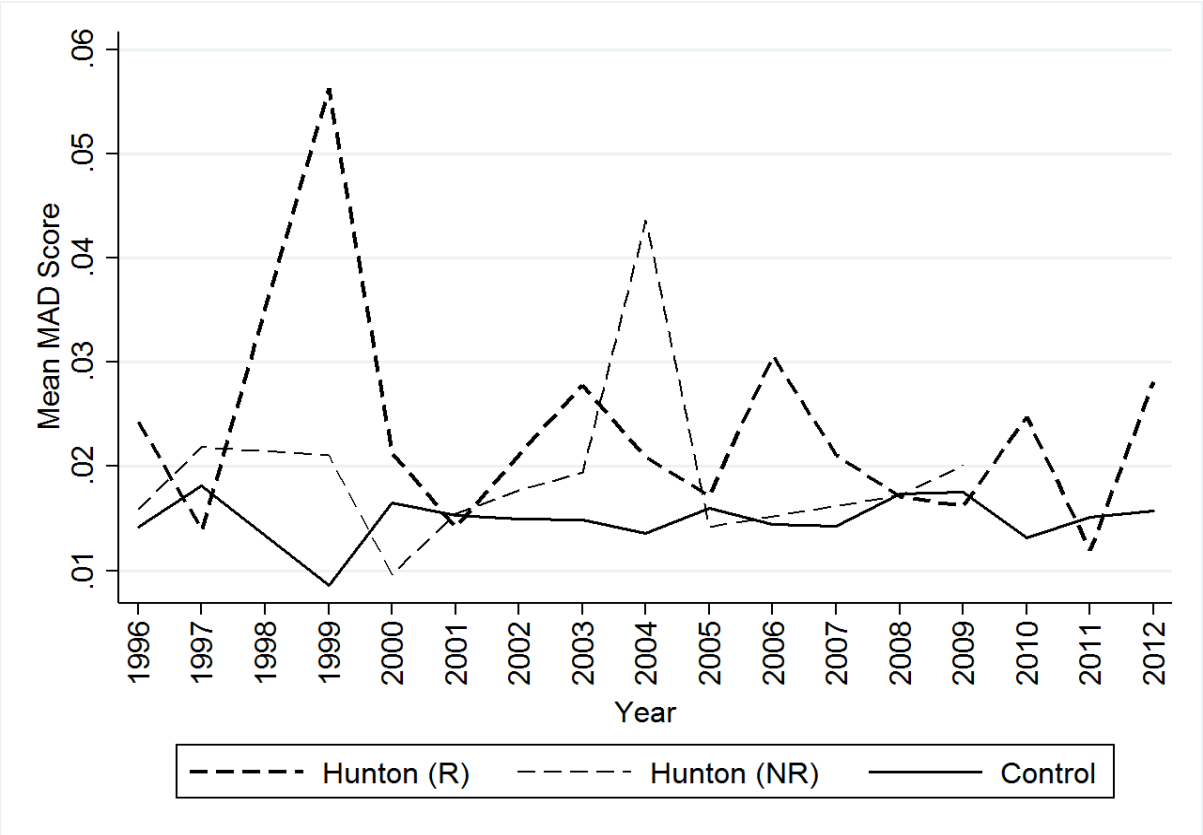

Table 1: Sample Size by Group (Hunton's Retracted, Non-Retracted Articles, Control Group), Year and Journal Ranking.

By Year	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	Total
Hunton_R	2	1	0	1	1	3	5	1	1	3	3	2	3	1	6	2	2	**37**
Control Group	7	6	0	0	1	7	5	5	2	18	10	5	24	0	40	16	10	**156**
Hunton_NR	2	3	0	1	2	3	1	1	2	1	0	0	1	1	0	0	0	**18**
Control Group	5	21	0	1	8	25	8	2	3	5	0	0	3	9	0	0	0	**90**
Total	**16**	**31**	**0**	**3**	**12**	**38**	**19**	**9**	**8**	**27**	**13**	**7**	**31**	**11**	**46**	**18**	**12**	**301**

By Journal Ranking	1	2	3	4	4*	Total
Hunton_R	8 (22%)	3 (8%)	12 (32%)	3 (8%)	11 (30%)	37
Control Group	21 (13%)	8 (5%)	35(22%)	25 (16%)	67 (43%)	156
Hunton_NR	4 (22%)	7 (39%)	2 (11%)	1 (6%)	4 (22%)	18
Control Group	6 (7%)	23 (26%)	17 (19%)	21 (23%)	23 (26%)	90

Table 2. List of retracted papers attributable to Hunton

	Title	Year	Journal	Rank	Authors	Type
1	Hierarchical and gender differences in private accounting practice	1996	Accounting Horizons	3	2	Survey
2	Performance of accountants in private industry: A survival analysis	1996	Accounting Horizons	3	1	Database
3	An Assessment of the Relation between Analysts' Earnings Forecast Accuracy, Motivational Incentives and Cognitive Information Search Strategy	1997	The Accounting Review	4*	1	Experimental
4	Is analyst forecast accuracy associated with accounting information use?	1999	Accounting Horizons	3	1	Experimental
5	The impact of electronic commerce assurance on financial analysts' earnings forecasts and stock price estimates	2000	Auditing: A Journal of Theory and Practice	3	3	Survey
6	Linking participative budgeting congruence to organization performance	2001	Behavioral Research in Accounting	3	1	Questionnaire
7	Mitigating the common information sampling bias inherent in small-group discussion	2001	Behavioral Research in Accounting	3	0	Experimental
8	The effects of small monetary incentives on response quality and rates in the positive confirmation of account receivable balances	2001	Auditing: A Journal of Theory and Practice	3	1	Questionnaire
9	Analysts' Reactions to Earnings Preannouncement Strategies	2002	Journal of Accounting Research	4*	2	Experimental
10	Investigating the Impact of Auditor-Provided Systems Reliability Assurance on Potential Service Recipients.	2002	Journal of Information System	1	1	Experimental
11	Promotion and performance evaluation of managerial accountants	2002	Journal of Management Accounting Research	2	2	Survey
12	Sampling practices of auditors in public accounting, industry, and government	2002	Accounting Horizons	3	2	Survey
13	The Reaction of Financial Analysts to Enterprise Resource Planning (ERP) Implementation Plans.	2002	Journal of Information System	1	2	Experimental
14	Extending the Accounting Brand to Privacy Services	2003	Journal of Information System	1	1	Experimental
15	Are Financial Auditors Overconfident in Their Ability to Assess Risks Associated with Enterprise Resource Planning Systems?	2004	Journal of Information System	1	2	Experimental
16	Behavioural Self-Regulation of Telework Locations: Interrupting Interruptions!	2005	Journal of Information System	1	0	Experimental
17	Capital Market Pressure, Disclosure Frequency-Induced Earnings/Cash Flow Conflict, and Managerial Myopia	2005	The Accounting Review	4*	2	Experimental
18	Does Graduate Business Education Contribute to Professional Accounting Success?	2005	Accounting Horizons	3	2	Survey
19	Does the Form of Management's Earnings Guidance Affect Analysts' Earnings Forecasts?	2006	The Accounting Review	4*	2	Experimental
20	Financial Reporting Transparency and Earnings Management	2006	The Accounting Review	4*	2	Experimental
21	Recognition v. Disclosure, Auditor Tolerance for Misstatement, and the Reliability of Stock-Compensation and Lease Information	2006	Journal of Accounting Research	4*	2	Experimental
22	Enterprise resource planning systems and non-financial performance incentives: The joint impact on corporate performance	2007	International Journal of Accounting Information Systems	2	2	Database
23	The Potential Impact of More Frequent Financial Reporting and Assurance: User, Preparer, and Auditor Assessments	2007	Journal of Emerging Technologies in Accounting	1	2	Experimental
24	Can directors' self-interests influence accounting choices	2008	Accounting, Organizations and Society	4*	1	Experimental
25	Potential Functional and Dysfunctional Effects of Continuous Monitoring	2008	The Accounting Review	4*	2	Experimental
26	Relationship Incentives and the Optimistic/Pessimistic Pattern in Analysts' Forecasts	2008	Journal of Accounting Research	4*	3	Experimental
27	The Impact of Client and Auditor Gender on Auditors' Judgments	2009	Accounting Horizons	3	2	Experimental
28	A Field Experiment Comparing the Outcomes of Three Fraud Brainstorming Procedures: Nominal Group, Round Robin and Open Discussion	2010	The Accounting Review	4*	1	Experimental
29	Continuous monitoring and the status quo effect	2010	International Journal of Accounting Information Systems	2	2	Experimental
30	Decision Aid Reliance: A Longitudinal Field Study Involving Professional Buy-Side Financial Analysts	2010	Contemporary Accounting Research	4	2	Database
31	R&D Capitalization and Reputation-Driven Real Earnings Management	2010	The Accounting Review	4*	1	Survey
32	The Impact of Alternative Telework Arrangements on Organizational Commitment: Insights from a Longitudinal Field Experiment	2010	Journal of Information System	1	1	Experimental
33	When Do Analysts Adjust for Biases in Management Guidance? Effects of Guidance Track Record and Analysts' Incentives	2010	Contemporary Accounting Research	4	2	Experimental
34	The Influence of Corporate Governance Ratings on Buy-Side Analysts' Earnings Forecast Certainty: Evidence from the United States and the United Kingdom	2011	Behavioral Research in Accounting	3	2	Experimental
35	The Relationship between Perceived Tone at the Top and Earnings Quality	2011	Contemporary Accounting Research	4	2	Survey
36	The Dark Side of Online Knowledge Sharing	2012	Journal of Information System	1	3	Survey
37	Will corporate directors engage in bias arbitrage to curry favor with shareholders?	2012	Journal of Accounting and Public Policy	3	1	Experimental

This table presents the list of retracted papers attributable to James E. Hunton. Papers 1-9 report Hunton as being employed by the University of South Florida, and papers 10-37 by Bentley University. Papers 17 and 31 were not written by Hunton directly but he was noted as supplying data for the studies. All papers are classified by Retraction Watch as Hunton retractions.

Table 3: Descriptive Statistics

Panel A: Descriptive Statistics for Hunton's Retracted Articles and Retracted Control Sample

		Hunton Retracted Sample (n=37)					Control Sample (n=156)					Difference	
Variable		**Mean**	**Std. Dev**	**Median**	**25th Q**	**75th Q**	**Mean**	**Std. Dev.**	**Median**	**25th Q**	**75th Q**	**t-stat**	**Wilcoxon**
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
By Group:													
All Digits	***Dependent Variables:***												
	MAD	0.0221	0.0116	0.0209	0.0126	0.0280	0.0148	0.0061	0.0148	0.0100	0.0181	5.33***	3.46***
	FSD-Score	0.1079	0.0656	0.0851	0.0586	0.1518	0.0798	0.0416	0.0726	0.0531	0.0935	3.26***	1.98**
	KS1	0.4324	0.5022	0.0000	0.0000	1.0000	0.2756	0.4482	0.0000	0.0000	1.0000	1.86*	1.85*
Descriptive Digits													
	MAD	0.0286	0.0156	0.0241	0.0176	0.0330	0.0212	0.0109	0.0191	0.0132	0.0253	3.34***	4.52***
	FSD-Score	0.1387	0.0814	0.1284	0.0773	0.1695	0.1119	0.0650	0.0976	0.0621	0.1366	2.14**	1.95*
	KS1	0.5675	0.5022	1.0000	0.0000	1.000	0.3012	0.4602	0.0000	0.0000	1.0000	3.10***	3.04***
Regression Digits													
	MAD	0.0306	0.0176	0.0258	0.0166	0.0375	0.0221	0.0121	0.0194	0.0143	0.0263	3.38***	3.09***
	FSD-Score	0.1571	0.0927	0.1292	0.0907	0.1989	0.1156	0.0716	0.0984	0.0653	0.1359	2.90***	2.90***
	KS1	0.3429	0.4815	0.0000	0.0000	1.0000	0.1866	0.3909	0.0000	0.0000	0.0000	2.03**	2.02**
	Control Variables:												
	Pool	5.6543	0.5419	5.5797	5.3844	5.9135	5.6987	0.6804	5.7021	5.2652	6.0637	0.72	0.72
	Num_Authors	2.6486	0.7155	3.0000	2.0000	3.0000	2.3205	0.9015	2.0000	2.0000	3.0000	2.06**	1.99**
	Experimental	0.6486	0.4839	1.0000	0.0000	1.0000	0.6602	0.4751	1.0000	0.0000	1.0000	0.13	0.13
	Survey	0.2162	0.4173	0.0000	0.0000	0.0000	0.1666	0.3738	0.0000	0.0000	0.0000	0.70	0.74
	Questionnaire	0.0541	0.2292	0.0000	0.0000	0.0000	0.0192	0.1377	0.0000	0.0000	0.0000	1.19	1.19
	Database	0.0810	0.2800	0.0000	0.0000	0.0000	0.1538	0.3619	0.0000	0.0000	0.0000	1.14	1.14
	Linear	0.8918	0.3148	1.0000	1.0000	1.0000	0.8269	0.3795	1.0000	1.0000	1.0000	0.96	0.96

Panel B: Descriptive Statistics for Hunton's Non-Retracted Articles and Non-Retracted Control Sample

		Hunton Non-Retracted Sample (n=18)					Control Sample (n=90)					Difference	
	Variable	**Mean** (1)	**Std. Dev.** (2)	**Median** (3)	**25th Q** (4)	**75th Q** (5)	**Mean** (6)	**Std. Dev.** (7)	**Median** (8)	**25th Q** (9)	**75th Q** (10)	**t-stat** (11)	**Wilcoxon** (12)
By Group:													
All Digits	***Dependent Variables:***												
	MAD	0.0200	0.0120	0.0180	0.0158	0.0211	0.0167	0.0066	0.0168	0.0119	0.0205	1.65*	1.03
	FSD-Score	0.1068	0.0520	0.0991	0.0823	0.1322	0.0846	0.0471	0.0745	0.0522	0.1026	1.79**	2.02**
	KS1	0.4444	0.5113	0.0000	0.0000	1.0000	0.1667	0.3747	0.0000	0.0000	0.0000	2.69***	2.61***
Descriptive Digits													
	MAD	0.0278	0.0152	0.0188	0.0188	0.0328	0.0218	0.0101	0.0209	0.0124	0.0282	2.12**	1.61
	FSD-Score	0.1322	0.0533	0.1297	0.1062	0.1668	0.1097	0.0626	0.0966	0.0643	0.1346	1.42	2.07**
	KS1	0.2777	0.4608	0.0000	0.0000	1.0000	0.1888	0.3936	0.0000	0.0000	0.0000	0.85	0.85
Regression Digits													
	MAD	0.0321	0.0154	0.0294	0.0216	0.0389	0.0271	0.0124	0.0241	0.0180	0.0338	1.45	1.61
	FSD-Score	0.1688	0.0848	0.1493	0.0989	0.2204	0.1369	0.0750	0.1143	0.0848	0.1751	1.58	1.67
	KS1	0.4444	0.5113	0.0000	0.0000	1.0000	0.0886	0.2859	0.0000	0.0000	0.0000	4.03***	3.74***
	Control Variables:												
	Pool	5.6018	0.6719	5.4930	5.2882	5.8749	5.1789	0.7188	5.1239	4.6913	5.7200	2.30**	2.17**
	Num_Authors	2.7222	1.1274	2.5000	2.0000	2.0000	2.4111	1.0694	2.0000	2.0000	2.0000	1.11	1.25
	Experimental	0.7777	0.4278	1.0000	1.0000	1.0000	0.9111	0.2861	1.0000	1.0000	1.0000	1.64	1.64
	Survey	0.0555	0.2357	0.0000	0.0000	0.0000	0.0333	0.1805	0.0000	0.0000	0.0000	0.45	0.45
	Questionnaire	0.0555	0.2357	0.0000	0.0000	0.0000	0.0111	0.1054	0.0000	0.0000	0.0000	1.27	1.27
	Database	0.1111	0.3233	0.0000	0.0000	0.0000	0.0444	0.2072	0.0000	0.0000	0.0000	1.12	1.12
	Linear	1.0000	0.0000	1.0000	1.0000	1.0000	0.8333	0.3747	1.0000	1.0000	1.0000	1.87*	1.86*

Panel C: Correlation Matrix among Conformity Measures, Hunton Papers and Article characteristics.

	1	2	3	4	5	6	7	8	9	10	11
1 *MAD*		0.6404*	0.4162*	0.1588	0.0483	0.0942	0.1911*	-0.0825	0.0089	-0.1994*	0.0103
2 *FSD-Score*	0.7797*		0.6500*	0.1716	-0.0606	0.1880	0.1790	-0.1291	0.0387	-0.1620	-0.1072
3 *KS1*	0.5009*	0.7241*		0.2529*	0.3281*	0.1126	0.0400	0.0177	0.0963	-0.1262	0.0781
4 *Hunton*	0.3599*	0.2296*	0.1340		0.2167*	0.1078	-0.1581	0.0439	0.1229	0.1085	0.1796
5 *Ln_Pool*	-0.3711*	-0.3326*	0.0907	-0.0522		-0.0654	-0.0310	0.0223	0.0358	0.0031	0.3560*
6 *Num_Authors*	0.1887*	0.2243*	0.2106*	0.1477*	-0.1071		-0.1218	0.0068	0.0686	0.1211	0.0484
7 *Experimental*	0.0632	0.1157	-0.0195	-0.0096	-0.2177*	-0.0961		-0.5547*	-0.3885*	-0.6860*	-0.1420
8 *Survey*	0.0537	0.1049	0.1655*	0.0512	0.1405	0.0927	-0.6415*		-0.0269	-0.0476	0.0788
9 *Questionnaire*	-0.0023	-0.0455	-0.0374	0.0863	-0.1222	-0.0342	-0.2262*	-0.0754		-0.0333	0.0552
10 *Database*	-0.1444*	-0.2526*	-0.1379	-0.0826	0.1993*	0.0452	-0.5594*	-0.1865*	-0.0658		0.0974
11 *Linear*	0.0066	-0.0023	0.0146	0.0697	0.1532*	-0.0180	0.1904*	-0.1311	-0.0175	-0.1083	

This table presents the summary statistics, difference tests and correlations of, and between, the different conformity measures, groups, and article characteristics. *, **, *** represent significance level of 10%, 5% and 1%, (two-tailed) respectively. For the correlation matrix, below the diagonal we report correlations for Hunton retracted analysis and above the diagonal Hunton's non-retracted analysis.

Table 4: Comparing Conformity to Benford Law (BL) of Hunton's Retracted Articles relative to their Control Group.

Conformity Measure (MAD or FSD-Score or KS1) $= \alpha + \beta_1 Hunton\_R + \beta_2 Pool + \beta_3 Num\_Authors + \beta_4 Experimental + \beta_5 Survey + \beta_6 Questionanaire + \beta_7 Linear +$ *Year F.E. + Journal F.E.*

Conformity Measure *MAD, FSD-Score, and KS1*

	MAD			*FSD-Score*			*KS1*		
	All Digits	Descriptive	Regression	All Digits	Descriptive	Regression	All Digits	Descriptive	Regression
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Hunton_R	**0.006*****	**0.007*****	**0.005****	**0.023*****	**0.025****	**0.023***	0.453	**1.286*****	**1.290****
	(4.65)	**(3.01)**	**(2.04)**	**(2.70)**	**(2.02)**	**(1.70)**	(0.88)	**(2.73)**	**(2.38)**
Ln_Pool	-0.005***	-0.007***	-0.007***	-0.026***	-0.042***	-0.038***	0.366	0.491*	1.220***
	(-5.39)	(-5.68)	(-6.09)	(-4.46)	(-5.69)	(-5.68)	(0.95)	(1.67)	(3.60)
Num_Author	0.001	-0.001	0.002	0.008*	0.004	0.007	0.784***	0.178	0.318
	(0.88)	(-0.21)	(1.52)	(1.94)	(0.77)	(1.10)	(2.93)	(0.82)	(1.19)
Experimental	0.002	0.003	0.006	0.020	0.039*	0.032	1.089	1.574*	0.143
	(0.62)	(0.75)	(1.43)	(1.28)	(1.75)	(1.26)	(1.03)	(1.72)	(0.14)
Survey	0.006*	0.001	0.009	0.061***	0.035	0.061*	3.143**	0.961	1.270
	(1.92)	(0.20)	(1.49)	(3.00)	(1.17)	(1.79)	(2.00)	(0.77)	(0.92)
Questionnaire	-0.001	-0.001	0.010	0.0021	0.017	0.036	2.953	1.541	
	(-0.03)	(-0.12)	(0.74)	(0.05)	(0.29)	(0.47)	(1.37)	(0.77)	
Linear	0.002	0.005*	0.001	0.013	0.017	0.006	0.749	0.230	1.579*
	(1.24)	(1.94)	(0.06)	(1.32)	(1.27)	(0.33)	(1.25)	(0.42)	(1.70)
Journal F.E.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year F.E.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	193	193	185	193	193	185	185	188	165
Adjusted R^2	0.379	0.233	0.249	0.243	0.210	0.271	0.263	0.164	0.193

This table reports the OLS and logit estimates for the BL conformity measure *MAD* (columns 1-3), conformity measure *FSD-Score* (columns 4-6), and conformity measure *KS1* (columns 7-9) for the period 1996 to 2012. The *MAD* score is defined as the mean of the absolute value of the difference between the frequency of each first digit within the sample, and the frequency as determined by BL. The *FSD-Score* is the Kolmogorov-Smirnov statistic which is defined as the maximum cumulative deviation from the theoretical distribution of BL. The larger the *MAD* or *FSD-Score* the higher the levels of non-conformity with BL. The *KS1* score is an indicator variable which equals one if the Kolmogorov-Smirnov test statistic is greater than the critical value at 1% significance and zero otherwise. A *KS1* value of one indicates the non-conformity with BL. *Hunton_R* is an indicator variable which equals one if the article is authored by Hunton and is retracted. *Ln_Pool* is the natural log of the number of first digits used in the calculation of the conformity score. *Num_Authors* is the number of authors on each article. *Experimental*, *Survey* and *Questionnaires* are all indicator variables which identifies the type of methodology employed in the articles. *Linear* is an indicator variable which takes the value of one if the analysis used in the article is a linear process and zero otherwise. t-statistics are reported in parentheses. *, **, *** represent significance level of 10%, 5% and 1%, respectively (two-tailed).

Table 5: Comparing Conformity to Benford Law (BL) of Hunton's Non-Retracted Articles relative to their Control Group.

Conformity Measure (MAD or FSD-Score or KS1) $= \alpha + \beta_1 Hunton\_NR + \beta_2 Pool + \beta_3 Num\_Authors + \beta_4 Experimental + \beta_5 Survey + \beta_6 Questionanaire + \beta_7 Linear$ *+ Year F.E. + Journal F.E.*

Conformity Measure *MAD, FSD-Score, and KS1*

	MAD			*FSD-Score*			*KS1*[*]	
	All Digits	Descriptive	Regression	All Digits	Descriptive	Regression	All Digits	Descriptive
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Hunton_NR	0.004	**0.006**[*]	**0.009**[**]	**0.027**[*]	0.023	**0.052**[**]	1.227	0.540
	(1.62)	**(1.93)**	**(2.55)**	**(1.92)**	(1.39)	**(2.60)**	(1.57)	(0.72)
Ln_Pool	0.001	-0.003[*]	-0.008[***]	0.001	-0.018[**]	-0.044[***]	1.529[***]	0.815[**]
	(0.34)	(-1.70)	(-4.99)	(0.14)	(-2.31)	(-4.38)	(2.92)	(2.11)
Num_Author	0.001	0.001	0.001	0.008	0.011[*]	0.005	0.451	0.095
	(0.66)	(0.46)	(0.45)	(1.62)	(1.98)	(0.74)	(1.56)	(0.36)
Experimental	0.014[*]	0.022[*]	-0.005	0.062	0.082	-0.083	17.78	0.389
	(1.68)	(1.80)	(-0.39)	(1.18)	(1.30)	(-1.08)	(0.00)	(0.00)
Survey	0.014	0.023	-0.013	0.059	0.051	-0.099	3.607	-14.610
	(1.21)	(1.31)	(-0.68)	(0.79)	(0.56)	(-0.93)	(0.00)	(-0.00)
Questionnaire	-0.001	0.009	-0.011	0.047	0.084	-0.082	19.36	
	(-0.02)	(0.57)	(-0.66)	(0.68)	(1.02)	(-0.83)	(0.00)	
Linear	-0.001	0.006	-0.004	-0.017	0.013	-0.021	-0.740	0.788
	(-0.10)	(1.61)	(-0.66)	(-1.07)	(0.70)	(-0.51)	(-0.70)	(0.84)
Journal F.E.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year F.E.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	108	108	97	108	108	97	92	102
Adjusted R^2	0.047	0.048	0.226	0.030	0.130	0.211	0.283	0.193

This table reports the OLS and logit estimates for the BL conformity measure MAD (columns 1-3), conformity measure *FSD-Score* (columns 4-6), and conformity measure KS1 (columns 7-9) for the period 1996 to 2012. The *MAD* score is defined as the mean of the absolute value of the difference between the frequency of each first digit within the sample, and the frequency as determined by BL. The *FSD-Score* is the Kolmogorov-Smirnov statistic which is defined as the maximum cumulative deviation from the theoretical distribution of BL. The larger the *MAD* or *FSD-Score* the higher the levels of non-conformity with BL. The *KS1* score is an indicator variable which equals one if the Kolmogorov-Smirnov test statistic is greater than the critical value at 1% significance and zero otherwise. A *KS1* value of one indicates the non-conformity with BL. *Hunton_NR* is an indicator variable which equals one if the article is authored by Hunton but is not retracted. *Ln_Pool* is the natural log of the number of first digits used in the calculation of the conformity score. *Num_Authors* is the number of authors on each article. *Experimental*, *Survey* and *Questionnaires* are all indicator variables which identifies the type of methodology employed in the articles. *Linear* is an indicator variable which takes the value of one if the analysis used in the article is a linear process and zero otherwise. t-statistics are reported in parentheses. *, **, *** represent significance level of 10%, 5% and 1%, respectively (two-tailed). [*]Due to the small sample size, several multicollinearity issues prevent us from adequately running the logit model for the regression analyses numbers attributable to this group of papers

Table 6: Comparing Hunton's Articles conformity to Benford Law (BL) relative to Control Group 2.

Conformity Measure (MAD, FSD-Score, KS1) $= \alpha + \beta_1 Hunton\_R + \beta_2 Hunton\_NR + \beta_3 Ln\_Pool + \beta_4 Num\_Authors + \beta_5 Experimental + \beta_6 Survey + \beta_7 Questionanaire + \beta_8 Linear + Year\ F.E. + Journal\ F.E.$

Conformity Measure *MAD, FSD-Score, and KS1*

	MAD			*FSD-Score*			*KS1*		
	All Digits	Descriptive	Regression	All Digits	Descriptive	Regression	All Digits	Descriptive	Regression
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Hunton_R	**0.008*****	**0.009*****	**0.007*****	**0.032*****	**0.035*****	**0.034****	**1.460*****	**1.788*****	**1.355****
	(4.24)	**(3.32)**	**(2.74)**	**(3.15)**	**(2.65)**	**(2.42)**	**(2.85)**	**(3.26)**	**(2.24)**
Hunton_NR	**0.005***	0.006	**0.008****	**0.033****	0.015	**0.046****	**1.66****	0.038	**2.344*****
	(1.93)	(0.093)	**(2.27)**	**(2.33)**	(0.84)	**(2.47)**	**(2.16)**	(0.05)	**(2.73)**
Ln_Pool	-0.003***	-0.006***	-0.008***	-0.020***	-0.029***	-0.037***	0.816***	1.305***	0.661**
	(-3.10)	(-5.10)	(-8.23)	(-4.17)	(-5.34)	(-6.75)	(3.02)	(4.94)	(2.42)
Num_Author	-0.002***	-0.003***	0.001	-0.007**	-0.010**	0.006	-0.274	-0.227	0.256
	(-3.53)	(-3.66)	(0.91)	(-2.42)	(-2.49)	(1.42)	(-1.49)	(-1.28)	(1.14)
Experimental	0.004**	0.008***	-0.001	0.014	0.034***	-0.009	0.457	1.438***	0.566
	(2.18)	(3.43)	(-0.42)	(1.56)	(2.91)	(-0.71)	(0.81)	(2.66)	(0.86)
Survey	0.007***	0.006	0.002	0.036***	0.027*	0.015	1.516**	1.226*	0.146
	(3.27)	(1.91)	(0.67)	(3.21)	(1.88)	(0.94)	(2.41)	(1.92)	(0.18)
Questionnaire	0.003	0.006	-0.004	0.010	0.044**	-0.011	1.131	2.706***	1.071
	(1.10)	(1.59)	(-1.05)	(0.72)	(2.40)	(-0.56)	(1.39)	(3.23)	(0.98)
Linear	-0.0025	-0.0012	-0.001	-0.005	-0.003	0.002	0.186	0.371	-0.029
	(-1.58)	(-0.56)	(-0.35)	(-0.64)	(-0.33)	(0.13)	(0.37)	(0.74)	(-0.04)
Journal F.E.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year F.E.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	382	382	352	382	382	352	288	294	223
Adjusted R^2	0.195	0.387	0.320	0.136	0.342	0.314	0.210	0.274	0.243

This table reports the OLS and logit regression estimates for the BL conformity measures *MAD* or *FSD-Score*, for the period 1996 to 2012. The larger the conformity score the higher the levels of non-conformity with BL. The *KS1* score is an indicator variable which equals one if the Kolmogorov-Smirnov test statistic is greater than the critical value at 1% significance and zero otherwise. *Hunton_R* is an indicator variable which equals one if the article is authored by Hunton and is retracted. *Hunton_NR* is an indicator variable which equals one if the article is authored by Hunton but not retracted. *Pool* is the number of first digits. *Num_Authors* is the number of authors on each article. *Experimental*, *Survey* and *Questionnaires* are all indicator variables which identifies the type of methodology employed in the articles. *Linear* is an indicator variable which takes the value of one if the analysis used in the article is a linear process and zero otherwise. t-statistics are reported in parentheses. *, **, *** represent significance level of 10%, 5% and 1%, respectively (two-tailed).

Table 7: Abnormal *MAD* Scores (*AB_MAD* = $\widehat{MAD}$ - *MAD*)

Panel A: *AB_MAD* scores descriptives

	Hunton_R				***Hunton_NR***				***Control Group***					
AB_MAD:	**n**	**Mean**	**Std. Dev**	**Median**	**n**	**Mean**	**Std. Dev**	**Median**	**n**	**Mean**	**Std. Dev**	**Median**	**Differences**	
		(1)	(2)	(3)		(4)	(5)	(6)		(7)	(8)	(9)	(7)-(1)	(7)-(4)
All Digits	37	-0.0078	0.0112	-0.0047	18	-0.0049	0.0134	-0.0026	573	0.0000	0.0067	0.0000	0.0078***	0.0049***
Descriptive	37	-0.0084	0.0148	-0.0061	18	-0.0060	0.0183	-0.0021	573	0.0000	0.0098	0.0000	0.0084***	0.0060**
Regression	35	-0.0070	0.0164	-0.0031	18	-0.0083	0.0136	-0.0065	528	0.0000	0.0092	0.0002	0.0070***	0.0083***

Panel B: Sign and Magnitude of Hunton's *AB_MAD* compared to Control Group

	Hunton_R				***Hunton_NR***				**Control Group**				
	n	**%**	**Mean**	**Std. Dev**	**n**	**%**	**Mean**	**Std. Dev**	**n**	**Mean**	**Std. Dev**	**Differences**	
MAD > $\widehat{MAD}$		**(1)**	**(2)**	**(3)**		**(4)**	**(5)**	**(6)**		**(7)**	**(8)**	**(7)-(2)**	**(7)-(5)**
All Digits	27	73%	-0.0119	0.0101	11	61%	-0.0099	0.0152	235	-0.0056	0.0061	0.0064***	0.0043**
Descriptive	24	65%	-0.0159	0.0129	10	56%	-0.0162	0.0189	243	-0.0084	0.0082	0.0075***	0.0078***
Regression	21	60%	-0.0158	0.0157	13	72%	-0.0135	0.0122	207	-0.0085	0.0077	0.0073***	0.0050**
MAD < $\widehat{MAD}$													
All Digits	10	27%	0.0035	0.0034	7	39%	0.0029	0.0020	338	0.0039	0.0036	0.0004	0.0010
Descriptive	13	35%	0.0055	0.0042	8	44%	0.0067	0.0045	330	0.0062	0.0052	0.0007	-0.0006
Regression	14	40%	0.0062	0.0033	5	28%	0.0053	0.0042	321	0.0055	0.0049	-0.0007	0.0001

This table presents the summary statistics for the abnormal *MAD* scores (*AB_MAD*) which are the difference between the predicted *MAD* score ($\widehat{MAD}$) and the actual *MAD* score. $\widehat{MAD}$ is determined using the estimates obtained from the following equation $MAD = \alpha + \beta_1 Pool + \beta_2 Num\_Authors + \beta_3 Experimental + \beta_4 Survey + \beta_5 Questionanaire + \beta_6 Linear$ + *Year fixed effects.* + *Journal fixed effects* using the full set of control papers (n=573). *, **, *** represent significance level of 10%, 5% and 1%, (two-tailed) respectively.

Table 8: Conformity Measure (*MAD*) for Prof Libby and Prof Wier relative to each Control Group.

	Control Group 1						Control Group 2					
	All Digits		Descriptive		Regression		All Digits		Descriptive		Regression	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Prof Libby	-0.001		-0.006		-0.002		-0.0001		-0.005		-0.002	
	(-0.07)		(-1.41)		(-0.38)		(-0.03)		(-1.11)		(-0.50)	
Prof Wier		0.001		-0.001		0.004		-0.003		0.0003		0.001
		(0.22)		(-0.12)		(0.73)		(-0.85)		(0.07)		(0.12)
Ln_Pool	-0.003***	-0.003***	-0.005***	-0.005***	-0.007***	-0.007***	-0.003***	-0.003***	-0.006***	-0.006***	-0.008***	-0.008***
	(-4.09)	(-4.08)	(-5.62)	(-5.96)	(-7.40)	(-7.43)	(-3.31)	(-3.40)	(-5.22)	(-5.09)	(-8.03)	(-8.02)
Num_Author	0.001	0.001	-0.000	-0.000	0.001	0.001	-0.002***	-0.002***	-0.003***	-0.003***	0.001	0.001
	(1.04)	(1.16)	(-0.20)	(-0.11)	(1.15)	(1.08)	(-4.27)	(-4.08)	(-3.85)	(-3.95)	(0.77)	(0.69)
Experimental	0.000	0.000	0.005	0.003	0.006	0.004	0.003	0.002	0.006**	0.006***	-0.003	-0.003
	(0.22)	(0.18)	(1.60)	(0.90)	(1.59)	(1.24)	(1.41)	(1.27)	(2.39)	(2.30)	(-1.37)	(-1.35)
Survey	0.000	-0.000	0.002	-0.000	0.003	0.002	0.006***	0.007***	0.004	0.004*	0.002	0.002
	(0.02)	(-0.10)	(0.40)	(-0.01)	(0.68)	(0.60)	(2.91)	(2.99)	(1.31)	(1.24)	(0.60)	(0.56)
Questionnaire	0.001	0.001	0.001	-0.002	0.007	0.004	0.003	0.002	0.005	0.005	-0.004	-0.004
	(0.14)	(0.12)	(0.12)	(-0.25)	(0.76)	(0.46)	(0.97)	(0.84)	(1.20)	(1.27)	(-1.05)	(-0.99)
Linear	0.001	0.001	0.004**	0.005**	-0.003	-0.002	-0.002	-0.002	-0.001	-0.001	-0.002	-0.002
	(1.07)	(0.93)	(2.39)	(2.54)	(-1.07)	(-0.85)	(-1.55)	(-1.54)	(-0.38)	(-0.34)	(-0.89)	(-0.86)
Journal F.E.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year F.E.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	256	256	256	256	238	238	327	327	327	327	299	299
Adjusted R^2	0.100	0.097	0.134	0.150	0.272	0.267	0.241	0.243	0.448	0.450	0.374	0.372

This table reports the ordinary least squares estimates for the BL conformity measure *MAD* for the period 1996 to 2012. The larger the conformity score the higher the levels of non-conformity with BL. *Prof Libby* is an indicator variable which equals one if the article is authored by Professor Libby. *Prof Wier* is an indicator variable which equals one if the article is authored by Professor Wier *Pool* is the number of first digits. *Num_Authors* is the number of authors on each article. *Experimental*, *Survey* and *Questionnaires* are all indicator variables which identifies the type of methodology employed in the articles. *Linear* is an indicator variable which takes the value of one if the analysis used in the article is a linear process and zero otherwise. t-statistics are reported in parentheses. *, **, *** represent significance level of 10%, 5% and 1%, respectively (two-tailed).

Table 9: Conformity Measure *MAD* and Number of Citations (#Citations)

	ln #Citations							
Sample	All Papers			Retracted v Control	Non-Retracted v Control	Retracted v Control	Non-Retracted v Control	Hunton v All Papers
Digits	*All*	*Descriptive*	*Regression*			*All*		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
MAD	1.691	-1.462	0.794	2.840	-3.286			
	(0.36)	(-0.38)	(0.33)	(0.17)	(-0.27)			
HUNTON						-0.414	0.137	-0.017
						(-1.22)	(0.42)	(-0.08)
Ln_Pool	0.158***	0.098**	0.114*	0.308**	-0.016	0.281**	-0.029	0.155**
	(3.04)	(2.17)	(1.86)	(2.89)	(-0.12)	(2.89)	(-0.25)	(2.80)
Num_Author	0.021	0.013	0.035	-0.026	0.045	-0.001	0.0357	0.020
	(0.39)	(0.25)	(0.60)	(-0.27)	(0.47)	(-0.02)	(0.35)	(0.40)
Experimental	-0.822***	-0.839***	-0.858***	-0.854***	-0.403	-0.797***	-0.525	-0.817***
	(-4.07)	(-4.14)	(-3.57)	(-3.91)	(-0.79)	(-3.49)	(-1.06)	(-4.07)
Survey	-0.328	-0.335	-0.202	-0.952**	1.148	-0.827*	1.053	-0.316
	(-1.60)	(-1.65)	(-0.91)	(-2.97)	(1.80)	(-2.11)	(1.74)	(-1.57)
Questionnaire	-0.565*	-0.566*	-0.525	-0.059	0.137	0.151	0.031	-0.559
	(-1.75)	(-1.74)	(-1.34)	(-0.29)	(0.27)	(0.53)	(0.06)	(-1.74)*
Linear	0.068	0.112	-0.080	-0.237**	0.598*	-0.181	0.587*	0.068
	(0.42)	(0.57)	(-0.33)	(-2.33)	(2.17)	(-1.33)	(2.18)	(0.42)
Year F.E	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Journal F.E.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subject F.E.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	628	628	581	193	108	193	108	628
Adjusted R^2	0.461	0.459	0.466	0.249	0.512	0.266	0.512	0.461

This table reports the ordinary least squares estimates for the number of citations (#Citations) each paper received between 1996 to 2012. *MAD* is the BL conformity measure, the larger the *MAD* score the less the digits conform to BL. *HUNTON* is an indicator variable taking the value of 1 if Hunton was an author and zero otherwise. *Pool* is the number of first digits. *Num_Authors* is the number of authors on each article. *Experimental*, *Survey* and *Questionnaires* are all indicator variables which identifies the type of methodology employed in the articles. *Linear* is an indicator variable which takes the value of one if the analysis used in the article is a linear process and zero otherwise. Heteroskedasticity-robust standard errors are clustered at journal level. t-statistics are reported in parentheses. *, **, *** represent significance level of 10%, 5% and 1%, respectively (two-tailed)

Table 10: Investigating Additional Authors with Retracted Papers

Panel A: Descriptive Statistics for Additional Retracted Articles and their Control Sample

		Retracted Sample (n=45)					Control Sample (n=166)					Difference	
Variable		**Mean**	**Std. Dev**	**25th Q**	**Median**	**75th Q**	**Mean**	**Std. Dev.**	**25th Q**	**Median**	**75th Q**	**t-stat**	**Wilcoxon**
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
By Group:													
All Digits	***Dependent Variables:***												
	MAD	0.0174	0.0078	0.0114	0.0147	0.0223	0.0158	0.0068	0.0115	0.0146	0.0186	1.34*	0.63
	FSD-Score	0.0904	0.0523	0.0597	0.0743	0.1037	0.0886	0.0491	0.0549	0.0772	0.1116	0.21	0.07
	KS1	0.2667	0.4472	0.0000	0.0000	1.0000	0.2229	0.4174	0.0000	0.0000	0.0000	0.61	0.61
Descriptive Digits													
	MAD	0.0215	0.0133	0.0122	0.0163	0.0240	0.0178	0.0079	0.0124	0.0163	0.0219	2.38***	0.73
	FSD-Score	0.1127	0.0817	0.0597	0.0799	0.1214	0.1019	0.0601	0.0590	0.0862	0.1317	0.98	0.14
	KS1	0.2889	0.4584	0.0000	0.0000	1.0000	0.2048	0.4048	0.0000	0.0000	0.0000	1.20	1.19
Regression Digits													
	MAD	0.0313	0.0201	0.0201	0.0249	0.0360	0.0253	0.0135	0.0159	0.0213	0.0316	1.56*	1.33*
	FSD-Score	0.1501	0.0862	0.0799	0.1208	0.2446	0.1354	0.0825	0.0835	0.1223	0.1529	0.69	0.56
	KS1	0.1500	0.3663	0.0000	0.0000	0.0000	0.2143	0.4133	0.0000	0.0000	0.0000	0.62	0.63
	Control Variables:												
	Pool	5.3704	0.5558	5.1299	5.4381	5.6630	5.2728	0.7519	4.7622	5.2677	5.7930	0.81	1.06
	Num_Authors	3.8222	1.3864	3.0000	4.0000	5.0000	4.8795	2.8173	3.0000	4.0000	6.0000	2.43**	1.78*
	Experimental	0.7111	0.4584	0.0000	1.0000	1.0000	0.6446	0.4801	0.0000	1.0000	1.0000	0.83	0.83
	Interview	0.0444	0.2084	0.0000	0.0000	0.0000	0.0060	0.0776	0.0000	0.0000	0.0000	1.93*	1.92*
	Survey	0.2444	0.4346	0.0000	0.0000	0.0000	0.3494	0.4782	0.0000	0.0000	1.0000	1.33	1.32
	Linear	0.9333	0.2523	1.0000	1.0000	1.0000	0.9699	0.1714	1.0000	1.0000	1.0000	1.13	1.13

Panel B: Comparing Conformity to Benford Law (BL) of Additional Retracted Articles relative to their Control Group.

	MAD			*FSD-Score*			*KS1*		
	All Digits	Descriptive	Regression	All Digits	Descriptive	Regression	All Digits	Descriptive	Regression
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Retracted	**0.002***	**0.005*****	0.006	0.012	**0.023***	0.007	**1.117****	**1.002***	0.061
	(1.67)	**(3.24)**	(1.57)	(1.27)	**(1.94)**	(0.30)	**(2.05)**	**(1.80)**	(0.06)
Ln_Pool	-0.003***	-0.005***	-0.012***	-0.015***	-0.017**	-0.048***	1.159***	0.994***	0.691*
	(-4.80)	(-5.34)	(-7.62)	(-2.67)	(-2.60)	(-4.54)	(3.45)	(2.98)	(1.73)
Num_Author	0.000*	0.001*	-0.002*	0.002	0.003	-0.015*	0.230**	0.253**	-0.622*
	(1.74)	(1.81)	(-1.86)	(1.15)	(1.26)	(-1.78)	(2.06)	(2.21)	(-1.67)
Experimental	-0.018	-0.017	-0.007	-0.139	-0.109	-0.126	14.32	16.68***	-0.035
	(-1.57)	(-1.22)	(-0.45)	(-1.54)	(-0.99)	(-1.15)	(0.00)	(6.73)	(-0.01)
Survey	-0.006	-0.005	0.001	-0.009	0.016	-0.061	15.84	17.04	-1.640
	(-1.02)	(-0.74)	(0.00)	(-0.20)	(0.28)	(-0.86)	(0.00)	(0.01)	(-0.71)
Linear	-0.007***	-0.005	-0.009*	-0.017	-0.018	-0.001	-0.147	-0.558	1.242
	(-2.66)	(-1.57)	(-1.88)	(-0.85)	(-0.75)	(-0.00)	(-0.15)	(-0.56)	(0.88)
Journal F.E.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year F.E.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	211	211	90	211	211	90	171	171	61
Adjusted R^2	0.208	0.338	0.469	0.013	0.143	0.192	0.197	0.198	0.183

This table reports the OLS and logit estimates for the BL conformity measure *MAD* (columns 1-3), conformity measure *FSD-Score* (columns 4-6), and conformity measure *KS1* (columns 7-9). The *MAD* score is defined as the mean of the absolute value of the difference between the frequency of each first digit within the sample, and the frequency as determined by BL. The *FSD-Score* is the Kolmogorov-Smirnov statistic which is defined as the maximum cumulative deviation from the theoretical distribution of BL. The larger the *MAD* or *FSD-Score* the higher the levels of non-conformity with BL. The *KS1* score is an indicator variable which equals one if the Kolmogorov-Smirnov test statistic is greater than the critical value at 1% significance and zero otherwise. A *KS1* value of one indicates the non-conformity with BL. *Retracted* is an indicator variable which equals one if the article is retracted. *Ln_Pool* is the natural log of the number of first digits used in the calculation of the conformity score. *Num_Authors* is the number of authors on each article. *Experimental*, *Survey* and *Questionnaires* are all indicator variables which identifies the type of methodology employed in the articles. *Linear* is an indicator variable which takes the value of one if the analysis used in the article is a linear process and zero otherwise. t-statistics are reported in parentheses. *, **, *** represent significance level of 10%, 5% and 1%, respectively (two-tailed).

Table 11: Output from DGP Analysis

	All	Descriptives	Regression
Naturally Occurring Numbers			
High Correlation			
For differing n=100, repeated 500 times	0.0120	0.0080	0.0120
For differing n=100, repeated 1000 times	0.0099	0.0088	0.0115
For differing n=500, repeated 500 times	0.0107	0.0126	0.0107
For differing n=500, repeated 1000 times	0.0103	0.0128	0.0087
DGP			
Beta distributed around 3			
For differing n=100, repeated 500 times	0.0307	0.0210	0.0489
For differing n=100, repeated 1000 times	0.0303	0.0191	0.0481
For differing n=500, repeated 500 times	0.0290	0.0339	0.0317
For differing n=500, repeated 1000 times	0.0294	0.0352	0.0317
Beta normally distributed			
For differing n=100, repeated 500 times	0.0152	0.0248	0.0248
For differing n=100, repeated 1000 times	0.0158	0.0101	0.0247
For differing n=500, repeated 500 times	0.0221	0.0353	0.0337
For differing n=500, repeated 1000 times	0.0220	0.0359	0.0335
Normally distributed & correlation 0.6			
For differing n=100, repeated 500 times	0.0133	0.0221	0.0320
For differing n=100, repeated 1000 times	0.0137	0.0229	0.0330
For differing n=500, repeated 500 times	0.0263	0.0289	0.0400
For differing n=500, repeated 1000 times	0.0272	0.0302	0.0413

This table provides the MAD scores from output of a bootstrap regression using two different sample sizes and repeating the regression 500 or 1000 times. The results are derived from one naturally occurring dataset and three different DGP datasets.

Table 12: MAD Scores after Manually Editing Regression Output

	Regression	% Diff
Naturally Occurring Numbers		
Sample Size = 100		
Unadjusted – data output	0.0027	
Adjusted 5% of betas (plus corresponding t-stats)	0.0034	+ 26%
Adjusted 7% of betas (plus corresponding t-stats)	0.0038	+ 42%
Adjusted 10% of betas (plus corresponding t-stats)	0.0041	+ 55%
Sample Size = 500		
Unadjusted – data output	0.0024	
Adjusted 5% of betas (plus corresponding t-stats)	0.0029	+ 24%
Adjusted 7% of betas (plus corresponding t-stats)	0.0032	+ 34%
Adjusted 10% of betas (plus corresponding t-stats)	0.0038	+ 60%
Sample Size = 600		
Unadjusted – data output	0.0025	
Adjusted 5% of betas (plus corresponding t-stats)	0.0036	+ 33%
Adjusted 7% of betas (plus corresponding t-stats)	0.0033	+ 45%
Adjusted 10% of betas (plus corresponding t-stats)	0.0041	+ 65%

This table provides the MAD score for regression output from a bootstrap regression of naturally occurring dataset at three different sample sizes and repeating the process 600 times. The adjusted MAD score is calculated after manually adjusting a proportion of betas. Adjusted beta is calculated by multiplying the standard errors by 1.96.

Table 13: Sensitivity of *MAD* to Controlled Manipulation.

Panel A: All Numbers

#Obs	Base	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
10	0.042	0.040	0.043	0.043	0.042	0.043	0.048	0.051	0.052	0.057	0.059
20	0.033	0.033	0.033	0.033	0.036	0.034	0.039	0.040	0.043	0.047	0.051
30	0.028	0.028	0.027	0.028	0.029	0.029	0.032	0.036	0.039	0.043	0.046
40	0.026	0.026	0.025	0.024	0.026	0.027	0.031	0.034	0.036	0.041	0.045
50	0.023	0.024	0.023	0.023	0.025	0.026	0.029	0.033	0.034	0.039	0.043
100	0.019	0.019	0.018	0.018	0.019	0.021	0.024	0.028	0.031	0.037	0.039
150	0.017	0.016	0.016	0.016	0.017	0.020	0.023	0.027	0.030	0.035	0.038
200	0.016	0.015	0.015	0.015	0.016	0.019	0.022	0.026	0.029	0.034	0.037
250	0.015	0.014	0.014	0.015	0.016	0.018	0.022	0.025	0.029	0.033	0.036
300	0.013	0.013	0.013	0.014	0.016	0.018	0.022	0.025	0.028	0.033	0.036
350	0.013	0.012	0.013	0.014	0.016	0.018	0.022	0.025	0.028	0.032	0.035
400	0.012	0.012	0.013	0.014	0.016	0.018	0.022	0.024	0.028	0.032	0.035
450	0.012	0.012	0.012	0.013	0.016	0.019	0.021	0.024	0.028	0.032	0.035
500	0.012	0.011	0.012	0.013	0.015	0.019	0.021	0.025	0.028	0.031	0.035
1000	0.011	0.009	0.011	0.012	0.015	0.018	0.022	0.025	0.028	0.030	0.033
5000	0.009	0.006	0.009	0.011	0.015	0.018	0.021	0.025	0.027	0.030	0.032
10000	0.009	0.006	0.009	0.011	0.015	0.018	0.021	0.025	0.027	0.029	0.032
100000	0.009	0.006	0.009	0.011	0.014	0.018	0.021	0.025	0.028	0.029	0.031

Panel B: Descriptive Statistics

#Obs	Base	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
10	0.041	0.045	0.040	0.039	0.042	0.043	0.046	0.051	0.052	0.056	0.058
20	0.032	0.036	0.033	0.032	0.036	0.035	0.037	0.040	0.044	0.045	0.052
30	0.030	0.030	0.028	0.027	0.029	0.030	0.032	0.035	0.039	0.040	0.048
40	0.027	0.028	0.026	0.025	0.026	0.028	0.029	0.033	0.037	0.038	0.046
50	0.026	0.025	0.024	0.022	0.024	0.026	0.028	0.032	0.036	0.037	0.044
100	0.022	0.019	0.020	0.017	0.019	0.022	0.025	0.028	0.033	0.035	0.040
150	0.019	0.017	0.017	0.016	0.018	0.020	0.023	0.027	0.031	0.033	0.037
200	0.017	0.016	0.015	0.015	0.016	0.019	0.023	0.026	0.030	0.032	0.036
250	0.017	0.015	0.014	0.014	0.016	0.019	0.022	0.025	0.030	0.032	0.036
300	0.016	0.014	0.013	0.014	0.015	0.018	0.022	0.025	0.029	0.031	0.035
350	0.015	0.013	0.013	0.014	0.015	0.018	0.022	0.025	0.029	0.032	0.035
400	0.015	0.013	0.012	0.014	0.015	0.018	0.022	0.025	0.029	0.031	0.035
450	0.015	0.012	0.012	0.014	0.015	0.018	0.022	0.025	0.029	0.031	0.035
500	0.015	0.012	0.011	0.013	0.015	0.018	0.022	0.024	0.028	0.031	0.035
1000	0.013	0.010	0.009	0.012	0.014	0.017	0.021	0.024	0.027	0.031	0.033
5000	0.012	0.008	0.007	0.011	0.013	0.017	0.021	0.024	0.027	0.030	0.032
10000	0.012	0.007	0.007	0.011	0.013	0.017	0.021	0.024	0.027	0.030	0.032
100000	0.011	0.007	0.007	0.011	0.013	0.017	0.021	0.024	0.028	0.030	0.032

Panel C: Regression Output

#Obs	Base	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
10	0.042	0.043	0.043	0.047	0.048	0.048	0.050	0.050	0.051	0.055	0.058
20	0.032	0.031	0.032	0.036	0.037	0.037	0.041	0.044	0.046	0.048	0.051
30	0.025	0.025	0.026	0.030	0.031	0.032	0.037	0.039	0.041	0.045	0.049
40	0.022	0.023	0.025	0.028	0.029	0.030	0.035	0.037	0.039	0.042	0.044
50	0.020	0.020	0.022	0.025	0.027	0.029	0.032	0.035	0.038	0.039	0.043
100	0.014	0.015	0.017	0.020	0.023	0.024	0.027	0.031	0.034	0.036	0.039
150	0.012	0.013	0.015	0.018	0.021	0.023	0.026	0.030	0.032	0.035	0.037
200	0.011	0.012	0.013	0.016	0.020	0.022	0.025	0.029	0.030	0.034	0.036
250	0.010	0.011	0.012	0.015	0.020	0.021	0.024	0.028	0.029	0.033	0.035
300	0.009	0.010	0.012	0.015	0.019	0.021	0.024	0.028	0.029	0.033	0.035
350	0.008	0.010	0.012	0.015	0.019	0.021	0.023	0.027	0.029	0.032	0.035
400	0.008	0.009	0.011	0.014	0.018	0.021	0.023	0.027	0.029	0.032	0.035
450	0.007	0.009	0.011	0.014	0.018	0.020	0.023	0.026	0.029	0.032	0.035
500	0.007	0.009	0.011	0.014	0.018	0.020	0.023	0.026	0.029	0.032	0.035
1000	0.006	0.008	0.010	0.014	0.017	0.019	0.022	0.025	0.028	0.031	0.033
5000	0.003	0.006	0.010	0.013	0.016	0.018	0.021	0.024	0.027	0.029	0.032
10000	0.003	0.006	0.009	0.013	0.015	0.018	0.021	0.023	0.026	0.029	0.032
100000	0.002	0.006	0.009	0.012	0.015	0.018	0.021	0.023	0.026	0.029	0.031

This table presents the results of *MAD* sensitivity to the controlled manipulation of numbers within the resulting descriptive and regression outputs. Columns depict the level of manipulation undertaken whilst the rows investigate various sample sizes. Each cell within the table represents the average *MAD* score based on 100 runs. Highlighted cells represent significant differences from the base level at the 5% level.