publications

MDPI

Article

Can Retracted Social Science Articles Be Distinguished from Non-Retracted Articles by Some of the Same Authors, Using Benford's Law or Other Statistical Methods?

Walter R. Schumm [1,*], Duane W. Crawford [1], Lorenza Lockett [2], Asma bin Ateeq [3] and Abdullah AlRashed [4]

[1] Department of Applied Human Sciences, Kansas State University, 1700 Anderson Avenue, Manhattan, KS 66506, USA
[2] Department of Sociology, Anthropology, and Social Work, Kansas State University, 1603 Old Claflin Place, Manhattan, KS 66506, USA
[3] Education Department, Arab East Colleges, 3310 Abdullah bin Umar, Al Qirawan, Riyahd 13544-6394, Saudi Arabia
[4] Security Studies Program, Graduate School, Kansas State University, Fairchild Hall, 1700 Anderson Avenue, Manhattan, KS 66506, USA
* Correspondence: schumm@ksu.edu

Abstract: A variety of ways to detect problems in small sample social science surveys has been discussed by a variety of authors. Here, several new approaches for detecting anomalies in large samples are presented and their use illustrated through comparisons of seven retracted or corrected journal articles with a control group of eight articles published since 2000 by a similar group of authors on similar topics; all the articles involved samples from several hundred to many thousands of participants. Given the small sample of articles (k = 15) and low statistical power, only 2/12 of individual anomaly comparisons were not statistically significant, but large effect sizes ($d > 0.80$) were common for most of the anomaly comparisons. A six-item total anomaly scale featured a Cronbach alpha of 0.92, suggesting that the six anomalies were moderately correlated rather than isolated issues. The total anomaly scale differentiated the two groups of articles, with an effect size of 3.55 ($p < 0.001$); an anomaly severity scale derived from the same six items, with an alpha of 0.94, yielded an effect size of 3.52 ($p < 0.001$). Deviations from the predicted distribution of first digits in regression coefficients (Benford's Law) were associated with anomalies and differences between the two groups of articles; however, the results were mixed in terms of statistical significance, though the effect sizes were large ($d \geq 0.90$). The methodology was able to detect unusual anomalies in both retracted and non-retracted articles. In conclusion, the results provide several useful approaches that may be helpful for detecting questionable research practices, especially data or results fabrication, in social science, medical, or other scientific research.

Keywords: research integrity; fraud; research misconduct; anomalous results; retraction

check for updates

Citation: Schumm, W.R.; Crawford, D.W.; Lockett, L.; Ateeq, A.b.; AlRashed, A. Can Retracted Social Science Articles Be Distinguished from Non-Retracted Articles by Some of the Same Authors, Using Benford's Law or Other Statistical Methods? *Publications* **2023**, *11*, 14. https://doi.org/10.3390/publications11010014

Academic Editor: Bart Penders

Received: 28 November 2022
Revised: 3 February 2023
Accepted: 24 February 2023
Published: 3 March 2023

1. Introduction

How may editors and their reviewers detect problems in submitted papers before those papers might be accepted and then retracted for methodological problems? Solutions that would allow editors or reviewers to detect such problems may not be easy or obvious.

The number of articles retracted on account of scientific misconduct has increased in recent decades, even in medicine [1–4]; some academics have had dozens of their articles retracted [5–11], in spite of the grave consequences caused by scientific conduct being exposed [12]. Sometimes the misconduct has involved apparent or possible fabrication of data, which is one of the most serious types of scientific misconduct, although relatively rare [13,14].

What are reviewers and journal editors to do? We would like to suggest several statistical methods for detecting data anomalies, which may reflect fabrication of data and/or results.

There have been few systemic studies of retracted papers, especially with respect to papers from authors with multiple retractions [15] (p. 277). Several anomalies had been noted in some of the several articles which were of concern to Pickett [16], published in top tier journals from 2000 to 2020. The particular concerns expressed by Pickett [16] are as follows: (1) a high ratio of beta coefficients and standard errors that were identical across multiple models; (2) a high number of "hand-calculated" *t*-test values; (3) an absence of zeros in second or third decimal points; (4) binary values that were impossible; and (5) frequent omissions of important statistical information. Some editors responded to Pickett's concerns by retracting or correcting certain articles, some of the problems having been acknowledged by those articles' authors. It has been argued that it would be very desirable to develop statistical measures to permit the identification of fabricated or manipulated data [17] (p. 193).

Therefore, our most general research question was to find new ways to detect potentially fraudulent research using statistical methods. More specifically, our primary objective was to ask this research question and test the following general hypothesis: *do retracted articles differ from control articles in statistically significant ways?* We tested three specific hypotheses:

Hypothesis 1. *The retracted group of articles will differ from the control group of articles in terms of six anomalies, measured as percentages and by ordinal breakdowns of those percentages, including two scales derived from two different sums of the six anomalies.*

Hypothesis 2. *Comparison of the two groups of articles using expected values of first digits of regression coefficients (Benford's Law) will yield larger deviations from expected values for the retracted group of articles than for the control group of articles.*

Hypothesis 3. *Measures of deviations from Benford's Law will be correlated significantly with the two scales derived from the six ratings of the anomalies.*

2. Methods

2.1. Sample

A sample of articles was developed cumulatively. Seven articles were among those retracted or corrected, as reported by Pickett [16], though six of which are also available in the Retraction Watch database [https://retractionwatch.com (accessed on 1 February 2023)]. The corrected article was by Mears, Stewart, Warren, and Simons [18]; the remaining six were retracted [19–24]. Thus, Pickett was the original instigator calling for the retraction of the articles [18–24] but journal editors made the final decisions on retractions or corrections. We selected a set of eight articles as control articles, written by many of the same authors who wrote the retracted articles, including Gertz, Mears, Pickett, and Simons [25–32]. The control articles were selected by searching Google Scholar for articles related to criminology by co-authors of Dr. Stewart. The total sample came to 15 articles, listed in Table 1.

Table 1. Basic descriptive information for articles reviewed.

Reference Number	Typology	Authors	Year	Sample Size	Grant-Funded	Google Cites
18	Corrected	Mears, Stewart, Warren, Simons	2017	784	YES	49
19	Retracted	Stewart	2003	10,578	NO	543
20	Retracted	Johnson, Stewart, Pickett, Gertz	2011	1184	NO	92

Table 1. *Cont.*

Reference Number	Typology	Authors	Year	Sample Size	Grant-Funded	Google Cites
21	Retracted	Stewart, Martinez, Baumer, Gertz	2015	1186	YES	67
22	Retracted	Stewart, Mears, Warren, Baumer, Arnio	2018	1441	NO	19
23	Retracted	Mears, Stewart, Warren, Craig, Arnio	2019	1301	NO	13
24	Retracted	Stewart, Johnson, Warren, Rosario, Hughes	2019	2408	NO	4
25	Control	Simons, Lin, Gordon, Brody, Murry, Conger	2002	841	YES	291
26	Control	Mears, Pickett, Golden, Chiricos, Gertz	2013	520	YES	32
27	Control	Pickett, Mancini, Mears, Gertz	2015	1308	NO	81
28	Control	Metcalfe, Pickett, Mancini	2015	540	NO	46
29	Control	Mancini, Pickett	2016	537	NO	24
30	Control	Pickett, Chiricos, Golden, Gertz	2012	1273	NO	41
31	Control	Shi, Lu, Pickett	2020	422,504	NO	25
32	Control	Pickett, Mancini, Mears	2013	499	NO	30

2.2. Measurement

2.2.1. Summary Descriptives

The number of Google citations for each article as of 5 January 2023 were recorded via a search of each article in Google Scholar. The year of publication of each article was recorded from an inspection of the article and the citation from Google Scholar. Whether the article said it was supported by a state or federal grant was determined through an inspection of each article and credits for grant support. The total number of authors of each article was included from a count of the authors listed for each article. Each article was coded as either a control article (coded as 0) or a retracted/corrected article (coded as a 1). Sample size was derived from author reports within each article, using the largest sample available if more than one sample were used. Total pages used was assessed through page counts. See Table 2.

Table 2. Sample data summary characteristics for 15 articles used in this study.

Variable	Mean	Standard Deviation	Median	Range of Scores
Google Citations (Jan 23)	115.60	146.99	51.00	4–548
Year Article Published	2013.87	5.33	2015.00	2002–2020
Total Number Authors	3.87	1.30	4.00	1–6
Total Pages Used	28.40	6.39	28.00	16–41
Sample Size	29,793.60	108,668.82	1186.00	499–422,504
Grant Supported	0.267	0.458	0.00	0–1
(0 = No, 1 = Yes)				

NOTE: Comparing the sample characteristics using *t*-test, and Mann–Whitney U test, across the retracted/control groups; none of the results were significant ($p < 0.05$).

2.2.2. Individual Measures of Anomalies

Several measures were created to identify possible anomalies in the 15 articles under consideration. These are reported as percentages in Table 3, but were analyzed in decimal form (i.e., 53.4% = 0.534).

Table 3. Characteristics of anomalies.

Reference Number	Typology	HC %	Zero %	SE Rate	B Rate	% Bad Binary	% Close Binary	% Close or Bad Binary
18	Corrected	0	2.78	Avoided	69.23	50.00	37.50	87.50
19	Retracted	100.00	0.00	93.33	93.33	100.00	0.00	100.00

Table 3. *Cont.*

Reference Number	Typology	HC %	Zero %	SE Rate	B Rate	% Bad Binary	% Close Binary	% Close or Bad Binary
20	Retracted	0	2.00	90.48	54.76	53.33	13.33	66.67
21	Retracted	91.38	1.75	94.44	66.67	54.55	9.09	63.64
22	Retracted	0	0.191	99.40	54.17	66.67	0.00	66.67
23	Retracted	0	12.05	54.31	18.97	Avoided	Avoided	Avoided
24	Retracted	0	0.00	97.62	56.35	58.82	5.88	64.71
25	Control	0	10.00	Avoided	Avoided	0	0	0
26	Control	0	10.96	10.00	0	0	12.50	12.50
27	Control	0	6.59	0	0	0	11.76	11.76
28	Control	0	8.14	0	0	0	0	0
29	Control	0	7.81	52.94	13.73	0	0	0
30	Control	0	8.96	38.68	14.15	0	8.33	8.33
31	Control	0	8.54	8.70	4.35	Avoided	Avoided	Avoided
32	Control	0	8.14	Avoided	7.81	0	0	0

2.2.3. Hand Calculation

Hand calculation was measured by dividing unstandardized regression coefficients by their standard errors [B/SE] and attending to whether the reported t-value was replicated exactly to two or three decimals. Brown and Heathers [33] have provided more details on this issue of hand calculation.

2.2.4. Excess Identical Unstandardized Regression Coefficients {Betas} or Standard Errors

An excess of identical betas or SEs was determined by calculating how many adjacent identical pairs were possible and creating a ratio of identical pairs to all possible identical adjacent pairs. If a table had five models, that would mean that each row could have four adjacent identical pairs, etc. If all SE's were identical across all rows and columns, then the ratio would be 1.0. Other approaches that we did not use might have counted how many parameters were the same across a row of results, even if not adjacent or counted as a match if the last parameter in a row matched the first parameter in that row.

2.2.5. Shortage/Excess of Zeroes in Terminal Digits of Regression Coefficients or Standard Errors

A shortage (or excess) of zeros in terminal digits [34,35] was determined by counting all the listed data points (regression coefficients, standard errors) in the regression tables (not including data from correlation matrices, factor loadings, odds ratios, *t*-tests, other test statistics [e.g., Exp(b)], intercept values, and squared variance values) that had two or three decimal points, and counting how many ended in a digit of zero. The ratio was turned into a percentage. We did not assess terminal zeroes in tables of means and standard deviations. Pickett [16] has reported that the retracted articles appeared to avoid zeroes as terminal digits; therefore, we focused on that issue rather than unusually high or low values for the digits 1 to 9, which could also suggest data problems [34]. Research in other situations might investigate shortfalls in all digits rather than just zero.

2.2.6. Mathematically Incorrect Standard Deviations for Binary Variables

All binary variables were checked to see if the standard deviations were computed correctly from the reported mean values. A ratio was determined from those not calculated correctly compared to all binary variables used and turned into a percentage. In one or two articles, the authors reported binary mean scores but did not report standard deviations.

Binary results were coded as "bad" if off by more than 0.02; incorrect results off by 0.02 or less were coded as close. We created two variables: one from the percentage of "bad" results, and one from the total percentage of "bad" and "close" results.

2.2.7. Benford's Law Deviations

Another method available for detecting fraudulent research has been discussed elsewhere in more detail [8,36,37]. Benford's law indicates that the left-most digits in a genuine set of data will follow a pattern of declining percentages from 1 to 9, as follows: 30.1029996, 17.6091259, 12.4938737, 9.6910013, 7.9181246, 6.6946790, 5.7991947, 5.1152522, and 4.5757491 ([36]: from Table 1, p. 110; Table 7, p. 118). However, Benford's Law may be most useful for fraud detection when fraud is rampant, when the first three digits are considered rather than just the first digit [38], and when using unstandardized regression coefficients [39]. Results have been mixed with respect to using Benford's Law for detecting scientific fraud [17]. Benford's Law has been used to validate, as well as to raise suspicions about, published research [40]. Absolute values of differences for initial digits in regression coefficients compared to expected values for Benford's Law were summed for nine (DIFF9) and three (DIFF3) digits. For example, suppose an article featured 60 regression coefficients, of which 20, 10, and 6 featured left-hand digits of 1, 2, and 3, respectively, for percentages of 33.3, 16.7, and 10.0. Taking the absolute differences between Benford's Law in decimal form, computed to the fifth decimal point, would yield the sum of absolute values of $[(0.33333 - 0.30103 = 0.03230) + (0.16667 - 0.17609 = 0.00942) + (0.10000 - 0.12494 = 0.02494)] = 0.06660$. Thus, DIFF3 for that article would be 0.06660; we did not divide by three to average the differences. We initially applied Benford's Law to means, standard deviations, regression coefficients, and standard errors but found little relationship with other anomalies except in the case of regression coefficients (mostly unstandardized in the retracted and control group articles).

2.3. Creation of Ordinal Anomaly Scales

To expedite an overall analysis of the anomalies, percentage values were converted to ordinal measures, using the terms no issue (coded 0), avoided (1), slight (2), moderate (3), and major (4), as presented in Table 4 below.

2.3.1. Missing Data

The term avoided was used when a series of parameters could have been reported but were not. In some articles, binary variable means were reported but not their standard deviations. In other cases, beta coefficients were reported, but not their standard errors. Because avoiding obvious statistics would be an issue in itself, we coded that situation as 1. In most cases, tables of results would present more than one column of data, each column representing a different model, allowing for comparison of regression coefficients and standard errors from one model in the table to another model in the same table. However, in one article [19], the two models were presented in separate tables and were thus compared.

2.3.2. Hand Calculation, Regression Coefficients, and Standard Errors

For the percentages associated with hand calculation, regression coefficients, and standard errors, ordinal items were created by coding the percentages for those items as follows: from 0.0 to 5.99% was coded as no issue, from 6.00 to 29.99% was coded as a slight issue, from 30.0% to 65.99% were coded as a moderate issue, with 66% or more being coded as a major issue. For example, if 34% of the possible adjacent standard errors were identical to three digits in an article, the standard error variable would be coded as "moderate" for that article. More leeway was granted for the other variables because some situations would be more likely to occur naturally and only more extreme situations would be indicative of serious problems.

2.3.3. Shortage/Excess of Zeroes

For the zero's variable, the recoding scheme was centered on the expected value of 10%, such that both sets of percentages, less than 3% and more than 20%, were coded as a major issue (major deviation from the expected value), and from 3% to 4.99% and from 15% to 19.99% were both coded as moderate deviations. Values between 5.00% and 6.99% and between 13% and 14.99% were coded as slight deviations. Values from 7% to 12.99% were coded as not an issue. The coding pattern was not symmetric because there were no values above 13.1% and we wanted to make some distinctions among those less frequent values rather than coding them identically. For example, if an article contained 200 regression coefficients and their standard errors and only 2 of them ended in a digit of zero, the value of 1.0% for zeroes would be coded as "major".

2.3.4. Binary Variable Standard Deviations Relative to Their Means

For the "bad" binary variable, coding was 0% (no issue), 0.01% to 24.99% was coded as a slight issue, 25.0% to 49.99% was coded as a moderate issue, and 50% or more was coded as a major issue. The coding was more sensitive because accurate computer calculations should seldom, if ever, make a major error in calculating the standard deviations for binary variables. For example, if there were ten binary variables reported in an article and four of the standard deviations were in error by 0.05 units and one was in error by 0.01 units, then the "bad" binary variable (i.e., errors > 0.02) would be coded as "moderate" while the total "bad and close" (i.e., errors ≥ 0.01) binary variable would be coded as "major".

2.3.5. Benford's Law Measurement

For each article, the percentage of first digits that were 1, 2, 3, 4, 5, 6, 7, 8, and 9 was calculated. Our first measure related to Benford's Law was those percentages averaged across all of the articles (k = 15), the retracted articles (k = 7), and the control group articles (k = 8). Those results were compared to the expectations of Benford's Law and the absolute values of the differences summed. Next, the absolute value of the difference between the expectations of Benford's Law and the results for each of the nine digits was calculated. For one measure, the sum of the absolute values of the differences across all nine digits was calculated (DIFF9); for a second measure, the sum of the absolute values was calculated for only digits 1, 2, and 3 (DIFF3).

2.3.6. Total Anomalies Scale

The total anomaly scale was computed by adding the ordinal scores for hand calculation, percentage of zeros, percentage of adjacent standard errors, percentage of adjacent betas, percentage of incorrect binary standard deviations (>0.02), and percentage of incorrect binary standard deviations (≥0.01). Measurement characteristics for this scale are reported in Section 3.3.3.

2.3.7. Anomaly Severity Scale

An anomaly severity scale score was also developed by coding avoided or slight ratings as 0.25, moderate ratings as 0.50, and major ratings as 1.0, summing them across each of the six measures of anomalies. Measurement characteristics for this scale are reported in Section 3.3.4.

2.4. Analyses

Pearson zero-order correlations were used to correlate the key variables, while *t*-tests were used to compare scores for the group of retracted articles versus the scores for the control group of articles. SPSS 28.0 was used for all statistical calculations, including the calculation of Cohen's d [41,42], to assess effect sizes, using the convention from 0.50 to 0.79 as a moderate effect size and 0.80 or greater as a large effect size. A repeated measures analysis with the group (retracted vs. control) as a between subjects variable and digit percentages over nine digits as a within subjects factor was used to assess main effects and the group via digit percentages interaction term. The SPSS SCALE/RELIABILITY program was used to calculate Cronbach's alpha, a measure of the internal consistency reliability of scales. A website [www.escal.site] was used to convert correlations to Cohen's d for the equivalent effect size of the correlations.

3. Results

3.1. Descriptive Data and Retraction Status

The authors, year of publication, sample size, grant funding, and google cites as of 5 January 2023 have been presented in Table 1.

3.2. Comparing Retracted and Control Articles' Data

Basic descriptive statistics from the variables in Table 1 were presented in Table 2. For the variables presented in Table 2, compared across the retracted and control group articles, there were no significant differences as a function of retracted status.

3.3. Measurement

3.3.1. Anomaly Percentage Values

Table 3 earlier presented the results of classifying each article under consideration in terms of percentage levels of each type of anomaly measured.

3.3.2. Anomaly Ordinal Values

Table 4 presents the results of classifying each article under consideration in terms of ordinal levels of each type of anomaly measured.

Table 4. Summary of anomalies in ordinal measurement.

Reference Number	Typology	HC	Zeros	SE	Beta	Bad Binary	Close or Bad Binary	Anomaly Scale	Severity Measure
18	Corrected	No issue	Major	Avoided	Major	Major	Major	17.00	4.25
19	Retracted	Major	Major	Major	Major	Major	Major	24.00	6.00
20	Retracted	No Issue	Major	Major	Moderate	Major	Major	19.00	4.50
21	Retracted	Major	Major	Major	Major	Major	Major	24.00	6.00
22	Retracted	No Issue	Major	Major	Moderate	Major	Major	19.00	4.50
23	Retracted	No Issue	No Issue	Moderate	Slight	Avoided	Avoided	7.00	1.25
24	Retracted	No Issue	Major	Major	Moderate	Major	Major	19.00	4.50
25	Control	No issue	No issue	Avoided	Avoided	No issue	No issue	2.00	0.50
26	Control	No issue	No issue	Slight	No issue	No issue	Slight	4.00	0.50
27	Control	No issue	Slight	No issue	No issue	No issue	Slight	4.00	0.50
28	Control	No issue	No issue	No issue	No issue	No issue	No issue	0.00	0.00
29	Control	No issue	No issue	Moderate	Slight	No issue	No issue	5.00	0.75
30	Control	No issue	No issue	Moderate	Slight	No issue	Slight	7.00	1.00
31	Control	No issue	No issue	Slight	No issue	Avoided	Avoided	4.00	0.75
32	Control	No issue	No issue	Avoided	Slight	No issue	No issue	3.00	0.50

3.3.3. Total Anomalies Scale

The characteristics of the total anomalies scale were a mean of 10.53 (SD = 8.63; range, 0–24; median, 7.00). The Cronbach alpha for the anomaly scale was 0.92 and would be 0.94 if the hand-calculated anomaly rating were not included in the scale. Deleting any one of the other five items changed the value of the alpha from between 0.89 to 0.92. The total anomaly scale scores ranged between zero and seven for the control group and between seven and 24 for the retracted/corrected group of articles. The correlation between retracted status and the total anomaly scale score was r = 0.89 ($p < 0.001$), an indication of predictive validity.

3.3.4. Anomaly Severity Scale

The characteristics of the anomaly severity scale were a mean of 2.37 (SD = 2.26, range 0–6, median = 1.00). The Cronbach alpha for the severity scale was 0.94 and would be 0.96 if the hand-calculated severity rating were not included in the scale. Deleting any one of the other five items changed the value of the Cronbach alpha between 0.92 and 0.94. The anomaly severity scores ranged between zero and 1.0 for the control group and between 1.25 and 6.00 for the retracted group of articles. This severity scale was also correlated 0.88 ($p < 0.001$) with retracted status, an indication of predictive validity.

3.4. Retraction Status and Anomalies

Table 5 shows the percentages (expressed in decimals (0 = 0%, 1 = 100%) for each of the six anomalies as well as the ordinal recoding of the percentages, compared across the two groups. Results for the hypotheses are presented below.

Table 5. Differences between control and retracted articles on anomaly variables.

Variables	Control (N = 8)		Retracted (N = 7)		d	t	df	*p*
	X	SD	X	SD				
Comparing Percentages in decimals for:								
Hand Calculated Tests	0	0	0.27	0.47	0.86	1.55	6	0.086
Wide Error Binary SDs	0	0	0.64	0.19	5.1	8.42	5	<0.001
Wide and Narrow Error	0.05	0.06	0.75	0.15	6.3	10.64	6.31	<0.001
Binary SDs Zeroes	0.09	0.01	0.03	0.04	1.94	3.75	13	0.001
Adjacent Identical B's	0.06	0.06	0.6	0.24	3.1	5.8	12	<0.002
Adjacent Identical SE's	0.18	0.22	0.88	0.17	3.55	6.15	10	<0.001
Comparing Anomaly Ratings for:								
Hand-Calculated	0	0	1.14	1.95	0.86	1.55	6	0.086
Wide Error Binary SDs	0.13	0.35	3.57	1.13	5.1	8.19	13	<0.001
Wide and Narrow Error Binary SDs	0.88	0.99	3.57	1.13	2.55	4.92	13	<0.003
Zeroes	0.25	0.71	3.43	1.51	2.76	5.34	13	<0.001
Adjacent Identical B's	0.88	0.99	3.29	0.76	2.71	5.23	13	<0.001
Adjacent Identical SE's	1.5	1.2	3.43	1.13	1.65	3.19	13	0.004
Anomaly Severity Scale	0.56	0.29	4.43	1.59	3.52	6.8	13	<0.001
Total Anomaly Scale	3.63	2.07	18.43	5.71	3.55	6.87	13	<0.001

One-sided *t*-tests were used given the a priori assumption that retracted articles would feature more problems than the control articles. When Levene's test for equality of variances was violated, separate variance estimates were used for the reported *t*-tests and degrees of freedom. Numbers are rounded up from 5 or higher in the third or fourth digit. Missing data are reflected in the degrees of freedom reported.

3.4.1. Hypothesis 1

The first hypothesis was that the retracted group of articles would differ from the control group of articles in terms of six anomalies, measured as percentages and by ordinal breakdowns of those percentages, including two scales derived from two different sums of the six anomalies.

Table 5 presents the results for hypothesis 1. The results for the anomalies in terms of decimal percentages were significant ($p < 0.05$) except for hand calculation ($p < 0.09$) using one-tailed tests. Using two-tailed tests, the other tests would remain significant. The effect sizes ranged between 0.86 and 6.30, above the "large" size [41,42]. Using nonparametric Mann–Whitney U tests to compare the percentage ratings, all results were significant ($p < 0.05$, two-tailed) except for hand calculation.

The results for the ordinal ratings of the anomalies were all significant ($p < 0.05$, one-tailed) except for hand calculation ($p < 0.09$). Effect sizes ranged between 0.86 and 5.10, above the "large" size [41,42]. Using Mann-Whitney U-tests to compare the ratings, all results were significant ($p < 0.05$, two-tailed) except for hand calculation.

The *t*-test results for the two overall measures of anomalies were both significant ($p < 0.001$, one-tailed) with effect sizes between 3.52 and 3.55, far above Cohen's [41,42] "large" size. Using the nonparametric Mann–Whitney U-test, both results were significant ($p = 0.006$, two-sided).

Even though our results for the hand-calculation anomaly were not significant, Appendix A illustrates the difference between computer-generated results and hand calculation; 20% of the computer generated t-values differed from the hand-calculated values.

Thus, our results supported hypothesis 1 in terms of statistical significance and in terms of large effect sizes for all anomaly variables and scales except for hand calculation.

3.4.2. Hypothesis 2

The second hypothesis was that a comparison of the two groups of articles using expected values of first digits of regression coefficients (Benford's Law) would yield larger deviations from expected values for the retracted group of articles than for the control group of articles.

Deviations from Benford's Law were assessed as shown in Table 6. Although the effect sizes were in the "large" range (0.90 and 1.08), the *t*-test results were marginally significant (0.053, 0.029, one-tailed). A Mann–Whitney U test obtained a two-sided exact significance result of 0.040 for DIFF3, while the result for DIFF9 was not significant.

Table 6. Using Benford's Law to compare retracted and control groups of articles.

Variables	Control Group		Retracted Group		df	t	*p*	d
	Mean	SD	Mean	SD				
Benford discrepancies Over all nine digits (DIFF9)	0.3445	0.1577	0.488	0.1618	13	1.74	0.053	0.9
Benford discrepancies Over digits 1, 2, and 3 (DIFF3)	0.1418	0.068	0.2392	0.1111	13	2.08	0.029	1.08

NOTE: T-tests are one-tailed. Effect sizes are reported, using Cohen's d. Repeated measures analyses of variance were computed but the only consistent significant effect was a linear ($p < 0.001$) and quadratic ($p < 0.05$) main effect trend for DIFF9, which decreased from 0.072 to 0.038 from DIFF1 to DIFF 9, with values of 0.061, 0.054, 0.044, 0.036, 0.036, 0.035, and 0.035 for DIFF2 through DIFF8, respectively.

Thus, our results supported hypothesis 2 in terms of effect sizes but only partially in terms of significance levels. However, in terms of DIFF3, both the effect size and significance level supported hypothesis 2.

3.4.3. Hypothesis 3

The third hypothesis was that measures of deviations from Benford's Law would be correlated significantly with the two scales derived from the six ratings of the anomalies.

DIFF3 and DIFF9 were correlated 0.599 ($p = 0.018$, two-tailed, d = 1.50) and 0.472 ($p = 0.076$, two-tailed, d = 1.07) with the total anomaly scale; the respective correlations for the anomaly severity scale were 0.605 ($p = 0.017$, two-tailed, d = 1.52) and 0.486 ($p = 0.066$, two-tailed, d = 1.11). DIFF9 and DIFF3 correlated 0.434 ($p = 0.106$, two-tailed, d = 0.97) and 0.500 ($p = 0.058$, two-tailed, d = 1.16), respectively, with retracted status, with neither being significant but with large effect sizes. The Spearman rho between DIFF3 and retracted status was 0.607 ($p = 0.016$, two-tailed, d = 1.53), however. DIFF9 and DIFF3 were correlated r = 0.810 ($p < 0.001$, two-tailed), Spearman rho = 0.646 ($p = 0.009$, two-tailed). Our results supported hypothesis 3 in terms of effect sizes, but partially with respect to significance levels using two-tailed tests.

3.4.4. Additional within Control Group Analysis

Visual inspection of the values for two of the control group articles [29,30] yielded a finding of relatively high values for beta and standard error anomalies. Results for comparing those two articles' anomalies with the same anomalies for the other six control articles are presented in Appendix B. On four of the six *t*-tests, there were significant differences between the two subdivided control groups, with Cohen's d ranging from 1.41 to 5.97. On the other two *t*-tests, had a difference variance estimate *t*-test been used, all six of the tests would have been significant. Comparing the two article control group with the retracted article group led to five of six tests being significant ($p < 0.05$), with Cohen's d ranging from 0.41 to 2.65, while comparing the six article control group with the retracted article group led to all of the comparisons being significant ($p < 0.005$) with Cohen's d ranging from 2.35 to 6.06.

3.4.5. Discriminant Analysis for Sensitivity and Specificity

We performed a discriminant analysis using the groups (Retracted/Control) and six anomaly variables. Entering all six variables at once into the analysis, 100% of the retracted articles were predicted as retracted (sensitivity), while 100% of the control articles were predicted as controls (specificity). However, one article in each group ([23], retracted; [27], controls) came close to being assigned to the other group; therefore, a more conservative approach would indicate sensitivity as low as 85.7% (6/7), with retracted articles predicted as retracted and specificity as low as 87.5% (7/8) with control articles predicted as controls.

4. Discussion

Good data are hard to fake. Good data may have systematic patterns but will also have randomness; therefore, too much or too little consistency may signal that something is amiss. Among the seven retracted or corrected articles discussed here, most had outstanding reviews of the literature, convincing theory, and reasonable, useful conclusions, even more total pages than might be typical. However, high quality of narrative portions, even the theoretical portions, and useful conclusions do not guarantee valid data or valid statistical analysis.

Most of the results for our hypotheses were statistically significant. Our total measures of anomalies or of anomaly severity yielded significant differences as a function of retracted status, with substantial (d > 3.50) effect sizes. Results for violations of Benford's Law were mixed, but promising for larger samples using unstandardized regression coefficients, especially for the use of deviations for the left-most digits of 1, 2, and 3 (DIFF3), for which correlations with both anomaly scales were significant ($p < 0.05$), while also significant for rho ($p < 0.05$) but not r ($p < 0.06$) with respect to retracted status. This methodology appears capable of detecting unusual anomalies with high sensitivity and specificity in both retracted [19–24] and non-retracted articles [29,30] even though the articles were published by a variety of scholars.

When time is not adequate to permit detailed testing for anomalies, we would suggest some rules of thumb for editors and reviewers: for apparent cases of hand calculation of results, or adjacent identical regression coefficients or standard errors, we would suggest using 50% or more as a rule of thumb to suggest serious problems. For binary variables, if 50% or more are inaccurate or inconsistent (e.g., means of 0.35 and 0.47 both have standard deviations of 0.50), then we would suspect serious problems. In the case of second or third (presumably random) decimal points for regression coefficients or standard errors, we would question any situation in which 2% or less of any number from 0 to 9 were represented. Benford's Law is more difficult to simplify, but we would suggest that if the percentage of left-most digits of "1" are found to be below 20% or above 40%, there should be further investigation. Such levels would be "red flags" to us; other levels might still raise questions, especially if several of these rules of thumb were violated at the same time within any one paper.

Thus, our research provides scholars with several ways to detect anomalies that may help detect falsified or fabricated data or results, using either several detailed statistical approaches or several more simple rules of thumb for assessing the extent of unusual anomalies.

Author Contributions: Conceptualization, W.R.S. and D.W.C.; methodology, W.R.S., D.W.C., L.L., A.A. and A.b.A.; formal analysis, W.R.S.; investigation, W.R.S.; writing-original draft preparation, W.R.S., D.W.C., L.L., A.A. and A.b.A.; writing-review and editing, W.R.S., D.W.C., L.L., A.A. and A.b.A.; supervision, W.R.S.; project administration, W.R.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was not supported by any external or internal funding.

Data Availability Statement: Data used in calculations in this report are available in the tables. Original raw data may be requested from Dr. Walter R. Schumm, schumm@ksu.edu.

Acknowledgments: The opinions, findings, and conclusions or recommendations made in this report are those of the authors, and may not reflect those of the Department of Applied Human Sciences (formerly the School of Family Studies and Human Services), the College of Health and Human Sciences (formerly the College of Human Ecology), or of Kansas State University or the National Council on Family Relations. The findings in part of this report were presented at the Theory Construction and Research Methodology Preconference Workshop at the Annual Conference of the National Council on Family Relations, Fort Worth, Texas, 20 November, 2019, by the third author. Other findings in this report were presented at the annual conference of the Society for the Improvement of Psychological Science, Victoria, British Columbia, 22 June, 2020, by the first author. Appreciation is expressed to Justin Pickett, Ben Silliman, and those scholars who reviewed this report for their very helpful insights and comments.

Conflicts of Interest: No potential conflict of interest was reported by the authors.

Appendix A. Example of Hand Calculation versus Computer Generation: Predicting the Total Anomaly Scale from Several Independent Variables

Independent	B	SE	t Values		β	*p*
Variable			COMP	HC		
Year Published	1.212	1.187	1.021	1.021	0.748	0.344
Grant Supported	7.04	6.914	1.018	1.018	0.373	0.335
Total Pages Used in Articles	0.469	0.418	1.122	1.122	0.347	0.291
Google Citation Count	0.05	0.046	1.096	1.087	0.85	0.301
Total Number of Authors Per Article	0.504	2.629	0.192	0.192	0.076	0.852
F(5, 9) = 0.557, *p* = 0.731, R Square = 0.236						

COMP = computer-generated t value; HC = hand-calculated t value. All values are computer-generated except for the HC t value.

Appendix B. Within Control Group Comparisons on Beta and SE Rates, Anomalies, and Severities

Variables	Control (N = 6)		Control (N = 2)		d	t	df	*p*
	X	SD	X	SD				
Beta Rate	0.02	0.04	0.14	0.00	3.62	7.19	4.14	0.002
Beta Anomaly	0.50	0.84	2.00	0.00	1.96	2.41	6.00	0.053
Beta Severity	0.08	0.13	0.25	0.00	1.41	3.16	5.00	0.025
SE Rate	0.05	0.05	0.46	0.10	5.97	5.39	1.30	0.076
SE Anomaly	1.00	0.89	3.00	0.00	2.45	3.00	6.00	0.024
SE Severity	0.17	0.13	0.50	0.00	2.83	6.33	5.00	0.001

All *t*-tests feature two-tailed significance levels. The *t*-test for beta anomaly had we used the equal variance *t*-test, t(5) = 4.39 ($p = 0.007$); we used the unequal variance *t*-test because in the Levene test for homogeneity of variance's $p = 0.07$. Because in the Levene test $p = 0.004$ for SE Rate, we reported the unequal variance *t*-test. Had we used the equal variance *t*-test, t(4) = 6.89 ($p = 0.002$). One-way analysis of variance tests, including the retracted article scores, were all significant, $p < 0.005$. Comparing the six control articles versus the retracted articles on the above six variables, all *t*-tests were significant, $p < 0.005$. Comparing the two control articles versus the retracted articles on the above six variables, all but SE anomaly were significant, $p < 0.05$.

References

1. Page, G.D.; Columb, M.O. Fake News, Zombie Papers, and Fabricated Evidence: A Thoroughly Modern Pandemic? *Eur. J. Anaesthesiol.* **2022**, *39*, 302–304. [CrossRef]
2. Bordewijk, E.M.; Li, W.; Van Eekelen, R.; Wang, R.; Showell, M.; Mol, B.W.; Van Wely, M. Methods To Assess Research Misconduct in Health-Related Research: A Scoping Review. *J. Clin. Epidemiol.* **2021**, *136*, 189–202. [CrossRef] [PubMed]
3. Boetto, E.; Golinelli, D.; Carullo, G.; Fantini, M.P. Frauds in Scientific Research and How to Possibly Overcome Them. *J. Med. Ethics* **2021**, *47*, e19. [CrossRef]
4. Yeo-The, N.S.L.; Tang, B.L. Sustained Rise in Retractions in the Life Sciences Literature During the Pandemic Years 2020 and 2021. *Publications* **2022**, *10*, 29. [CrossRef]
5. Fanelli, D. How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta- Analysis of Survey Data. *PLoS ONE* **2009**, *4*, e5738. [CrossRef] [PubMed]
6. Fanelli, D. Why Growing Retractions Are (Mostly) a Good Sign. *PLoS Med.* **2013**, *10*, e1001563. [CrossRef] [PubMed]
7. Hartgerink, C.H.J.; Wicherts, J.M. Research Practices and Assessment of Research Misconduct. *Sci. Open Res.* **2016**, 1–10. [CrossRef]
8. Horton, J.; Kumar, D.K.; Wood, A. Detecting Academic Fraud Using Benford's Law: The Case of Professor James Hunton. *Res. Policy* **2020**, *49*, 104084. [CrossRef]
9. Steen, R.G.; Casadevall, A.; Fang, F.C. Why Has the Number of Scientific Retractions Increased? *PLoS ONE* **2013**, *8*, e68397. [CrossRef]
10. Stroebe, W.; Postmes, T.; Spears, R. Scientific Misconduct and the Myth of Self-Correction in Science. *Perspect. Psychol. Sci.* **2012**, *7*, 670–688. [CrossRef]
11. Wiedermann, C.J. Inaction Over Retractions of Identified Fraudulent Publications: Ongoing Weakness in the System of Scientific Self-Correction. *Account. Res.* **2018**, *25*, 239–253. [CrossRef] [PubMed]
12. Stern, A.M.; Casadevall, A.; Steen, R.G.; Fang, F.C. Financial Costs and Personal Consequences of Research Misconduct Resulting in Retracted Publications. *eLife* **2014**, *3*, e02956. [CrossRef] [PubMed]
13. Poutoglidou, F.; Stavrakas, M.; Tsetsos, N.; Poutoglidis, A.; Tsentemeidou, A.; Fyrmpas, G.; Karkos, P.D. Fraud and Deceit in Medical Research: Insights and Current Perspectives. *Voices Bioeth.* **2022**, *8*, 1–6. [CrossRef]
14. Nurunnabi, M.; Hossain, M.A. Data Falsification and Questions on Academic Integrity. *Account. Res.* **2019**, *26*, 108–122. [CrossRef]
15. Mistry, V.; Grey, A.; Bolland, M.J. Publication Rates After the First Retraction for Biomedical Researchers with Multiple Retracted Publications. *Account. Res.* **2019**, *26*, 277–287. [CrossRef]
16. Pickett, J.T. The Stewart Retractions: A Quantitative and Qualitative Analysis. *Econ. Watch J.* **2020**, *17*, 152–190.
17. Auspurg, K.; Hinz, T. Social Dilemmas in Science: Detecting Misconduct and Finding Institutional Solutions. In *Social Dilemmas, Institutions, and the Evolution of Cooperation*; Jann, B., Przepiorka, W., Eds.; DeGruyter Oldenbourg: Berlin, Germany; Boston, MA, USA, 2017.
18. Mears, D.P.; Stewart, E.A.; Warren, P.Y.; Simons, R.L. Culture and Formal Social Control: The Effect of the Code of the Street on Police and Court Decision-Making. *Justice Q.* **2017**, *34*, 217–247. [CrossRef]
19. Stewart, E.A. School Social Bonds, School Climate, and School Misbehavior: A Multilevel Analysis. *Justice Q.* **2003**, *20*, 575–604. [CrossRef]
20. Johnson, B.D.; Stewart, E.A.; Pickett, J.; Gertz, M. Ethnic Threat and Social Control: Examining Public Support for Judicial Use of Ethnicity in Punishment. *Criminology* **2011**, *49*, 401–441. [CrossRef]
21. Stewart, E.A.; Martinez, R., Jr.; Baumer, E.P.; Gertz, M. The Social Context of Latino Threat and Punitive Latino Sentiment. *Soc. Probl.* **2015**, *62*, 68–92. [CrossRef]
22. Stewart, E.A.; Mears, D.P.; Warren, P.Y.; Baumer, E.P.; Arnio, A.N. Lynchings, Racial Threat, and Whites' Punitive Views Toward Blacks. *Criminology* **2018**, *56*, 455–480. [CrossRef]
23. Mears, D.P.; Stewart, E.A.; Warren, P.Y.; Craig, M.O.; Arnio, A.N. A Legacy of Lynchings: Perceived Criminal Threat Among Whites. *Law Soc. Rev.* **2019**, *53*, 487–517. [CrossRef]
24. Stewart, E.A.; Johnson, B.D.; Warren, P.Y.; Rosario, J.L.; Hughes, C. The Social Context of Criminal Threat, Victim Race, and Punitive Black and Latino Sentiment. *Soc. Probl.* **2019**, *66*, 194–221. [CrossRef]
25. Simons, R.L.; Lin, K.H.; Gordon, L.C.; Brody, G.H.; Murry, V.; Conger, R.D. Community Differences in the Association Between Parenting Practices and Child Conduct Problems. *J. Marriage Fam.* **2002**, *64*, 331–345. [CrossRef]
26. Mears, D.P.; Pickett, J.T.; Golden, K.; Chiricos, T.; Gertz, M. The Effect of Interracial Contact Whites' Perceptions of Victimization Risk and Black Criminality. *J. Res. Crime Delinq.* **2013**, *50*, 272–299. [CrossRef]
27. Pickett, J.T.; Mancini, C.; Mears, D.P.; Gertz, M. Public (Mis)understanding of Crime Policy: The Effects of Criminal Justice Experience and Media Reliance. *Crim. Justice Policy Rev.* **2015**, *26*, 500–522. [CrossRef]
28. Metcalfe, C.; Pickett, J.T.; Mancini, C. Using Path Analysis to Explain Racialized Support for Punitive Delinquency Policies. *J. Quant. Criminol.* **2015**, *31*, 699–725. [CrossRef]
29. Mancini, C.; Pickett, J.T. The Good, the Bad, and the Incomprehensible: Typification of Victims and Offenders as Antecedents of Beliefs about Sex Crime. *J. Interpers. Violence* **2016**, *31*, 257–281. [CrossRef]

30. Pickett, J.T.; Chiricos, T.; Golden, K.M.; Gertz, M. Reconsidering the Relationship Between Perceived Neighborhood Racial Composition and Whites' Perceptions of Victimization Risk: Do Racial Stereotypes Matter? *Criminology* **2012**, *50*, 145–186. [CrossRef]
31. Shi, L.; Lu, Y.; Pickett, J.T. The Public Salience of Crime, 1960–2014: Age-Period-Cohort and Time-Series Analyses. *Criminology* **2020**, *58*, 568–593. [CrossRef]
32. Pickett, J.T.; Mancini, C.; Mears, D.P. Vulnerable Victims, Monstrous Offenders, and Unmanageable Risk: Explaining Public Opinion on the Social Control of Sex Crime. *Criminology* **2013**, *51*, 729–759. [CrossRef]
33. Brown, N.J.; Heathers, J.A. *Rounded Input Variables, Exact Test Statistics (RIVETS): A Technique for Detecting Hand-Calculated Results in Published Research*; Unpublished Paper; Bouve College of Health Sciences, Northeastern University: Boston, MA, USA, 2019.
34. Mosimann, J.E.; Dahlberg, J.E.; Davidian, N.M.; Krueger, J.W. Terminal Digits and the Examination of Questioned Data. *Account. Res.* **2002**, *9*, 75–92. [CrossRef]
35. Mosimann, J.E.; Wiseman, C.V.; Edelman, R.E. Data Fabrication: Can People Generate Random Digits? *Account. Res.* **1995**, *4*, 31–55. [CrossRef]
36. Dutta, A.; Choudhury, M.R.; De, A.K. A Unified Approach to Fraudulent Detection. *Int. J. Appl. Eng. Res.* **2022**, *17*, 110–124. [CrossRef]
37. Varian, H.R. Benford's Law. *Am. Stat.* **1972**, *26*, 65–66.
38. Diekmann, A. Not the First Digit! Using Benford's Law to Detect Fraudulent Scientific Data. *J. Appl. Stat.* **2007**, *34*, 321–329. [CrossRef]
39. Bauer, J.; Gross, J. Difficulties Detecting Fraud? The Use of Benford's Law on Regression Tables. *Jahrb. Fur Natl. Und Stat.* **2011**, *231*, 733–748.
40. Koch, C.; Okamura, K. Benford's Law and COVID-19 Reporting. *Econ. Lett.* **2020**, *196*, 109573. [CrossRef]
41. Cohen, J. A Power Primer. *Psychol. Bull.* **1992**, *112*, 155–159. [CrossRef]
42. Cohen, J. Statistical Power Analysis. *Curr. Dir. Psychol. Sci.* **1992**, *1*, 98–101. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.