This draft: 07/29/2012

Just post it: The lesson from two cases of fabricated data detected by statistics alone.

Uri Simonsohn
The Wharton School
University of Pennsylvania
uws@wharton.upenn.edu

Abstract.
I argue that journals should require authors to post the raw data supporting their published results. I illustrate some of the benefits of doing so by describing two cases of fraud I identified exclusively through statistical analysis of reported means and standard deviations. Analyses of the raw data provided important confirmation of the initial suspicions, ruling out benign explanations (e.g., reporting errors; unusual distributions), identifying additional signs of fabrication, and also ruling out one of the suspected fraudster's explanations for his anomalous results. If we want to reduce fraud, we need to require authors to post their raw data.

Data and SAS code behind all results in this paper will be posted upon the paper being accepted for publication.

Supplemental materials are available here.

Academic misconduct is a rare event, but not rare enough. Its occurrence challenges the credibility of the findings in our journals, and our mission as scientists more generally. While prevention is important, some misconduct is likely to occur no matter what steps are taken to prevent it. Measures that facilitate identifying such cases can help mitigate their negative consequences. The risk of detection, furthermore, may constitute the ultimate deterrent.

To undetectably fabricate data is difficult. It requires both (i) a good understanding of the phenomenon being studied (what do measures of this construct tend to look like?, which variables do they correlate with and by how much?, how often and what type of outliers are observed?, etc.), and (ii) a good understanding of how sampling error is expected to influence the data (e.g., how much variation and of what kind should estimates of this construct exhibit for the observed sample size and design?).

This paper makes two main points. First, information already included in most psychology publications, e.g., means and standard deviations, can be analyzed in light of points (i) and (ii) above, to identify likely cases of fraud. Second, the availability of raw data is an invaluable complement to verify the presence, and identify the precise nature of, fabrication. The paper illustrates these points through the analyses used to uncover two cases of fraud in psychology.

Both cases were first identified exclusively through the statistical analysis of means and standard deviations reported in published papers, by examining whether estimates across independent samples were too similar to each other to have originated in random samples (see also Carlisle, 2012; Fisher, 1936; Gaffan & Gaffan, 1992; Roberts, 1987; Sternberg & Roberts, 2006).

Supplementary analyses conducted on the raw data behind some of those means and standard deviations provided invaluable confirmation of the likely corrupt nature of the data. The analyses triggered committees that investigated possible academic misconduct at two different institutions, and were followed by the resignation of the researchers in charge of the suspected studies.

If we want to reduce fraud, the path forward is simple: journals should require authors to post the raw data behind published results. There are obvious challenges to implementation. Sometimes datasets are proprietary, sometimes variables could identify individual participants, sometimes datasets will be used in future investigations by the same authors, etc. These are exceptions. The majority of datasets could probably be posted without negative repercussions, while fraud deterrence and detection is just one of the positive ones (see e.g., Wicherts & Bakker, 2012).

Journals, it follows, should make the posting of raw data the default policy to which authors abide, unless they are waived from such requirement by the editorial team, when presented with reasonable exempting circumstances. This is not a utopian proposal. Journals publishing research similar to that of most psychology journals have already implemented both the required default policy and the reasonable exception clause (e.g., *Judgment and Decision Making*, and the *American Economic Review*).

It is the status quo that rests on utopian premises: journals have no protection against fraud, researchers have tremendous incentives to commit it, and yet we all conveniently assume it does not happen. As the examples that follow painfully demonstrate, this assumption is inconsistent with evidence published in some of our most respected journals.

Introduction to simulations (a card-dealing analogy)

Simulations are used to answer questions such as "how likely is X to happen?" in situations where math-based answers are not known or easy to apply. Suppose, for example, we had a deck of cards and wanted to know the odds of drawing 4 cards: two black cards, two red cards, and the four cards add up to 17. Computing the exact answer is possible, but suppose we did not know how. The simulation approach is to take a deck of cards (or a computer program that behaves like one), deal four cards many times, and use the percentage of times we dealt the described pattern as the probability estimate. Simulations turn probability questions we cannot easily answer into simple tasks involving but two steps: set up the problem, count.

When raw data were not available, for my simulations I drew cards from normal distributions with means and standard deviations matching those in the analyzed study. I rounded and bounded the randomly drawn numbers to better match the variables of interest (e.g., when simulating counts, the drawn number was rounded to 0 decimals and negative draws were censored at 0). I then also rounded the resulting means and standard deviations to the level of precision in the simulated studies (see Rubin & Stigler, 1979).

When raw data were available I did not rely on the normal distribution. Instead, I created a new deck, one where each observation in the study becomes a card. This type of simulation, bootstrapping, captures any idiosyncrasies of the observed data such as skewness, outliers, etc. (Boos, 2003; Efron & Tibshirani, 1994).

Case Study 1: Hot sauce and standard deviations.

The first anomalous findings.

Sanna and colleagues (2011) predicted, based on the metaphorical relationship between morality and altitude (e.g., *higher* moral ground), that people randomly assigned to be in higher elevation would act more prosocially.

For example, in experiment 3 participants walked up to the stage of a theatre, down to its orchestra pit, or stayed at the baseline elevation in a control condition, and then poured hot sauce for a supposed fellow participant to consume in a taste test; the prediction was that those on the stage would be more kind and serve less hot sauce.

The paper included three experiments summarized in their Table 1, reprinted here as Figure 1. It reveals a troubling anomaly: while means differ dramatically across conditions the SDs are almost identical. Consider Study 3. Between the High and Low conditions, means differ by about 115%, the SDs by just 2%. I next estimate whether such extreme similarity of SDs is compatible with data having been collected from random samples.

Fig.1

Simulating hot sauce.

To quantify how similar SDs were within a study, I computed *their* standard deviation: SD(SDs). For example, in Study 3, SD(25.09, 24.58, 25.65)=.54. The goal of the simulations is assessing the probability of a study leading to SD(SDs)≤.54. I simulated samples drawing from normal distributions with μ's matching the reported sample means, and with σ equal the pooled SD:

High (n=15):	N(39.74, 25.11)
Low (n=15):	N(85.74, 25.11)
Control (n=15):	N(65.73, 25.11)

Using the same σ in all three distributions is extremely conservative. We are trying to assess if the SDs are too similar in the samples, and we are going to be comparing them to simulations drawn from populations with an *identical* σ; true σs cannot be any more similar than identical, but they could easily be less similar.

Simulating 100,000 Study 3s, I found that only 1.3% of them had SD(SDs)≤.54. We conclude that the SDs in this study are too similar to have originated from random samples (p=.013). Proceeding analogously with the other two studies the individual p-values for each of them originating in random samples are p=.053 and p=.026.

While suggestive, these results at the individual study level are not powerful enough for concerns as serious as academic misconduct. It is hence useful to consider a more powerful test and assess the likelihood of the paper as a whole arising from random samples. I averaged the SD(SDs) across the three studies asking how likely is it for three studies in a single paper to arrive at such similar SDs. Before aggregating, however, we must get around the problem that studies differ in scale (e.g., grams of sauce and number of fish) and sample size (n=15 vs. n=20).

An easy way to do this is to divide SD(SDs) by the standard error of the (pooled) standard deviation, leading to an intuitive measure of deviation: by how many standard errors do SDs differ in this study? Let's refer to this number as Ψ. For the hot sauce

study we had that SD(SDs) was .54, dividing by the standard error (4.58) we get to Ψ_3=11.7%.[1] That is, SDs differ by 11.7% of a standard error from each other in Study 3.

For studies 2 and 4, Ψ_2=23.8% and Ψ_4=16.7%. Taking a simple average of Ψ_2,Ψ_3,Ψ_4 we get $\Psi_{overall}$=17.4%. Out of 100,000 simulated papers only 15 had $\Psi \leq 17.4\%$, so the data for this paper as a whole are inconsistent with random sampling (p=.00015), see Figure 2.

Fig.2*

Other authors, same paradigms, no anomalies.

One concern is that dependent variables used in this paper may have some unusual property that leads to SDs being more similar than those from simulations based on normal distributions. The fact that the three studies involve such different dependent variables (time in minutes, grams of hot sauce and number of fish in a computer game) already alleviates this concern somewhat. To address it further, I collected SDs from papers by other authors using similar paradigms. Figure 3 shows than they obtained dramatically more varied SDs.

Fig.3

Same authors, other papers, same anomaly.

Another concern regarding the simulations is their post-hoc nature. I performed them *because* the SDs struck me as too similar. While the low p-value, p=.00015,

[1] The standard error of the standard deviation, assuming normality, is SE(SD)=SD/√2n (Yule, 1923). For the hot sauce study, for example, we had SD=25.11 and n=15 so SE(SD)=25.11/√30=4.58. Note that the SE is used merely for scaling so while only approximate, the results do not hinge on possible deviations from such approximation.

provides some protection, the concern is important enough to warrant addressing more explicitly through replications.

When looking for articles for Figure 3, I searched for papers citing the same articles cited by Sanna et al., 2011. Among the papers I found, three were also authored by Sanna and colleagues, two of which report SDs. One included a single three-condition experiment (Sanna et al.,2009), providing limited statistical power to detect anomalous results. Nevertheless, the SDs were improbably similar to each other Ψ=23%, p=.056. The other included three relevant experiments with between 6 and 9 conditions each (Sanna et al.,2003). Its overall Ψ of 16.8% is virtually impossible to obtain from random samples, p<1 in 58 billion.

Convenience-sample of papers.

For one final comparison I obtained the SDs from the first few papers in the, at the time, most recent issue of the *Journal of Experimental Social Psychology* (September, 2011). To maximize comparability I computed Ψs based on sets of 2-3 conditions rather than entire papers or experiments, leading to 17 estimates for studies reported by Sanna and colleagues, and 39 by the control set. All by Sanna et al. had $\Psi \leq 25\%$, but only four in the control set did. This difference is statistically significant by a proportions test, p<1 in 5 billion. Whatever is behind the excessive similarity of standard deviations in the papers by Sanna et al., it is not something that appears to be present in work by any other authors.

Analyses based on raw data

Upon sharing all these analyses with the authors I requested and received the raw data behind the embodiment of altitude paper. This allowed a series of additional analyses that further suggest the data did not originate in random sampling.

Simulations redone. Upon verifying that there were no reporting errors, ruling out the most benign of explanations, I reran all simulations, now drawing cards from decks representing raw data rather than normal distributions. For each experiment I created one card per observation and placed them all in a single deck (e.g., one deck with 45 cards for Study 3). [2] I then would draw a card, take a note of its value, return it to the deck, reshuffle, and draw another card, until an entire simulated sample was created. I would then do the same for the other two samples. Data simulated this way, despite originating in the raw data, also almost never led SDs as similar as those reported in the paper, rejecting again the null of random sampling (p=.00011).[3]

Not only SDs are too similar. Finally, the raw data revealed that also the range of values the dependent variable takes was too similar. These differences of Max-Min across each of three conditions in the three studies were (10,10, 10), (76, 76, 74) and (28, 26,28). Maxima and minima are extremely volatile statistics when obtained from random samples.

Case 2. Colored folders and similar means

The first anomalous finding.

[2] To simulate from the raw data under the null of equal variances one first must subtract the condition's mean from each observation (for more details see Boos & Brownie, 1989)

[3] These simulations include both #-of-fish and mood from Study 4. Without the latter, p=.00083

Smeesters and a junior colleague (2011) predicted, based on the notion that the color red leads to avoidance and the color blue to approach, that priming participants with such colors could switch on and off contrast and assimilation effects. Their paper reports a single 3x4 between-subject experiment (N=169). Instructions were given in a red, blue or white folder, inviting participants to write about one of four possible targets: Kate Moss (a fashion model), Albert Einstein, a model, a professor. This task was then followed by 20 multiple-choice general knowledge questions. The number of correct answers was the dependent variable.

The authors expected performance to be high in six of the conditions, and low in the remaining six. The results are printed in Table 1. It shows data consistent with the predictions, but also a troubling anomaly: means predicted to be similar are exceedingly so. [4]

Table 1

Simulating the colored folders experiment

To quantify how similar means predicted to be similar were, I computed their standard deviation, SD(Ms). For the high conditions $SD(Ms)_{high}$=SD(11.43,11.71,…,12.07)=.23. For the lows $SD(Ms)_{low}$=.24. As before, we can divide by the standard error to obtain a scale free metric. Averaging across the high and low conditions we arrive at Ψ=30.8%; means predicted to be similar are 30.8% of a standard error apart from one another.

I then performed simulations to assess how likely $\Psi \leq 30.8\%$ would be if the data originated in random sampling. I drew samples of the reported sizes from normal

[4] Some of these were obtained from Dr. Smeesters.

distributions with μ equal to the pooled mean of the six predicted-to-be-similar conditions, and σ equal to each sample's SD. For example, the average number of correct answers in the *high* conditions was 11.81 and the SD of one of them was 2.79. I simulated this condition drawing from N(11.81, 2.79).

Note again that assuming μ is the same across conditions is extremely conservative. Even if the authors' hypothesis were correct there is no reason to expect, for example, that the combination of a red folder with a Kate Moss prime leads to the exact same average performance as that of an Einstein prime with a white folder. Only 21 of 100,000 simulations had $\Psi \leq 30.8\%$, see Figure 4.

Fig.4

Same authors, other papers, same anomaly.

Concerned that the previous analysis was at least somewhat post-hoc, I examined mean similarity *because* means seemed too similar, I analyzed two other articles by Dr. Smeesters. I selected them because they had been completed recently and had multiple conditions predicted to be similar.[5]

Both papers contained multiple studies, but each study had few conditions predicted to be similar, so I examined similarity of means across studies. This makes the assumption that the true means are the same even more conservative. If the true means are different, as they almost certainly are, the observed results are even less likely.

For each paper, (i) I identified a dependent variable employed across all studies, (ii) selected all conditions where such variable was predicted to show no effect/be at baseline, (iii) analyzed the degree of similarity for those means, using simulations

[5] Liu, Smeesters, and Trampe (2012); Smeesters, Wheeler, and Kay (2009)

analogous to those presented above. The null of random sampling was rejected for both papers, p=.00011 and p=.0023 (see supplement).

Analysis of raw data for the colored folders study

I requested and promptly received the raw data for the folders study, allowing a series of additional analyses that further suggested the data did not originate in random sampling.

Simulations redone. Upon verifying that there were no reporting errors, I reran all the simulations now drawing cards from the raw data instead of normal distributions. I did two versions of this. In both I created two decks of cards, one with the 86 observations from the *low* conditions, and one with the 83 from the *high* condition, and generated *high* and *low* samples by drawing from the respective decks.

In one version, "with replacement," I would as before draw one card, make a note of it, return it, reshuffle, draw another, until all conditions had been simulated. In the other version, "without replacement," I would shuffle the cards and deal them all at once into the six corresponding conditions, keeping therefore the exact set of observations constant, and the only thing varying is which *high* condition gets which *high* card, and analogously for the *lows*. The estimated p-value for the null of random sampling are .00030 and .00018 for the with/without replacement respectively, comparable to the .00021 from assuming normality.

Lack of sampling error of a different kind. One of the most well-known judgment-biases is the "belief in the law-of-small-numbers" (Tversky & Kahneman, 1971). Where even small samples are believed to closely resemble underlying populations. For

example, when people are asked to imagine four coin tosses, they tend not to imagine all heads or all tails, imagining instead something closer to the underlying 50:50 expectation. While four identical coin tosses is an unlikely event (12.5%), it is more likely than the nearly 0% observed in people's "fake" coin tosses.

If a person believing in the law of small numbers were asked to generate scores for the 12 conditions of general knowledge questions, we might analogously expect her to avoid too many of the same scores in a given condition, to generate sets of scores that are excessively even. To examine this prediction I needed a metric of how evenly distributed the data were. I focused on one of the simplest possible: the frequency of the mode.

For example, the fourteen scores for one of the twelve conditions were: [6,7,7,8,8,9,9,10,10,10,12,12,14,15]. The mode here is 10 and it appears three times. Across the twelve conditions nine had the mode appearing 3 times, and three just 2 times. The sum of mode frequencies, **F**, is hence **F**=9*3+3*2= 33.

How unlikely is it for **F**≤33? It occurs just 21 times in the 100,000 bootstraps with replacement, see Figure 5, and 93 in those without. This test of the data having originated in random sampling, then, also rejects such null. It is interesting to establish how independent the excessive similarity of means (Ψ) and the excessive evenness of scores (**F**) are. Are we observing the same red flag twice or are we seeing two red flags? For each of 100,000 simulations we have a Ψ_i and a $\mathbf{F}_i$. The correlation of these two metrics, it turns out, is quite low, $r(\mathbf{F},\Psi)$=.16 when drawing cards with replacement, and $r(\mathbf{F},\Psi)$=.18 when without. So closer to two flags. Not a single simulation has both $\mathbf{F}_i \leq 33$ *and* $\Psi_i \leq 30.8\%$, random sampling would seem to never lead to the observed data.

Finally, the official report of the investigation of possible academic misconduct by Smeesters notes that a student of his conducted a conceptual replication of this study. It failed, and the pattern of excessively evenly distributed scores was not obtained (Zwaan et al., 2012, p. 27).

Non-explanations for the anomalies.

When interviewed by the Erasmus committee examining possible misconduct on his part (Zwaan, et al., 2012), Smeesters said he "possibly made a coding mistake in two questions" (pp.4) and that it is possible that "people who answered a difficult question correctly always answered another difficult question correctly." (pp.3). While both of these things may be true, neither could account for the excessive similarity of means nor of evenness of scores across conditions. If anything, these features of the data generating process would cause the opposite pattern: more rather than less variation.

First, using the wrong key to grade questions increases rather than decreases noise. In the extreme, if the entire key was wrong, then the set of scores being analyzed would be only noise. We are trying to explain the opposite, the lack of statistical noise.

Second, if the likelihood of a person getting one question right were correlated with that of her getting another question right, then we would see more rather than fewer people with same scores within conditions. In the extreme, if those knowing one answer knew all of them, we would only see people scoring 20s and 0s. We again are trying to explain the opposite, that there are too few people with the same scores per condition, not too many.

Finally, and just as importantly, neither sloppy coding nor correlated answers to the questions can possibly account for simulations that show the data are incompatible with random samples, because these simulations are drawing from those raw data. The simulations already take into account *all* such idiosyncrasies. If six people got higher scores because of sloppy grading, for example, then there will be six cards with higher scores because of sloppy grading, and we will be drawing from them when estimating how likely the observed result is. Similarly, if everyone getting question 3 right also got question 6 right, all cards for a person with question 3 right will also have question 6 right. Etc.

Fig.5

Analysis of raw data for a willingness-to-pay study.

One of the two papers analyzed earlier to replicate the similarity of means analysis, [6] contains studies where participants were asked to indicate their maximum willingness-to-pay (WTP) for each of two black t-shirts with very similar designs, see Figure 6. I obtained the raw data for one of them (Study 3) and compared it to data from several other papers eliciting WTP: two published in journals that post data (Fudenberg, Levine, & Maniadis, 2012; Newman & Mochon, 2012), four by colleagues who at some point had emailed me their WTP data (Frederick, 2012; Keren & Willemsen, 2009; Simonson & Drolet, 2004; Yang, Vosgerau, & Loewenstein, 2012), a previous paper of mine (Simonsohn, 2009), and a new study I run just for this analysis, where participants indicated their WTP for *the same* two t-shirts from Figure 6. Smeesters' data markedly

[6] Liu, et al. (2012)

stand out from all of these benchmarks in a variety of ways. In what follows I discuss three examples.

Fig.6

Multiples of $5. A striking pattern in the WTP data of the Smeesters study is the low frequency of valuations expressed as multiples of $5. Because people often round up or down when answering pricing questions, the percentage of such valuations is typically much higher than the rate expected if people answered randomly (20%).

Figure 7 shows this is the case for all benchmark valuations, including two also run in the Netherlands and employing Euros as the currency, and the three excluding participants based on the instruction-manipulation-check (see upcoming "another non-explanation" section), but not in Smeesters' study where the rate of multiples of $5 is at the level expected when numbers are chosen randomly. That paper, recall, also exhibits excessive similarity of means, an orthogonal anomaly. The upward trend shows that multiples of $5 are more common among pricier valuations

Fig.7

Correlation of valuations. Recall that every participant indicated their WTP for both t-shirts in short succession. Differences across participants in income, liking of t-shirts, attentiveness, etc. should lead these WTPs to be correlated. The benchmark papers show strong correlations between even completely unrelated items that are valued back to back by the same respondents. Using Cronbach's α to summarize overall correlations: WTP for air purifier, DVD box, chocolate bar and candles have α=.48 (Frederick, 2012), toaster, phone, backpack and headphones α=.62 (Simonson & Drolet,

2004), and planner, keyboard, calculator, book, chocolates and computer mouse α=.52 (Fudenberg, et al., 2012).

In contrast, the correlation between the valuation of the two nearly identical t-shirts in the suspicious paper is *negative*, $r = -.67$.[7] In the replication study the correlation for those same t-shirts was, as is to be expected, positive and high, $r=.80$. The difference between these latter two correlations is highly significant, $Z=13.82$, p is effectively 0.

Correlation of $5. Combining the ideas behind the previous two analyses we can examine if the use of multiples of $5 is correlated within subject. There are again many reasons to expect this would be the case in a sample from real valuations (e.g., respondents may differ in the tendency to use round numbers, or uncertainty of the valuations of t-shirts.).

We see that indeed such tendencies are correlated in the benchmark studies α=.64, .58, and .57 respectively. The correlation in the suspected study, however, is $r=-.04$. In the replication with the same t-shirts $r=.62$. The difference between these two is again highly significant, $Z=5.52$, $p<1$ in 58 million.

Another non-explanation

Smeesters also told the Erasmus committee that dropping participants failing an instruction-manipulation-check, henceforth IMC (Oppenheimer, Meyvis, & Davidenko, 2009) may explain the excessive similarity of means in his data. This is nonsense.

[7] The correlation is also negative within the 4 conditions where there is neither a predicted nor observed difference in valuations of both t-shirts ($r = -.64$).

Excluding noisy answers lowers standard deviations, making means to seem more *different* from each other (because their difference is benchmarked against the standard deviation). This is related to the impact of trimming means on standard errors (Keselman et al., 2004; Yuen, 1974). Intuitively, when we eliminate noisy observations we increase statistical power, making differences of means become more rather than less detectable.

Even if a researcher were to delete the highest few observations in conditions predicted to be low, and the lowest few in conditions predicted to be high, the result would be that means predicted to be high would become more *different* from each other when benchmarked against their standard deviations, same for the lows. Again, then, the explanation Smeesters provided for his anomalous data makes them be, if anything, more anomalous.

I nevertheless empirically assessed the impact of excluding participants based on IMCs by analyzing whether the means across conditions in the original demonstration of this technique (Oppenheimer, et al., 2009), and on a recent paper using it across seven experiments (Yang, et al., 2012) were too similar. They were not (see supplement).

It is worth noting that dropping inattentive participants also cannot explain any of the other irregularities discussed before (e.g., use of non-round numbers, lack of correlation in valuations), and that Smeesters explicitly indicated to the committee he only used this technique in three papers brought to his attention as suspicious. Two of the papers reanalyzed here that show impossible similarity of means are not in that set. Whatever is behind the overwhelming set of irregularities in the data, it has nothing to do with the completely acceptable filtering out of inattentive participants.

Discussion

Two cases studies of fraudulent data detected by statistical analyses alone show that if we desire to prevent and detect fraud, journals should require authors to post the raw data behind the results they report.

Why require it?

While working on this project I solicited data from a number of authors, sometimes due to suspicion, sometimes in the process of creating some of the benchmarks, sometimes due to pure curiosity. Consistent with previous efforts of obtaining raw data (Wicherts, 2011; Wicherts et al., 2006), the modal response was that they were no longer available. Hard disk failures, stolen laptops, ruined files, server meltdowns, corrupted spreadsheets, software incompatibility, sloppy record keeping, etc., all happen sufficiently often, self-reports suggest, that a forced backup by journals seems advisable.

Is raw data really needed?

The two cases were detected by analyzing means and standard deviations, why do we need raw data then? There is a third case, actually, where fraud took place with almost certainty, but due to lack of access to raw data, the suspicions cannot be properly addressed. If raw data were available, additional analyses could vindicate the author, or confirm her/his findings should be ignored. Because journals do not require raw data these analyses will never be conducted.

Furthermore, I've come across a couple of papers where data suggest fabrication but other papers by the same authors do not show the pattern. One possibility is that these

are mere coincidences. Another is that other people, e.g., their research assistants, tampered with the data.

Our research is often conducted by assistants whose honesty is seldom evaluated, and who have minimal reputation concerns. How many of them would we entrust with a bag filled with an uncounted number of $100 bills for a study? Trustworthy evidence is worth much more than $100. The availability of raw data would allow us to detect and prevent also these cases.

Concluding remarks; witch-hunting

From pencils to internet connections, all tools can be used for harm. This does not mean we should not produce or use tools, but it does mean we should take precautions to avoid nefarious uses. The use of statistical analyses for detecting potential fraud is no exception. Few scholarly goals are as important as eradicating fraud from our journals, and yet few actions are as regrettable as publicly accusing an innocent scholar of fraud.

I took a great many measures to pursue the former while preventing the latter. These measures should be easy to emulate, and improve upon, by anyone conducting these types of analyses in the future:(1) Replicate analyses across multiple papers before suspecting foul play by a given author, (2) Compare suspected studies to similar ones by other authors, (3) Extend analyses to raw data, (4) Contact authors privately, transparently, and give them ample time to consider your concerns, (5) Offer to discuss matters with a trusted statistically savvy advisor, (6) Give the authors more time. (7) If after all this suspicions remains, convey them only to entities tasked with investigating such matters, and do so as discretely as possible.

Table 1. Means (SD) for 12 conditions in Smeesters et al. (2011)

Predicted low	9.07 (2.55)	9.43 (2.82)	9.43 (3.06)	9.56[a] (2.83)	9.64 (3.03)	9.78 (2.66)
Predicted high	11.43 (2.79)	11.71 (2.87)	11.77[b] (3.03)	11.85 (2.66)	12.00 (3.37)	12.07 (2.78)

Note: Summary statistics for number of correct answers (out of 20) in a general knowledge task taken by 169 participants assigned to 12 conditions, six conditions were predicted to have high means, the other low. Each condition had *n*=14, except those with superscripts. [a] n=16, [b] n=13.

Figure 1. Reprint of Table 1 in Sanna et al. (2011)

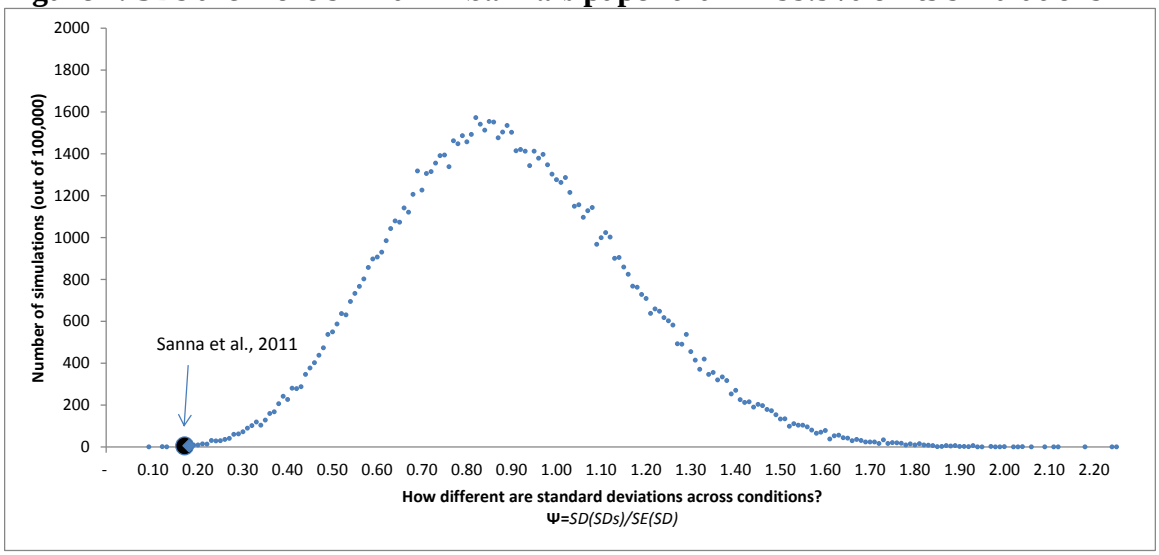
Table 1
Charitable contributions, helping, compassion, and cooperating and moods by physical (vertical) height.

Study/measure	Physical (vertical) height		
	High	Low	Control
Study 1			
Proportion contributing	.16 (59/368)	.07 (26/391)	.11 (37/350)
Study 2			
Mean time helping (minutes)	11.36 (2.82)	6.77 (2.75)	8.74 (2.96)
Study 3			
Mean compassion (hot sauce grams)	39.74 (25.09)	85.74 (24.58)	65.73 (25.65)
Study 4			
Mean cooperating (fish returned)	32.93 (9.24)	20.60 (9.54)	23.66 (9.82)
Mean moods	5.70 (1.13)	5.46 (1.19)	5.59 (1.11)

Note. Proportions rounded to nearest decimal with numbers contributing and totals in parentheses for Study 1; standard deviations in parentheses for Studies 2–4.

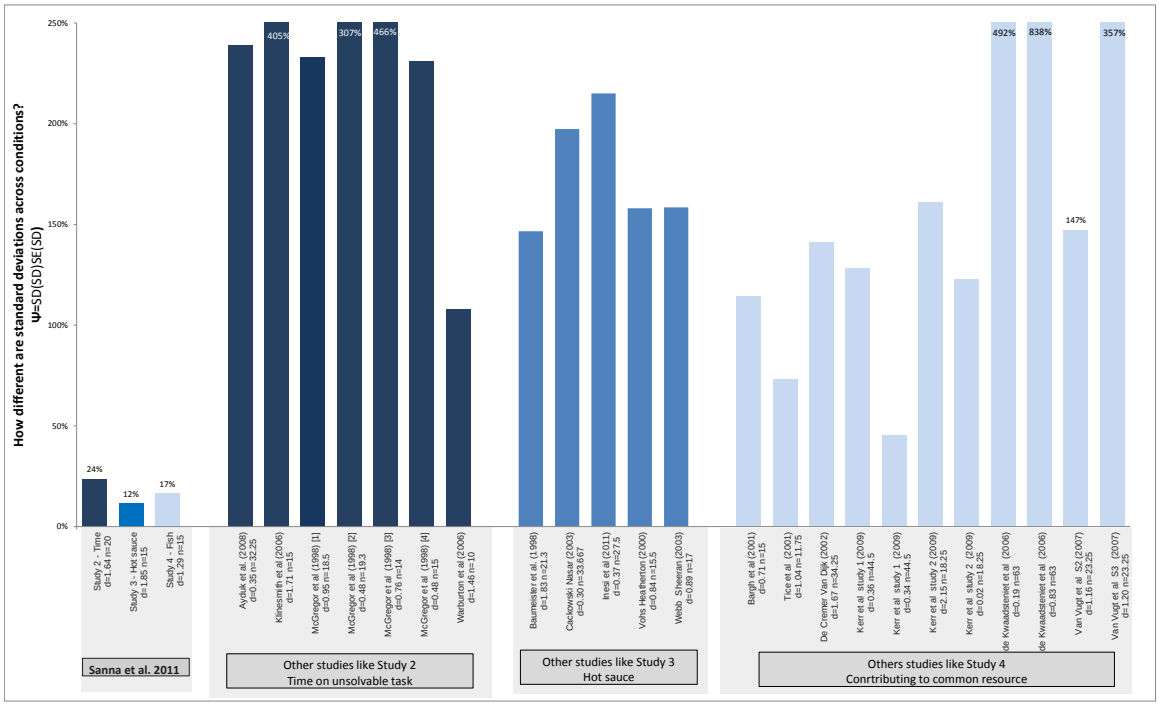
Notes: Rectangles added for emphasis, showing striking similarity of standard deviations across conditions. Study 1 is not an experiment and involved a binary variable. It is not discussed here.

Figure 2. SDs are more similar in Sanna's paper than in 99.9% of its simulations



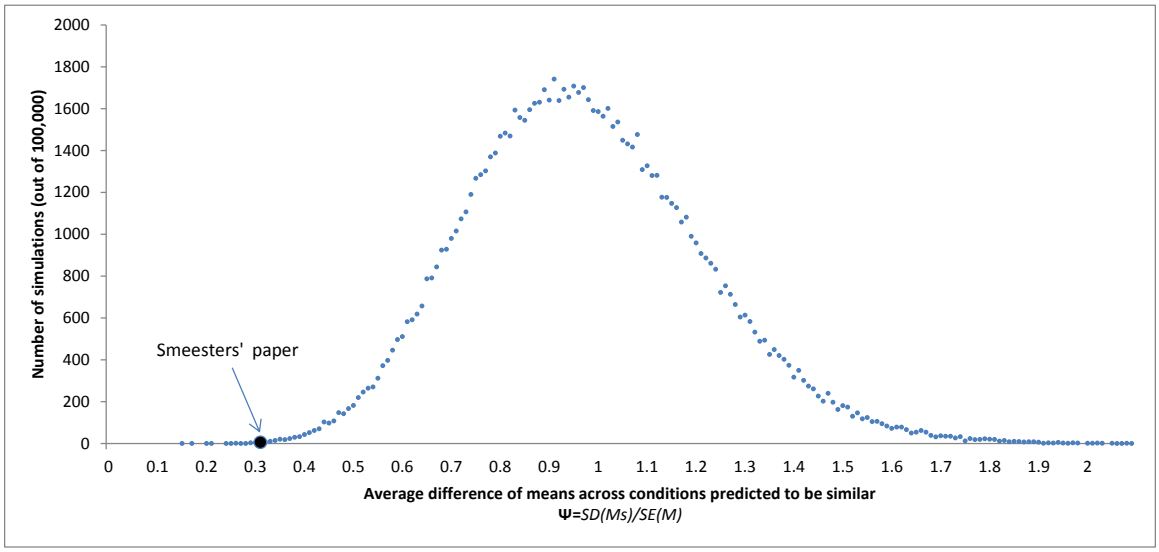
Note: 100,000 simulations where performed of all three experiments in Sanna et al., 2011. Each condition in each experiment is simulated drawing from normal distributions with the sample mean for that condition, and the pooled SD across all conditions. The standard deviations of the standard deviations for each simulated experiment is divided by the standard error and averaged across the three studies to arrive at Ψ. The figure reports the frequency with which Ψs were obtained across the simulations. Only 15 simulations have Ψ≤17.4%, the value from Sanna's paper.

Figure 3. Only Sanna obtains such similar SDs among researchers conducting similar studies.



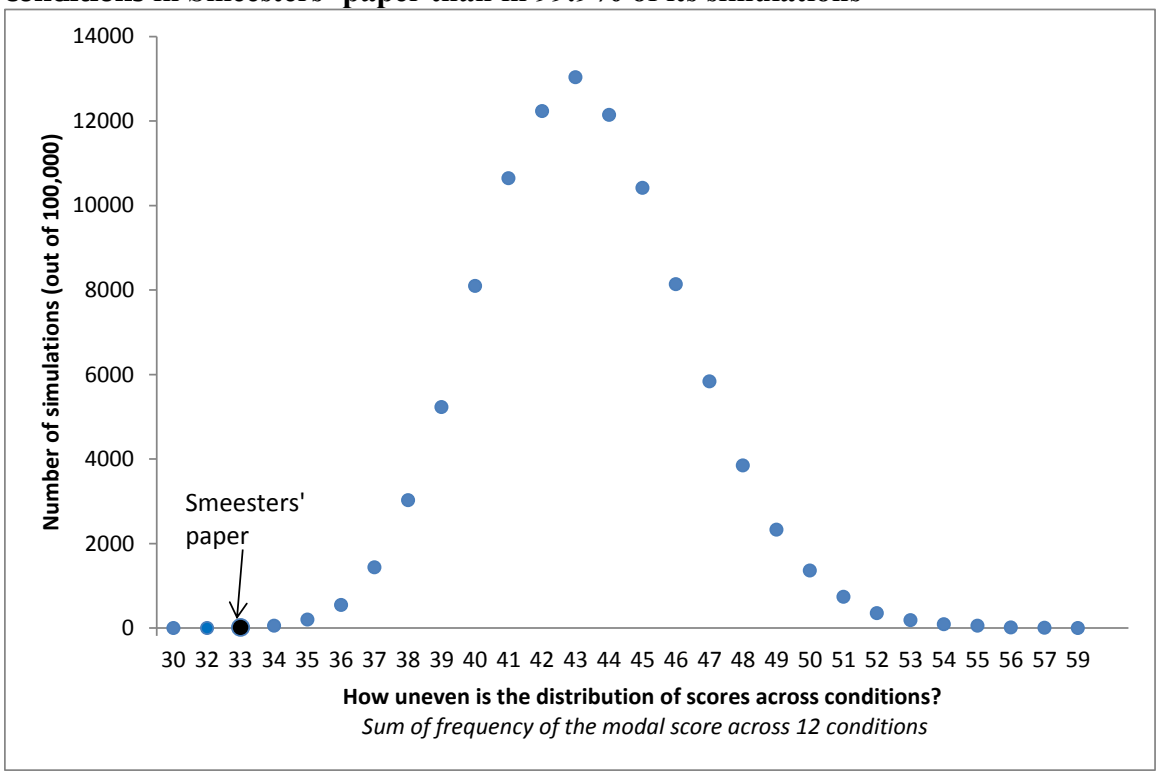
Note: Each bar represents a study's degree of variability of standard deviations across conditions (measured in standard errors of the pooled SD). *n*: per-condition sample size, *d*: effect size (Cohen-d).

Figure 4. Means are more similar in Smeesters' paper than in 99.9% of its simulations



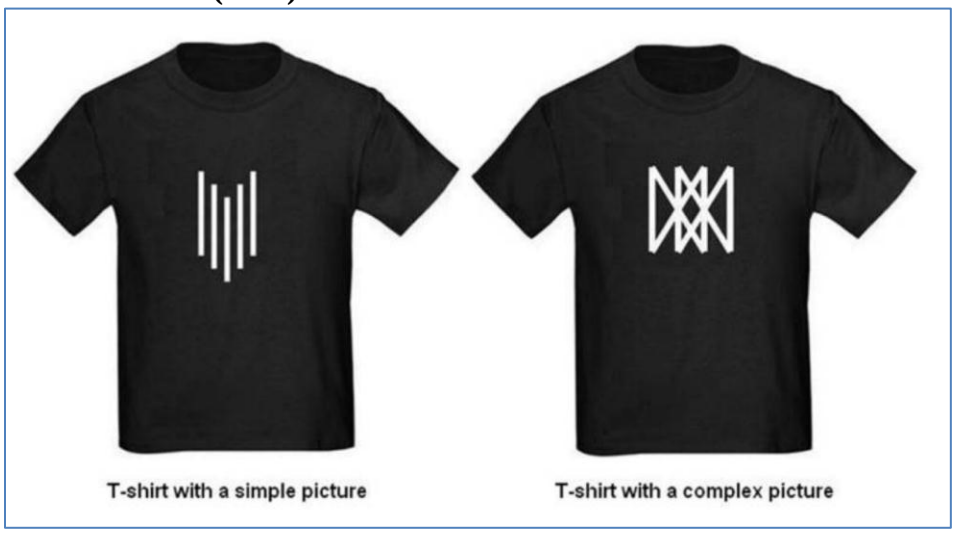
Note: 100,000 simulations where performed of the single study in Smeesters et al., 2011. Each of the six *high* conditions is simulated by drawing from normal distributions with the pooled mean as μ and the sample's SD as σ, analogously for the six *lows*. The analyzed paper reports Ψ=30.8%, means predicted to be similar differed by just 30.8% of a standard error from each other. Only 21 of the 100,000 simulations show such low Ψ.

Figure 5. General knowledge scores are more evenly distributed across 12 conditions in Smeesters' paper than in 99.9% of its simulations



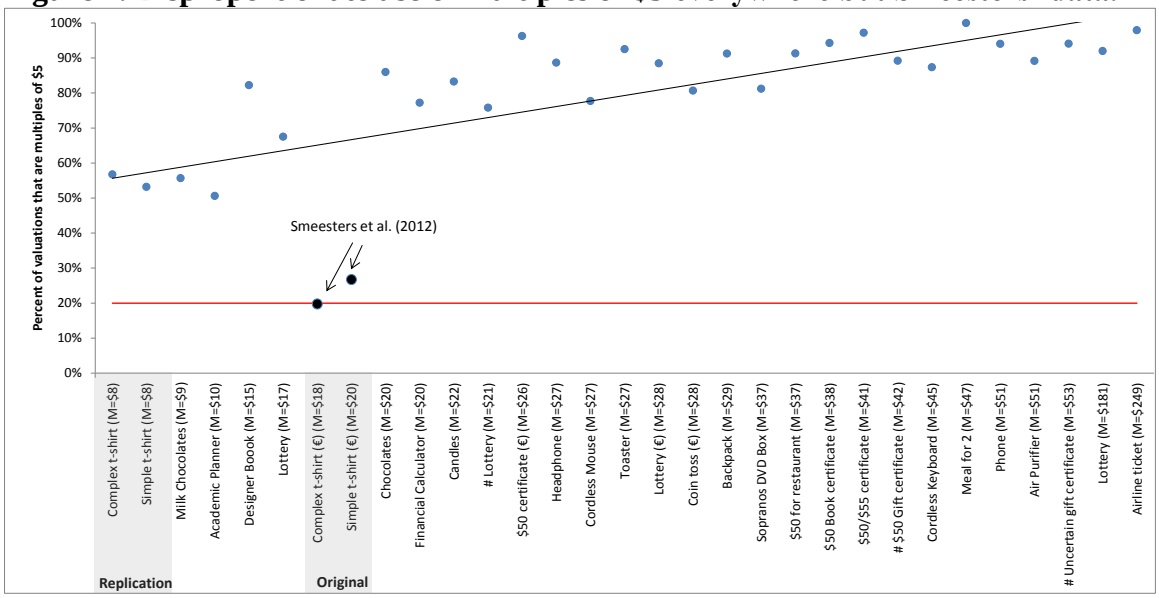
Note: 100,000 simulations where performed of the single study in Smeesters et al., 2011, drawing cards with replacement from the raw data. The x-axis corresponds to the sum of frequencies for the mode of each condition. For example, if the mode appears four times in each of 12 conditions, then such sum is 48, if there are no repeated scores in any conditions, then the sum is 12. In the raw data the sum is 33. Only 21 of 100,000 simulations had such a low sum (or lower).

Figure 6. T-shirts participants were asked their maximum willingness-to-pay for by Smeesters et al (2012).



Note: these were the two items that Smeesters et al (2012) reported asking participants to express their willingness to pay for. In the replication study I used this exact image. In both studies participants saw only one image at a time, in random order.

Figure 7. Disproportionate use of multiples of $5 everywhere but Smeesters' data.



Notes: Each dot indicates the percentage of valuations in a given study that are multiples of $5. Items are sorted by average valuation (indicated with M after item's description). The first two items are from a replication study asking participants to value the same two shirts used in Smeesters' paper. The remaining valuations come from various papers for which raw WTP data were available. The horizontal line represents expectations if valuations were chosen randomly. The three items with an "#" were obtained in a study eliminating 30% of participants failing an Instructional-Manipulation-Check (IMC).

References

Boos, D. D. (2003). Introduction to the Bootstrap World. *Statistical science, 18*(2), 168-174.

Boos, D. D., & Brownie, C. (1989). Bootstrap Methods for Testing Homogeneity of Variances. *Technometrics*, 69-82.

Carlisle, J. (2012). The Analysis of 169 Randomised Controlled Trials to Test Data Integrity. *Anaesthesia*.

Efron, B., & Tibshirani, R. J. (1994). *Chapter 1. An Introduction to the Bootstrap*.

Fisher, R. A. (1936). Has Mendel's Work Been Rediscovered? *Annals of Science, 1*, 115-137.

Frederick, S. (2012). Overestimating Others' Willingness to Pay. *Journal of Consumer Research, 39*(1), 1-21.

Fudenberg, D., Levine, D. K., & Maniadis, Z. (2012). On the Robustness of Anchoring Effects in Wtp and Wta Experiments. *American Economic Journal: Microeconomics, 4*(2), 131-145.

Gaffan, E., & Gaffan, D. (1992). Less-Than-Expected Variability in Evidence for Primacy and Von Restorff Effects in Rats' Nonspatial Memory. *Journal of Experimental Psychology: Animal Behavior Processes, 18*(3), 298-301.

Keren, G., & Willemsen, M. C. (2009). Decision Anomalies, Experimenter Assumptions, and Participants' Comprehension: Revaluating the Uncertainty Effect. *Journal of Behavioral Decision Making, 22*(3), 301-317. doi: 10.1002/bdm.628

Keselman, H., Othman, A. R., Wilcox, R. R., & Fradette, K. (2004). The New and Improved Two-Sample T Test. *Psychological Science, 15*(1), 47-51.

Liu, J., Smeesters, D., & Trampe, D. (2012). Effects of Messiness on Preferences for Simplicity. *Journal of Consumer Research, 39*(1), 199-214.

Newman, G. E., & Mochon, D. (2012). Why Are Lotteries Valued Less? Multiple Tests of a Direct Risk-Aversion Mechanism. *Judgment and Decision Making, 7*(1), 19-24.

Oppenheimer, D., Meyvis, T., & Davidenko, N. (2009). Instructional Manipulation Checks: Detecting Satisficing to Increase Statistical Power. *Journal of Experimental Social Psychology, 45*(4), 867-872.

Roberts, S. (1987). Less-Than-Expected Variability in Evidence for Three Stages in Memory Formation. *Behavioral neuroscience, 101*(1), 120.

Rubin, D. B., & Stigler, S. M. (1979). Dorfman's Data Analysis. *Science, 205*(4412), 1204-1125.

Sanna, L. J., Chang, E. C., Miceli, P. M., & Lundberg, K. B. (2011). Rising up to Higher Virtues: Experiencing Elevated Physical Height Uplifts Prosocial Actions. *Journal of Experimental Social Psychology, 47*, 472-476.

Sanna, L. J., Chang, E. C., Parks, C. D., & Kennedy, L. A. (2009). Construing Collective Concerns. *Psychological Science, 20*(11), 1319.

Sanna, L. J., Parks, C. D., & Chang, E. C. (2003). Mixed-Motive Conflict in Social Dilemmas: Mood as Input to Competitive and Cooperative Goals. *Group Dynamics: Theory, Research, and Practice, 7*(1), 26.

Simonsohn, U. (2009). Direct Risk Aversion: Evidence from Risky Prospects Valued Below Their Worst Outcome. *Psychological Science, 20*(6), 686-692. doi: 10.1111/j.1467-9280.2009.02349.x

Simonson, I., & Drolet, A. (2004). Anchoring Effects on Consumers' Willingness-to-Pay and Willingness-to-Accept. *Journal of Consumer Research, 31*(3), 681-690.

Smeesters, D., & Liu, J. E. (2011). The Effect of Color (Red Versus Blue) on Assimilation Versus Contrast in Prime-to-Behavior Effects. *Journal of Experimental Social Psychology, 47*(3), 653-656.

Smeesters, D., Wheeler, S. C., & Kay, A. C. (2009). The Role of Interpersonal Perceptions in the Prime-to-Behavior Pathway. *Journal of Personality and Social Psychology, 96*(2), 395.

Sternberg, S., & Roberts, S. (2006). Nutritional Supplements and Infection in the Elderly: Why Do the Findings Conflict? *Nutrition journal, 5*(1), 30.

Tversky, A., & Kahneman, D. (1971). Belief in the Law of Small Numbers. *Psychological Bulletin, 76*(2), 105.

Wicherts, J. M. (2011). Psychology Must Learn a Lesson from Fraud Case. *Nature, 480*(7375), 7.

Wicherts, J. M., & Bakker, M. (2012). Publish (Your Data) or (Let the Data) Perish! Why Not Publish Your Data Too? *Intelligence*.

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The Poor Availability of Psychological Research Data for Reanalysis. *American Psychologist, 61*(7), 726.

Yang, Y., Vosgerau, J., & Loewenstein, G. F. (2012). *The Influence of Framing on Willingness to Pay as an Explanation of the Uncertainty Effects*.

Yuen, K. K. (1974). The Two-Sample Trimmed T for Unequal Population Variances. *Biometrika, 61*(1), 165-170.

Yule, G. U. (1923). An Introduction to the Theory of Statistics London, 1922. *Chas. Griffen and Co., Ltd*.

Zwaan, R. A., Groenen, P. J. F., van der Heijden, A. J., & te Lindert, R. (2012). Report by the Committee for Inquiry into Scientific Integrity (English Translation, June 28, 2012).