Psychology Research and Applications, Vol. 3, No. 2, June 2021
https://dx.doi.org/10.22606/pra.2021.32001

Nine Ways to Detect Possible Scientific Misconduct in Research with Small (N < 200) Samples

Walter R. Schumm*, Duane W. Crawford, Lorenza Lockett, Abdullah AlRashed and Asma bin Ateeq

College of Health and Human Sciences, Kansas State University, Justin Hall, Manhattan, Kansas, USA
Email: schumm@ksu.edu

Abstract. Some scientists have fabricated their data, yet have published their fake results in peer-reviewed journals. How can we detect patterns typical of fabricated research? Nine relatively less complex ways for detecting potentially fabricated data in small samples (N < 200), are presented, using data from articles published since 1999 as illustrations. Even with smaller samples, there are several ways in which scholars, as well as their undergraduate and graduate students, can detect possible fabrication of data as well as other questionable research practices (QRPs). However, with larger samples, other techniques may be needed.

Keywords: research integrity, fraud, research misconduct, anomalous results, methods for detecting fraudulent research

1 Introduction

The number of articles retracted on account of scientific misconduct has increased in recent decades; some academics have had dozens of their articles retracted (Fanelli, 2009, 2013; Hartgerink & Wicherts, 2016; Steen, Casadevall, & Fang, 2013; Stroebe, Postmes, & Spears, 2012; Wiedermann, 2018), in spite of the serious consequences when scientific conduct is exposed (Pickett, 2020; Pickett & Roche, 2018; Stern, Casadevall, Steen, & Fang, 2014). For example, Diederik Stapel's research misconduct led to the retraction of 58 articles (Hartgerink & Wicherts, 2016). Joachim Boldt had nearly 90 articles retracted (Brainard, You, & Bonazzi, 2018). Sometimes scientific misconduct has involved fabrication of data, one of the most serious types of scientific misconduct, even if relatively rare (Nurunnabi & Hossain, 2019; Pickett, 2020; Pickett & Roche, 2018; Reisig, Holtfreter, & Berzofsky, 2020). Nurunnabi and Hossain (2020) listed over two dozen articles that have recently been retracted or had statements of concern dealing with scientific misconduct, including data fabrication. Authors of one or more retracted papers often have future papers retracted, suggesting there may be a pattern to their scientific misconduct (Kuroki and Ukawa, 2018).

2 Methods

What are reviewers and journal editors to do? We would like to suggest several formal and informal statistical methods for detecting data errors, including possible data misconduct or fabrication, for small samples. All methods will have limitations, of course, but our methods will give editors and reviewers more techniques with which to evaluate research papers and request appropriate explanations or material from authors, as needed. This way small technical errors can be corrected before publication, and any doubts about the legitimacy of data or methods may be clarified as well. We do recognize that there are far more advanced approaches being developed in certain fields of social science. However, we wanted to limit our approaches to those that could be assessed with hand calculations or readily available website calculations. Thus, we will present nine methods for detecting unusual results by (1) explaining the issue, (2) illustrating with examples, (3) noting limitations, and (4) making recommendations for each of the nine methods. Those readers interested in more complex statistical approaches to detecting data fabrication than presented here may want to consult Hartgerink, Wicherts, and van Assen (2016) and Hartgerink, Voelkel, Wicherts, and van Assen (2019).

3 Nine Approaches or "Clues" for Detecting Anomalous Data or Results

3.1 First Clue – Absence of Basic Statistical Information

Issue. Our first clue occurs when authors do not report their full results, even though detailed guidance has been provided by professional organizations on how to report data completely in journal articles (Appelbaum, Cooper, Kline, Mayo-Wilson, Nezu, & Rao, 2018). Even though we were all probably trained to report means, standard deviations, actual and theoretical ranges, degrees of freedom, sample sizes, significance levels, and statistical results (t, F, chi-square, etc.), some authors have not done so (Schumm, Crawford, Fawver, et al., 2019). If standard deviations are not reported, for example, one cannot independently calculate basic statistical tests (e.g., t-tests) nor can one determine effect sizes. Sometimes, authors report so little, it's hard to know what they found; an author might say only that a t-test was not significant (no means, no standard deviations, no sample sizes, no degrees of freedom reported).

Examples. We examined several retracted articles by Diederik Stapel (Crocker & Cooper, 2011; Stapel & Lindenberg, 2011a, b; Vogel, 2011) and for some of them we could not find any evidence of tampering with their data except for the absence of much of the statistical information that authors should provide. Another article we studied (Stewart, Simons, & Conger, 2000) did not report means or standard deviations for its binary variables, and did not report standard errors, so it was difficult to evaluate for questionable data. Authors who have fabricated their data may simply cover their tracks by refusing to provide any of their original data to journal editors or other scholars.

Limitations. A limitation of this first clue is that an absence of statistics doesn't prove a research report to be in error or to have been fabricated. It carries the limitation that editors or peer reviewers would have to ask for more information before processing manuscripts for further review, which would complicate and lengthen a journal's review process.

Recommendations. If authors, when contacted, are willing to provide more details about their results or their data, that is a positive sign of their good faith as scholars. On the other hand, if scholars are not willing to provide such material, then one might begin to question the validity of their data or their results. We have requested data from some scholars who have refused to provide anything, presumably lest we find flaws in their data or analyses. Other scholars have provided further information on their data, to their credit (Schumm & Crawford, 2020). Editors and peer reviewers, in a more powerful position than individual scholars, should not hesitate to ask for more information if a manuscript lacks basic statistical data.

3.2 Second Clue - Means or Percentages That Don't Make Sense

Issue. Our second clue, suggested by James Heathers (Brown & Heathers, 2017; Marcus & Oransky, 2018), labeled the granularity-related inconsistency of means test or the GRIM test (Brown, Kaiser, & Allison, 2018), is that, if the raw data involves only integers, the mean scores in a research report should be divisible by the sample size, meaning that if you multiply the mean score by the sample size, you should come close to a whole number (e.g. 4.50 x 100 = 450.0). Furthermore, the SPRITE program can check for illogical data in terms of either the mean or the standard deviation (Heathers, Anaya, van der Zee, & Brown, 2018). Both programs are free and web-based. Sometimes a mean might pass the GRIM test but the standard deviation would not be a possible value given the mean and sample size. It may be easier to check percentage results by hand, but they could also be checked by the GRIM and SPRITE tests.

Examples. As another example, Ruggiero, Steele, Hwang, & Marx (2000), whose article was later retracted for falsification of data (Ruggiero, Steele, Hwang, & Marx, 2001), reported many results, for which we will focus on two. One involved a mean of 4.14, SD = 2.67, n = 27. Those results fail the GRIM test since 4.14 x 27 = 111.78, *not* an integer. The second case involved a mean score of 1.42 with a standard deviation of 1.04 (n = 30), which also failed the GRIM test (1.42 x 30 = 42.6). Furthermore, in Ruggiero et al. (2000, p. 1275) for some rows in their Table 1, the mean scores fell within very narrow ranges: three rows of means, six values per row, with means in the top row to the bottom row with minimums and maximums as follows – (0.70 to 1.03), (8.57 to 9.02), and (5.47 to 5.70). It might be a

coincidence, for example, but the bottom row had two values of 5.67 and two of 5.47 out of six possibilities. The mean scores of 5.47 both had the same SD of 3.46. The first row had two means of 0.70 and two of 1.00, of six possibilities. The data just would seem to be "too neat". Real data are randomly messy. On a more positive note, Milevsky (2020) reported an average age of 50.55 in a sample of 40 adults; 40 x 50.55 = 2022.00, indicating that his data did pass a GRIM test. With respect to percentages, suppose an article reported a value of 21% in a situation where N = 18 (e.g., Durwood et al., 2017). Twenty-two percent would only occur when 4 cases met the criterion (4/18 = 22%). Either there had to be several missing cases or the value of 21% could not be correct (in this case, the authors informed us that 22% was the correct value).

Limitations. As Brown and Heathers (2017) have admitted, this works best for small samples. One of the main limitations for the GRIM test and others is that test validity declines with larger sample sizes. Suppose you get a result close to an integer with a sample of 1,200. If you don't know how many cases were missing for that particular mean score or standard deviation, it may not be possible to obtain a definitive result with the GRIM test. Even if you know the exact sample size, it may still be difficult to use the GRIM test effectively if the sample is much larger than 200, in our estimation. For an example of this issue, Milevsky (2019) reported an average age of 18.23 for 432 college students, which does not yield an integer (18.23 x 432 = 7875.36); rather, integers from 7,874 to 7,877 would yield an average age of 18.23. Furthermore, if a mean score is reported as an integer (e.g., 4.00), the GRIM test is meaningless as an integer times the integer for the sample size will always yield an integer, even if the mean has been fabricated. With respect to checking the accuracy of percentages, such data cannot be checked unless the sample size (including missing data) is reported as well as the percentage value.

Recommendations. If researchers are fabricating their data, it is unlikely, we think, that they will fabricate it so well that all their data pass the GRIM and SPRITE tests. It might be easy to make mean scores fit a desired pattern but getting the standard deviations to fit as well? The GRIM and SPRITE tests are easy to use and can detect data that may involve impossible values; therefore, we recommend them for use by editors and reviewers. However, those tests do have several limitations, as we have noted.

3.3 Third Clue – Standard Deviations

A third set of clues revolves around inspection of standard deviations, whose value is often overlooked in statistics education (Schumm, Bosch, & Doolittle, 2009). We have already noted that the SPRITE test can check for the validity of a given standard deviation for a given sample size and mean score. Here we will discuss other issues with standard deviations, using the same order of discussion – issue and then examples, but holding limitations, and recommendations to the end of this section. However, some authors may try to hide behind the use of standard errors rather than standard deviations or only report standard deviations/standard errors in supplementary materials. For example, LaCour and Green (2014), who article was retracted by Science in 2015 (McNutt, 2015), reported mean scores and standard errors only in their supplemental material. Standard deviations are found, of course, by multiplying standard errors by the square root of the sample size.

Unusual similarities in standard deviations/standard errors

Our first way of checking standard deviations is keyed to the idea that from random data, one would expect randomness in standard deviations as well as mean scores. If articles report too many identical standard deviations or identical standard errors that might be a clue to problems. Using the same standard deviations for multiple scores might simplify calculating effect sizes (Wilson, n.d.) for an author, but such a result might be very unlikely. If one accepts the last digit of a standard deviation as a random variable, then p = .10 and a binomial test could be used to rest the probability of getting a large number of identical SDs or SEs.

We found one case (Gartrell, Bos, and Koh, 2018) in which two matched samples (N = 77 each) were being compared on 21 outcomes. Surprisingly, despite many theoretical and methodological differences between the two samples, over 85% of the standard errors, across the two samples, were identical to two decimal points. The odds of identical matches for 18 or more of 21 cases would be very small using the binomial test ($z = 11.20$, $p < .000001$), even though the journal's editor didn't find those results to be problematic. In a retracted article, in their Table 1 (Ruggiero, Steele, & Marx, 2000, p. 1275), out of 36

values, there were 6 standard deviations in the narrow range of 0.90 to 1.10, which would seem like an unusual situation with another 7 between 0.50 and 1.50. Among the remaining scores, 10 featured standard deviations between 3.30 and 3.70. Extending the upper range to 3.81 would add another two cases. Again, the data would seem to be too clean, as real data are randomly messy.

Identical pairs of means and standard deviations

As a second check, not only can means or standard deviations be remarkably similar within reported tables, *pairs* of means and standard deviations can also be reported as *identical*. As shown by Hartgerink and Wicherts (2016, p. 6), one of Stapel's articles (Ruys & Stapel, 2008) featured 15 of 32 cells of means and standard deviations featured identical pairs; Hartgerink and Wicherts commented that "Finding exact duplicates is extremely rare for even one case if the variables are a result of probabilistic processes as in sampling theory" (p. 6). We would suspect that having both means and standards being identical would be an even greater indication of possible fraud than having either means or standard deviations being unusually similar.

Ranges and standard deviations.

Our third check involves both theoretical and actual ranges of scores as compared to standard deviations. One check is to see if the standard deviations make sense with respect to the *theoretical* range of the scores. If the variable has a theoretical range of 1 to 7 and the mean and the standard deviation are both 4, it would seem that all of the data should occur within plus or minus one standard deviation, an unlikely situation; authors should be asked to provide frequency data in such situations. As an example, LaCour and Green reported a standard error of .15 with a sample of 519 and a mean of 3.17 for a variable with a theoretical range of 1 to 5. The square root of 519 is 22.78 which multiplied by 0.15 = 3.42. Thus plus or minus one standard deviation would extend beyond the possible range of the variable on both sides of the mean, clearly an unlikely circumstance. For another variable, LaCour and Green reported a mean of 61.50, with a standard error of 2.42 with n = 519. The standard deviation of 55.13 was large enough to encompass all of the data on the high end of the variable and most of it on the low end, again an unlikely situation.

A second, related check is based on the fact that the ratio between the *actual* range of the variable and the standard deviation usually ranges between 3 and 8 (Schumm, Higgins, et al., 2017). Values outside those ranges are less likely and should be queried. In the case of LaCour and Green, 4 (the range from a maximum of 5 – 1, the minimum value) divided by 3.42 = 1.17, a very unlikely ratio of the range to the standard deviation. For the second variable we used before, the ratio (100/55.13) was found to be 1.81, again well below typical values.

Standard deviations for binary variables.

A fourth check for standard deviations involves binary variables. Although we mentioned this issue earlier in 2018 (Schumm et al., 2018), Heathers and Brown (2019) reported the creation of a formal DEBIT test to detect impossible values for binary data. Recently, we have published examples of how published articles had reported dozens of impossible standard deviations for binary data, leading eventually to the retraction of at least five articles in sociological journals (Schumm, Crawford, & Lockett, 2019a, b; also see Pickett, 2020). In general, the curve depicting the relationship between binary means and SDs is a semicircle, with points (0,0) and (1,0) at the low and high ends of the mean score and a maximum point (for the SD) of about 0.50 for a mean of 0.50, for larger samples. Thus, once you know the mean for a binary variable, then there is only *one* mathematical possibility for the standard deviation because the SD can be predicted exactly from the mean and the sample size. For larger samples (n > 100), the maximum SD for a binary variable should be about .50, when the mean is between .40 and .60. For very small samples, the maximum SD might be larger (e.g., .70). Elsewhere, we have presented a chart that shows the maximum SD for binary variables for given sample sizes (Schumm, Crawford, & Lockett, 2019b). For examples, readers may consult (Pickett, 2020; Schumm, Crawford, & Lockett, 2019a, b).

Limitations. The main limitation to this assessment of standard deviations is that some scholars can sidestep the issue by simply not reporting standard deviations for binary or other variables. One way this is done is to report percentages for demographic variables rather than mean scores, simply omitting standard deviations. It will be more difficult to check for range/SD ratios if only SEs are reported. It

does take some calculation to determine such ratios. Some authors may not report both actual and theoretical ranges. Without actual ranges, the range/SD ratio cannot be checked. If editors or peer reviewers were not familiar with the association between means and SDs for binary variables, it would be easy to miss these problems. Sometimes, errors with standard deviations might only be typographical; nevertheless, they should be corrected before publication.

Recommendations. If data are being fabricated in research, it may be more difficult to "fix" standard deviations than to report mean scores that seem reasonable. Therefore, we recommend that editors and peer reviewers be alert unusual patterns of standard deviations/standard errors as well as for unusual patterns for mean scores. They should also look for impossible or unusual values of binary means and standard deviations and, at the very least, insist that both means and SDs for binary variables be reported, even when the means are reported as percentages. Some authors may report theoretical ranges but not actual ranges, but they should provide both to editors and peer reviewers. Having large numbers of identical pairs may also indicate that a test of the randomness of terminal digits might be fruitful (Mosimann, Dahlberg, Davidian,, & Krueger, 2002; Mosimann, Wiseman, & Edelman, 1995) for detecting anomalies. If editors or peer reviewers detect unusual patterns of standard deviations, the authors should be queried and asked to provide printouts of or data files for their results and/or of their raw data. In the case of sample descriptions, authors should not provide merely the number of research participants, but additional information on age, education, income (with means, standard deviations, and ranges) to permit ratio checks on the demographic variables as well as the study's outcome variables. Simonsohn (2013) has provided more formal, detailed methods for calculating the probabilities of finding unusual patterns of standard deviations.

3.4 Fourth Clue - Reconstruction of Data.

Issue. Another check is whether the data appear to make sense and perhaps can be reconstructed. Here one might start with reconstructing the mean score for the N cases and then try to obtain the reported SD by adjusting the mean score by increasing/decreasing data points by similar amounts to retain the same mean.

Examples. Ruggiero et al. (2000) reported a mean score of 4.00 with a standard deviation of 3.81 for a sample of 29 cases for a variable ranging from 0 to 10. The sample size is small enough that a likely set of values can be reconstructed. The mean must be 4.00, of course. But creating data such that the standard deviation is nearly 4 as well is more of a challenge. First of all, such a bifurcated set of data would seem unlikely. Second, even if the data were not fabricated, the distribution would be far from normally distributed (Kolmogorov-Smirnov test, $p < .001$), an assumption usually required for the analysis of variance used in their analyses. Thus, even if their data were valid, their parametric analysis was not. Another example in the same article was a mean of 9.02, SD = 0.96, n = 30. First, we could not obtain a mean of 9.02 but only 9.03. One set of data that yielded a mean of 9.03 and SD = 0.96 featured 1 value of 7, 10 8's, 6 9's, and 13 10's. Again, the data were bifurcated; trying to spread out the data to reduce the bifurcation increased the standard deviation; as before, such data would be significantly ($p < .001$) non-normal and violate an assumption of the analysis of variance. A third example was a mean of 0.70 with a standard deviation of 0.70, n = 30. This result can be reproduced with 13 0's, 13 1's, and 4 2's, but such data would also be non-normal ($p < .001$). In the three examples, actual ranges were 8, 3, and 2. Dividing those ranges by the standard deviations yields ratios of 2.0, 3.13, and 2.86, which are all on the low side of what would be expected to such ratios.

Limitations. The main limitation of reconstructing data is that it may be impossible for large samples. Even for small samples, reconstruction may be difficult and time consuming. Although the SPRITE test will do this, it may find more than one way, even dozens of ways, to reconstruct the data to obtain the desired mean and SD. If the mean scores are impossible (GRIM test), attempts at data reconstruction will be a waste of time.

Recommendations. Reconstruction of data may be in the "too hard" box for editors and reviewers but does provide another approach to checking the validity of data. It may only be useful in exceptional situations.

3.5 Fifth Clue – Recalculation of Significance Tests and Effect Sizes.

Issue. A fifth check involves recalculation of significance testing. A number of tests are available on the web that allow for determining t or F values with appropriate degrees of freedom and significance levels if one can input the sample size, mean, and SD for each of two or more groups. Web tests for t values and F values can also reveal unlikely results. Another check can involve estimation of effect sizes. Rough estimates of effect sizes can be made, comparing the differences between the two means divided by a rough average of the two standard deviations while precise estimates can be made from website calculators (Wilson, no date).

Examples. For example, Ruggiero et al. (2000, p. 1275) reported an $F(1, 138) = 17.71$ ($p < .001$) from comparing one group of scores (M = 2.17, $SD = 1.78$) versus another (M = 1.59, $SD = 2.00$). When we used an internet calculator for t-tests to recalculate the significance level, assuming equal groups of 70 participants (based on the degrees of freedom in the F test, although the two samples should have had 90 each), we found $t(178) = 1.81$ ($p < .08$, Cohen's d (1992a, b) = 0.31). Our recalculation did not find the difference to be statistically significant.

Ruggiero et al. (2000, p. 1275) also cited their comparison of one score (M = 4.14, $SD = 2.67$) versus another (M = 1.42, $SD = 1.04$) yielding $F(1, 155) = 99.77$ ($p < .001$). Those values would yield an approximate effect size of 2.72/1.86 = 1.46, a value nearly twice as large as what Cohen (1992) deemed a "large" effect. That effect size would seem to be "too good to be true". In another paper, Ruggiero & Marx (1999, Table 1, p. 778; 2001) reported four significant effects ($p < .001$) but did not report effect sizes. Our calculations found the effect sizes to run between 1.14 and 1.41, again "too good to be true".

Limitations. Even if data are fabricated, one might expect the statistical analyses to be correct, given the data. One would hope that effect sizes would be calculated and reported, even if the data were fabricated. However, even statistical results can be incorrect, so one cannot assume their validity without checking. If authors do not report means or standard deviations, recalculation of their statistical tests may not be possible. Even with the availability of web-based t-tests, it can be burdensome to repeat multiple mean/SD comparisons to check an author's results. Although effect sizes can be estimated easily, exact results may require another website visit and multiple inputs.

Recommendations. We recommend that all authors be required to provide all the necessary data and should be asked to respond when statistical results don't seem to fit the data or are much better than what might be expected. Authors should discuss how their results compared in terms of effect sizes with prior literature.

3.6 Sixth Clue - Incorrect Statistics/Incorrect Solutions

Issue. Sometimes reported results just don't make sense. Perhaps two groups are being compared but while each group has 50 cases, the F-test's degrees of freedom are 1, 200, an impossible value. The square root of an F value (for two groups) is equivalent to a t value, so if the F is 100, a t value of 5 would look like a problem. F tests have to have two degrees of freedom, so if an F test is reported with only one number for the two sets of degrees of freedom that suggests a problem.

Examples. What else was also wrong with Ruggiero et al.'s (2000) results? Ruggiero et al. (2000) compared those two data points (means of 4.14 and 1.42) with an F-test, reporting a finding of $F(1, 155) = 99.77$ ($p < .001$). We calculated a t-test from the two sets of scores, obtaining $t(55) = 5.17$ ($p < .0001$). However, for a test of two groups, the t test is the square root of the F test. However, the square root of $F = 99.77$ is 9.99, not 5.17. Furthermore, the degrees of freedom for the F-test should have been 1, 55 rather than 1, 155. Thus, we see, for Ruggerio's data, that the mean scores were invalid, the standard deviations were not possible, nonparametric statistical tests should have been used but were not, the F values were incorrect, and the degrees of freedom for the F-test were incorrect. In other articles we have found F tests with only one degree of freedom [F(2195)] and means that didn't make sense with standard deviations (Schumm, Crawford, and Fawver, et al., 2019). For another example if a mean of 59 is reported with SD = 10 but only 5% of the subjects scored above 63, you have to wonder what is going on. Usually, about 34% of data would fall between 59 and 69, but only 5% is falling above 63? It turned out that the 59 was an error; the correct mean score was 49, not 59.

Limitations. It may be difficult for journal editors and reviewers to believe that some scholars would report non-sensical statistical results, but that is a real issue. Once a scholar starts to cheat with their

data, one error may lead to another, in a cascade of problems. It may be sad that editors and reviewers have to be on the alert for such cascades, but they do.

Recommendations. Sample sizes, degrees of freedom, statistical tests, and levels of significance, as well as effect sizes, are interrelated. If results do not reflect that interrelationship, authors should be challenged on their results, so that corrections can be made before publication. The "statcheck" package in R can be used to evaluate inconsistencies between p values and test statistics (Brown, Kaiser, & Allison, 2018).

3.7 Seventh Clue – Significant Results that Are Declared Not to be Significant

Issue. The seventh clue may not suggest fraud in terms of fabrication of data, but may suggest data falsification or misrepresentation. There are many clever ways of proving or disproving a null hypothesis. If an author is known to have a pre-existing bias in one direction or the other (yes, sometimes they do have such a bias!), reviewers and editors should be on the alert for methods that would favor obtaining the desired outcome. Briefly, if one wants to find in favor of the null hypothesis, use small samples, use two-sided tests, use conservative post hoc tests, or ignore significant findings by omission (Schumm, 2021). If you want to reject the null hypothesis, use large samples, one-sided tests, use more liberal post hoc tests, and report trivial findings that are nonetheless significant statistically.

For example, Gartrell et al. (2018) concluded that "there were no significant differences in measures of mental health" between the children since birth of lesbian mothers and a comparison sample from a national study. However, in their Table 1, they reported significant differences for anxiety/depression (p = .01) and for internalizing (p = .02), with (our calculation) effect sizes of 0.46 and 0.39 (0.34 if you accept the data from Koh et al. (2019) as more accurate). The reason is that they used a conservative (i.e., Bonferroni) post hoc procedure to control for multiple tests. Even so, small-to-medium effect sizes just don't "go away" in small samples with low statistical power. In Schumm and Crawford (2020) we showed how some scholars "found" non-significant results comparing groups of children by ignoring the significant difference between the groups overall (they split them into three subgroups but didn't report the overall differences), dividing alpha by three to reduce it to 0.016, which meant that two results that were significant at p = .02 were deemed non-significant.

3.8 Eighth Clue – Inconsistencies Across Studies Using the Same Data

Issue. Some years ago, we observed that studies based on the same data had reported different time periods for data collection and had reported different sample sizes (Schumm, Nazarinia, & Bosch, 2009). Sometimes authors will report inconsistent results across different journal articles, perhaps hoping no one will detect the discrepancies.

Examples. Gartrell, Bos, and Koh (2018) reported means of 13.67 and 9.46 for internalizing and externalizing scales, respectively; but for 77 cases, those means failed the GRIM test and, thus, also the SPRITE test. In a later study, Koh, Bos, and Gartrell (2019) reported the scores on the same variables for male (N = 39) and female (N = 38) children and all four means passed the GRIM test, but when the four scores were combined for the total of 77 children, the means/SDs came to 13.21/8.64 (internalizing) and 8.79/6.33 (externalizing) rather than the scores of 13.67/8.60 and 9.46/7.37 reported in the 2018 report. The combined scores passed the GRIM/SPRITE tests, indicating that any problems with the scores were limited to the 2018 report. At www.statstodo.com/CombineMeansSDs_pgm.php one can find a calculator to combine subgroup scores (means/SDs/N's) to find total group results, which also was useful for calculating total self-worth scores as discussed in Schumm and Crawford (2020).

Limitations. While reviewers are provided with the paper they are to assess, they probably won't know about other papers that they could cross-check. This will be a difficult problem to solve because the initial papers won't have the later papers for cross-checking available. Post-publication review may be the best option available.

Recommendations. Editors may require authors to discuss how their submitted paper fits in with other already published reports or other papers in progress. One way to deal with this is to ask if authors have considered breaking down their results by gender (if they haven't) or other obvious categories. Some authors will publish one study from a data set and then proceed to do the same analysis with gender, race, sexual orientation, or other categories. In a first paper, at least editors should

require authors to indicate they plan future analyses with the same data in those other ways. This could alert other scholars for upcoming results that might be checked for consistency with the initial report(s).

3.9 Ninth Clue – Bias in Terminal Digits

Issue. Terminal digits in means or standard deviations/standard errors should be nearly random (Mosimann et al., 1995). Thus, another concern is when authors fabricate data but favor certain last digits over others. Unlike the case discussed where terminal digits were the same, here terminal digits are not the same but some are much more prevalent than others, when they should be nearly random and equally distributed from 0 to 9. This, too, can be tested with a chi-square test for the occurrences of all of the numbers or by checking for the chances of so few zeros (or any other digit) if p = .10.
Example. Someone might "hate" zeros in the last digit of standard deviations and use the other nine numbers instead as Pickett (2020) has demonstrated with respect to one author's research.
Limitations. However, the major limitation is that even if you prove that the outcomes are very unlikely, it's difficult to prove the data were fabricated. Unlike the previous clues, this clue will work as well, if not better, for large samples because the chances of having disparities in the percentages of terminal digits will be reduced with large sample sizes. However, testing for this clue probably involves the most effort as the terminal digits have to be summarized and then analyzed, which could take a great deal of time if the research involves hundreds of statistical coefficients.
Recommendations. This clue is probably best reserved for a follow-up test if one suspects that data are manufactured. It can be recommended for larger data sets as well as smaller samples.

Overall Limitations

One limitation here is that if the author's biases and those of reviewers and editors match, it may be easier to overlook ways of twisting data to promote either the null hypothesis or its rejection. All sorts of excuses can be made for things like small samples or not reporting all available results.

It is incumbent, we think, for editors and peer reviewers to take their own biases into account when evaluating statistical results and to be wary of the various ways results can be twisted or selectively applied in order to "find in favor of" the desired outcomes.

4 Discussion

We have presented several informal and formal methods for detecting "fishy" results or data. Some "catches" may be only simple, innocent mistakes or even typographical errors introduced by the printers. In other cases, a find of multiple "fishy" results may be an indication of fraudulent data. We doubt that all fabrications of data can be detected with these nine methods, especially if the reports do not provide much information about their statistical results.

We recognize the chance that authors of integrity might occasionally have their data questioned, but if so, they should be glad to provide requested information. Some clever deceitful authors may navigate their way through our suggested "checks" for fabricated data or results, but certainly some may be detected, saving themselves (as well as reviewers and editors) the embarrassment of having their papers later retracted. We hope that widespread awareness of these checks on fabrication of data may deter its occurrence in the future (deterrence theory). We welcome future advances by other scholars to improve upon the checks discussed here.

While some controversial articles can draw "heat" for unpopular conclusions (e.g., Regnerus, 2012), we have found a number of articles that featured "fishy" results that were not detected during peer review. This situation leads us to question the soundness of much of what happens during peer review. One journal retracted a paper recently because one reviewer had recommended to the editor that the paper be given a good statistical review, but that review didn't happen. In such a case, we'd blame the editor more than the reviewers since the editor didn't ask for a review from a statistical expert. But editors need to be able to detect some of these issues before they send papers out for review, rejecting some manuscripts prior to peer review, if they feature substantial anomalies.

5 Conclusion

We have discussed ways of detecting data anomalies in small samples, with a focus on mean scores and standard deviations that don't make sense. While our approaches to detecting data anomalies and/or scientific fraud are not as complex as some, we think that fraud will be deterred more if a larger number of scientists and the public understand the basics of fraud detection compared to more precise knowledge being limited to a very few scientists with elite methods. That is to say, if a scholar is considering cutting corners and committing scientific misconduct, the risk of getting caught probably comes into consideration (deterrence theory); that risk will be higher, if a larger number of scholars, or even journalists, are able to detect possible misconduct. Regardless of sample size, manuscripts that omit important statistical data should be viewed with caution, since such omissions sidestep the chance of others detecting many possible data anomalies. Authors and co-authors should be familiar enough with the raw data that they can detect data anomalies rather than leaving most of the data analysis and reporting to one author, simply assuming everything is fine; when one co-author can report results with serious deficiencies without other co-authors detecting the problems, there is a problem with authorship assignment (Penders & Shaw, 2020). Both reviewers and editors of journals should keep these issues in mind when considering manuscripts for publication and do their best to intercept papers with substantial numbers of anomalies or papers with substantial amounts of omitted statistics before they are accepted for publication (Lanier, 2020). Once papers are published, they should still be examined for inconsistencies or anomalies, such as those we have illustrated. When fabrication of data is suspected, both the accused and the whistleblowers deserve due process and protection (Bouter & Hendrix, 2017; Malek, 2010; Willcox, 1992). Early detection of scientists who may be making a career of research misconduct is important lest their pattern be permitted to continue unhindered for years or even decades (Mistry, Grey, & Bolland, 2019). Teachers may want to ask students to take examples from this report or more recent ones and test them against the nine clues.

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. The findings in 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, November 20, 2019, by the third author.

Disclosure statement. No potential conflict of interest was reported by the authors.

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

ORCID. Walter R. Schumm http://orcid.org/0000-0003-3097-3551

References

1. Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board Task Force Report. *American Psychologist*, *73*(1), 3-25.
2. Brainard, J., You, J., & Bonazzi, D. (2018). Rethinking retractions. *Science*, *362*(6413), 390-393. Brown, A. W., Kaiser, K. A., & Allison, D. B. (2018). Issues with data and analyses: Errors, underlying themes, and potential solutions. *PNAS*, *115*(11), 2563-2570.
3. Brown, N. J. L., & Heathers, J. A. J. (2017). The GRIM test: A simple technique detects Numerous anomalies in the reporting of results in psychology. *Social Psychological and Personality Research*, *8*(4), 363-369.
4. Brown, N. J. L., & Heathers, J. A. J. (2019). 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, 360 Huntington Avenue, Boston, MA 02115.

5. Bouter, L. M., & Hendrix, S. (2017). Both whistleblowers and the scientists they accuse are vulnerable and deserve protection. *Accountability in Research*, *24*(6), 359-366.
6. Cohen, J. (1992a). A power primer. *Psychological Bulletin*, *112*, 155-159.
7. Cohen, J. (1992b). Statistical power analysis. *Current Directions in Psychological Science*, *1*, 98-101.
8. Crocker, J., & Cooper, M. L. (2011). Addressing scientific fraud. *Science*, *334(6060)*, 1182. doi: 10.1126/science.1216775.
9. Fanelli, Daniele. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PloS One*, *4*, e5738.
10. Fanelli, D. (2013). Why growing retractions are (mostly) a good sign. *PLoS Medicine*, *10*(12), e1001563. doi: 10.1371/journal.pmed.1001563.
11. Gartrell, N., Bos, H., & Koh, A. (2018). National Longitudinal Lesbian Family Study – Mental health of adult offspring. New *England Journal of Medicine*, *379*(3), 297-299.
12. Hartgerink, C. H. J., Voelkel, J. G., Wicherts, J. M., van Assen, M. A. L. M. (2019). *Detection of data fabrication using statistical tools*. Unpublished report. Tilburg, Netherlands: Tilburg University.
13. Hartgerink, C. H. J., & Wicherts, J. M. (2016). Research practices and assessment of research misconduct. *ScienceOpenResearch*. Doi: 10.14293/S2199-1006.1.SOR-SOCSCI.ARYSBI,v1.
14. Hartgerink, C. H. J., Wicherts, J. M., & van Assen, M. A. L. M. (2016). The value of statistical tools to detect data fabrication. *Research Ideas and Outcomes*, *2*, e8860, 1-17.
15. Heathers, J. A. J., Anaya, J., van der Zee, T., & Brown, N. J. L. (2018). Recovering data from summary statistics: Sample parameter reconstruction via iterative techniques (SPRITE). *Peerj Preprints*. doi.org/10.7287/peerj.preprints.26968v1.
16. Heathers, J. A. J., & Brown N. J. L. (2019) DEBIT: A simple consistency test for binary data. Unpublished paper, Bouve College of Health Sciences, Northeastern University, 360 Huntington Avenue, Boston, MA 02115.
17. Koh, A., Bos, H. M. W., & Gartrell, N. K. (2019). Predictors of mental health in emerging adult offspring of lesbian-parent families. *Journal of Lesbian Studies*, *23*(2), 257-278.
18. Kuroki, T., & Ukawa, A. (2018). Repeating probability of authors with retracted scientific publications. *Accountability in Research*, *25*(4), 212-219.
19. LaCour, M. J., & Green, D. P. (2014). When contact changes minds: An experiment on transmission of support for gay equality. *Science*, *346*(6215), 1366-1369. doi: 10.1126/science.1256151.
20. Lanier, W. L. (2020). Dealing with inappropriate-, low-quality-, and other forms of challenging peer review, including hostile referees and inflammatory or confusing critiques: Prevention and treatment. *Accountability in Research*, online advance.
21. Malek, J. (2010). To tell or not to tell? The ethical dilemma of the would-be whistleblower. *Accountability in Research*, *17*, 115-129.
22. Marcus, A., Oransky, I. (2018). www.sciencemag.org/news/2018/02/meet-data-thugs-out-exposes-hoddy-and-questionable-research.com.
23. McNutt, M. (2015a). Editorial expression of concern. *Science*, *348*(6239), 1100. Doi: 10.1126/Science.aac.6184.
24. McNutt, M. (2015b). Editorial retraction. Science, 348(6239), 1100. doi: 10.1126/science.aac6638.
25. Milevsky, A. (2019). Parental factors, psychological well-being, and sibling dynamics: A meditational model in emerging adulthood. *Marriage & Family Review*, *55*(5), 476-492.
26. Milevsky, A. (2020). Sibling dynamics in adulthood: A qualitative analysis. *Marriage & Family Review*, *56*(2), 91-108.
27. Mistry, V., Grey, A., & Bolland, M. J. (2019). Publication rates after the first retraction for biomedical researchers with multiple retracted publications. *Accountability in Research*, *26*(5), 277-287.
28. Mosimann, J. E., Dahlberg, J. E., Davidian, N. M., & Krueger, J. W. (2002). Terminal digits and the examination of questioned data. *Accountability in Research*, *9*, 75-92.
29. Mosimann, J. E., Wiseman, C. V., & Edelman, R. E. (1995). Data fabrication: Can people generate random digits? *Accountability in Research*, *4*, 31-55.
30. Nurunnabi, M., & Hossain, M. A. (2019). Data falsification and question on academic integrity. *Accountability in Research*, *26*(2), 108-122.
31. Penders, B., & Shaw, D. M. (2020). Civil disobedience in scientific authorship: Resistance and insubordination in science. *Accountability in Research*, online advance.
32. Pickett, J. T. (2020). The Stewart retractions: A quantitative and qualitative analysis. *Econ Watch Journal*, *17*(1), 152-190.

33. Pickett, J. T., & Roche, S. P. (2018). Questionable, objectionable or criminal? Public opinion on data fraud and selective reporting in science. *Science and Engineering Ethics*, *24*, 151-171.
34. Reisig, M. D., Holtfreter, K., & Bersofsky, M. E. (2020). Assessing the perceived prevalence of research fraud among faculty at research-intensive universities in the USA. *Accountability in Research*, advance online.
35. Regnerus, M. (2012). How different are the adult children of parents who have same-sex relationships? Findings from the New Family Structures Study. *Social Science Research*, *41*, 752-770.
36. Ruggiero, K. M., & Marx, D. M. (1999). Less pain and more to gain: Why high-status group members blame their failure on discrimination. *Journal of Personality and Social Psychology*, *77*, 774-784. doi: 10.1037/0022-3514.77.4.774.
37. Ruggiero, K. M., & Marx, D. M. (2001). Less pain and more to gain: Why high-status group members blame their failure on discrimination: Retraction. *Journal of Personality and Social Psychology*, *81*, 178. doi: 10.1037/0022-3514.81.2.178.
38. Ruggiero, K. M., Steele, J., Hwang, A., & Marx, D. M. (2000). "Why did I get a 'D'?" The effects of social comparisons on women's attributions to discrimination. *Personality and Social Psychology Bulletin, 26*, 1271-1283. doi: 0.1177/0146167200262008.
39. Ruggiero, K. M., Steele, J., Hwang, A., & Marx, D. M. (2001). "Why did I get a 'D'?" The effects of social comparisons on women's attributions to discrimination: retraction. *Personality and Social Psychology Bulletin, 27, 1237.*
40. Ruys, K. L., & Stapel, D. A. (2008). Emotion elicitor or emotion messenger?: Subliminal priming reveals two faces of facial expressions [retracted]. *Psychological Science*, *19*(6), 593-600.
41. Schumm, W. R. (2021). Confirmation bias and methodology in social science: an editorial. *Marriage & Family Review*, 57(4), 285-293. Doi: 10.1080/01494929.2021.1872859.
42. Schumm, W. R., Bosch, K. R., & Doolittle, A. (2009). Explaining the importance of statistical variance for undergraduate students. *Psychology and Education – An Interdisciplinary Journal*, *46*(3/4), 1-7.
43. Schumm, W. R., & Crawford, D. W. (2020). Is research on transgender children what it seems? Comments on recent research on transgender children with high levels of parental support. *Linacre Quarterly*, *87*(1), 9-24.
44. Schumm, W. R., Crawford, D. W., Fawver, M. M., Gray, N. K., Niess, Z. M., & Wagner, A. D. (2019). Statistical errors in major journals: Two case studies used in a basic statistics class to assess understanding of applied statistics. *Psychology and Education – An Interdisciplinary Journal*, *56, 1/2*, 35-42.
45. Schumm, W. R., Crawford, D. W., Higgins, M., Lockett, L., AlRashed, A., & Ateeq, A. B. (2018). Estimating the standard deviation from the range: A replication of analysis of demographic data reported in *Marriage & Family Review*, 2016-2017. *Marriage & Family Review*, *54*, 777-792.
46. Schumm, W. R., Crawford, D. W., & Lockett, L. (2019a). Using statistics from binary variables to detect data anomalies, even possibly fraudulent research. *Psychology Research and Applications*, *1*(4), 112-118.
47. Schumm, W. R., Crawford, D. W., & Lockett, L. (2019b). Patterns of means and standard deviations with binary variables: A key to detecting fraudulent research. *Biomedical Journal of Scientific and Technical Research*, *23*(1), 17151-17153.
48. Schumm, W. R., Higgins, M., Lockett, L., Huang, S., Abdullah, N., Asiri, A., Clark, K., & McClish, K. (2017). Does dividing the range by four provide an accurate estimate of a standard deviation in family science research: A teaching editorial. *Marriage & Family Review*, *53*, 1-23. doi: 10.1080/0149.
49. Schumm, W. R., Nazarinia, R. R., & Bosch, K. R. (2009). Unanswered questions and ethical issues concerning U.S. biodefence research. *Journal of Medical Ethics*, 35, 594-598.
50. Simonsohn, U. (2013). Just post it: The lesson from two cases of fabricated data detected by statistics alone. *Psychological Science*, *24*(10), 1875-1888. doi: 10.1177/0956797613480366.
51. Stapel, D. A., & Lindenberg, S. (2011a). Coping with chaos: How disordered contexts promote stereotyping and discrimination. *Science*, *332*, 251. doi: 10.1126/science.1201068.
52. Stapel, D. A., & Lindenberg, S. (2011b). Retraction. *Science*, *334*, 1202.
53. Steen, R. G., Casadevall, A., & Fang, F. C. (2013). Why has the number of scientific retractions increased? *PLoS One*, *8*(7), e68397. doi: 10.1371/journal.pone.0068397.
54. Stern, A. M., Casadevall, A., Steen, R. G., & Fang, F. C. (2014). Financial costs and personal consequences of research misconduct resulting in retracted publications. *eLife*, *3*, e02956. doi: 10.7554/eLife.02956.
55. Stewart, E. A., Simons, R. L., & Conger R. D. (2000). The effects of delinquency and legal sanctions on parenting behaviors. *Families, Crime, and Criminal Justice*, *2*, 257-279.
56. Stroebe, W., Postmes, T., & Spears, R. (2012). Scientific misconduct and the myth of self-correction in science. *Perspectives on Psychological Science*, 7, 670-688. Doi: 10.1177/1745691612460687.

57. Vogel, G. (2011). Psychologist accused of fraud on 'astonishing scale'. *Science*, *334*, 579.
58. Wiedermann, C. J. (2018). Inaction over retractions of identified fraudulent publications: Ongoing weakness in the system of scientific self-correction. *Accountability in Research*, *25*(4), 239-253.
59. Willcox, B. L. (1992). Fraud in scientific research: The prosecutor's approach. *Accountability in Research*, *2*(2), 139-151.
60. Wilson, D. B. (n.d.) *Practical meta-analysis effect size calculator* [online calculator]. https://campbellcollaboration.org/effect-size-calculator/research-resources/research-for-resources/effect-size-calculator.html.

Appendix

Table 1. Summary of statistical methods for detecting research anomalies

Method	Examples
1. Absence of Basic Statistical Information	Stewart, Simons, & Conger, 2000
2. Means or Percentages that Don't Make Sense	Ruggiero, Steele, Hwang, & Marx (2000)
3. Standard Deviations – General	
Unusual similarities in standard deviations	Gartrell, Bos, & Koh, 2018; Ruggiero et al. (2000); LaCour & Green, 2014
Unusual similarities in pairs of means and standard deviations	Ruys & Stapel, 2008
Ratio between actual range and standard deviation	LaCour & Green, 2014
Binary variables	Regnerus, 2012; Schumm, Crawford, & Lockett, 2019a, b
4. Reconstructing data	Ruggiero et al. (2000)
5. Recalculating statistical tests and effect sizes	Ruggiero et al. (2000)
6. Incorrect statistics	Ruggiero et al. (2000); Schumm, Crawford, et al. (2019)
7. Significant results declared not significant	Gartrell et al. (2018); Koh et al. (2019)
8. Inconsistencies in means/standard deviations across different studies using the same data	Gartrell et al. (2018); Koh et al. (2019)
9. Improbable Percentages for Terminal Digits	Pickett (2020)

