General Article

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Psychological Science
24(10) 1875–1888

DOI: 10.1177/0956797613480366
pss.sagepub.com
SAGE

Just Post It: The Lesson From Two Cases of Fabricated Data Detected by Statistics Alone

Uri Simonsohn
The Wharton School, University of Pennsylvania

Abstract

I argue that requiring authors to post the raw data supporting their published results has the benefit, among many others, of making fraud much less likely to go undetected. I illustrate this point by describing two cases of suspected fraud I identified exclusively through statistical analysis of reported means and standard deviations. Analyses of the raw data behind these published results provided invaluable 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 fraud's explanations for his anomalous results. If journals, granting agencies, universities, or other entities overseeing research promoted or required data posting, it seems inevitable that fraud would be reduced.

Keywords

judgment, decision making, scientific communication, fake data, data sharing, data posting

Received 7/20/12; Revision accepted 1/30/13

Academic misconduct is a rare event, but not rare enough. Its occurrence challenges the credibility of research, and the mission of science more generally. Although 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. Furthermore, the risk of detection may constitute the ultimate deterrent.

To undetectably fabricate data is difficult. It requires both (a) a good understanding of the phenomenon being studied (e.g., what measures of a construct tend to look like, which variables they correlate with and by how much) and (b) a good understanding of how sampling error is expected to influence the data (e.g., how much variation and the kind of variation the estimates of interest should exhibit given the observed sample size and design). In this article, I show that although means and standard deviations can be analyzed in light of these two criteria to identify likely cases of fraud, the availability of raw data makes the task of detection easier and more diagnostic, and hence that of fabrication more difficult and intimidating.

Posting data has many advantages unrelated to, and possibly more valuable than, prevention and detection of fraud. For example, as Wicherts and Bakker (2012) have noted, when raw data are posted, scientific evidence is preserved for longer periods of time, more researchers get to analyze and hence learn from a given amount of scientific evidence, and reporting errors become easier to prevent and detect.

In this article, I illustrate how raw data can be analyzed for identifying likely fraud through two case studies. Each began with the observation that summary statistics reported in a published article were too similar across conditions to have originated in random samples, an approach to identifying problematic data that has been employed before (Carlisle, 2012; Fisher, 1936; Gaffan & Gaffan, 1992; Kalai, McKay, & Bar-Hillel, 1998; Roberts, 1987; Sternberg & Roberts, 2006).[1] These preliminary analyses of excessive similarity motivated me to contact the authors and request the raw data behind their results. Only when the raw data were analyzed did these suspicions rise to a level of confidence that could trigger

Corresponding Author:
Uri Simonsohn, The Wharton School, University of Pennsylvania, 3730 Walnut St., 500 Huntsman Hall, Philadelphia, PA 19104
E-mail: uws@wharton.upenn.edu

the investigations of possible misconduct that were eventually followed by the resignation of the researchers in question.

There was a third case of exceedingly similar summary statistics, but I was unable to obtain the raw data behind them. The main author reported losing them, and the coauthors of the article did not wish to get involved. The concerns of possible fraud behind this third case remain unaddressed.

The availability of raw data, then, causally led to the retraction of existing publications with invalid data and in all likelihood prevented additional ones from being published. The absence of raw data, in contrast, led to suspicions of fraud in another case not being acted on. If journals, granting agencies, universities, or other entities overseeing research promoted or required data posting, it is hard to imagine that fraud would not be reduced.

There are, at the same time, obvious challenges to data posting. For example, sometimes data sets are proprietary, sometimes variables could identify individual participants, and sometimes data sets will be used in future investigations by the same authors. These are exceptions. A majority, and quite possibly the vast majority, of psychological research is based on data that seem to pose minimal to no challenges for data posting.

Consider, for instance, a recent issue of the *Journal of Personality and Social Psychology* (October 2012). Five of the 10 articles in that issue are based on simple scenario studies, which pose no challenges to data posting. Another 2 used images of participants as stimuli. Because of privacy considerations, these should be unposted, but the actual data, the rows of numbers fed to SPSS, seem to be entirely postable. Another 2 articles employed original and expensive data sets collected by the authors themselves (a multimonth panel and a multicountry survey). The authors arguably deserve the fruit of their labor, and perhaps data posting could be delayed for a grace period of a few years. The 10th article is, fittingly and coincidentally, the retraction of an article in which Smeesters was one of the coauthors.[2] The data behind that article would have been trivial to post.

Why not just continue with the American Psychological Association's policy of making data available on request? While working on this project, I solicited data from a number of authors, sometimes because of suspicion, sometimes in the process of creating benchmarks, and sometimes because of pure curiosity. As has been found in previous efforts of obtaining raw data (Wicherts, 2011; Wicherts, Borsboom, Kats, & Molenaar, 2006), the modal response was that they were no longer available. Harddisk failures, stolen laptops, ruined files, server meltdowns, and so forth all happen sufficiently often, self-reports suggest, that data posting seems advisable. The posting of data could perhaps become the default policy of journals, universities, or granting institutions—a default from which authors might be exempted.

The proposal of data posting for psychology is not a utopian one. Some behavioral-science journals have already implemented both this required default policy and a reasonable-exception clause (e.g., *Judgment and Decision Making* and the *American Economic Review*). Data posting is policy also in fields with vastly larger data sets that are more expensive to collect, and in which competing teams study closely related questions, such as gene sequencing (see e.g., the National Center for Biotechnology Information's Sequence Read Archive, http://www.ncbi.nlm.nih.gov/sra, and Gene Expression Omnibus, http://www.ncbi.nlm.nih.gov/geo/). There are no obvious features of psychological data that make them less postable than genomics data, and there are several that make them more so.

It is the status quo that rests on utopian premises: As a discipline, psychology has no protection against fraud, researchers—especially incompetent ones—have tremendous incentives to commit it, and yet everyone conveniently assumes that it does not happen. As the examples that follow painfully demonstrate, this assumption is inconsistent with evidence published in some of psychology's 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 in which simple math-based techniques are not easy to apply. Suppose, for example, that we have a deck of cards and want to know the odds of drawing two black cards and two red cards with all four cards adding up to 17. Computing the exact answer is possible, but suppose we do 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. Thus, simulations turn probability questions we cannot easily answer into simple tasks involving but two steps: Set up the problem and then 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., for simulated 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 analyzed study.

This approach prevents one from concluding that the results are too similar across samples merely because authors have not reported them with enough decimals for differences to be detected (see Rubin & Stigler, 1979) or because the dependent variable does not have the level of granularity the statistical test assumes. For example, it is less shocking if a publication reports two standard deviations of 1.3 than if it reports two standard deviations of 1.321, and this approach to simulations allows taking this level of reporting precision into account.

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

Case Study 1: Embodiment of Morality and Standard Deviations

The first anomalous findings

On the basis of the metaphorical relationship between morality and altitude (e.g., *higher* moral ground), Sanna, Chang, Miceli, and Lundberg (2011; retraction: Sanna, Chang, Miceli, & Lundberg, 2013) predicted that people randomly assigned to be in higher elevation would act more prosocially. For example, in their Study 3, participants walked up to the stage of a theater, walked down to its orchestra pit, or stayed at the baseline elevation (control condition) and then poured hot sauce for a supposed fellow participant to consume in a taste test; the prediction was that participants on the stage would be more kind and serve less hot sauce.

The article reported three experiments summarized in a table, reprinted here as Figure 1. It reveals a troubling anomaly: Although means differed dramatically across conditions, the standard deviations were almost identical. Consider Study 3. Between the high and low conditions, means differed by about 115%, but the standard deviations differed by just 2%. I next estimated whether such extreme similarity of standard deviations is compatible with data having been collected from random samples.[3]

Simulating the studies

To quantify how similar standard deviations were within a study, I computed the standard deviation of the standard deviations. For example, in Study 3, the standard deviation of the three standard deviations (25.09, 24.58, 25.65) was 0.54. The goal of the simulations was to assess the probability of a study leading to a standard deviation

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.

Fig. 1. Reprint of Table 1 in Sanna's article on the embodiment of morality (see the text). Rectangles are added to emphasize the striking similarity of standard deviations across conditions. Study 1 was not an experiment and involved a binary variable. It is not discussed here.

of the standard deviations that was 0.54 or less. I simulated samples drawing from normal distributions with population means equal to the reported sample means and with σ equal to the pooled standard deviation:

High condition (n = 15): N(39.74,25.11)

Low condition (n = 15): N(85.74,25.11)

Control condition (n = 15): N(65.73,25.11)

Using the same σ in all three distributions was extremely conservative. I was trying to assess whether the standard deviations were too similar in the samples, and I was going to be comparing them with 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 a standard deviation of the standard deviations that was 0.54 or less. The standard deviations in this study are too similar to have originated from random samples, p = .013. Proceeding analogously with the other two studies, I found that their individual p values for originating in random samples were .053 and .026.

Although suggestive, these results at the individual-study level were not compelling enough for concerns as serious as academic misconduct. It was hence useful to consider a more statistically powerful test and consider the joint hypothesis that all three studies combined arose from random samples. I averaged the standard deviation of the standard deviations across the three studies, asking how likely it was for three studies in a single article to arrive at such similar standard deviations. Before aggregating, however, I had to get around the problem that the studies differed in scale (e.g., grams of hot sauce vs. number of fish) and sample size (n = 15 vs. n = 20).

An easy way to do this was to divide the standard deviation of the standard deviations by the standard error of the (pooled) standard deviation. This yielded an intuitive measure of deviation: the number of standard errors by which the standard deviations differed within a given study. I refer to this number as Ψ. For the hot-sauce study, the standard deviation of the standard deviations was 0.54, and dividing by the standard error (4.58) yielded a Ψ of 0.117.[4] That is, standard deviations differed from each other by 0.117 of a standard error in that study.

For Studies 2 and 4, Ψ was 0.238 and 0.167, respectively. The simple average of Ψ_2, Ψ_3, and Ψ_4 was 0.174. Out of 100,000 simulated articles, only 15 had a Ψ of 0.174 or less, so the data for this article as a whole are inconsistent with random sampling, p = .00015 (see Fig. 2).

Other authors, same paradigms, no anomalies

One concern is that the dependent variables in this article may have some unusual property that leads to standard deviations being more similar than those from

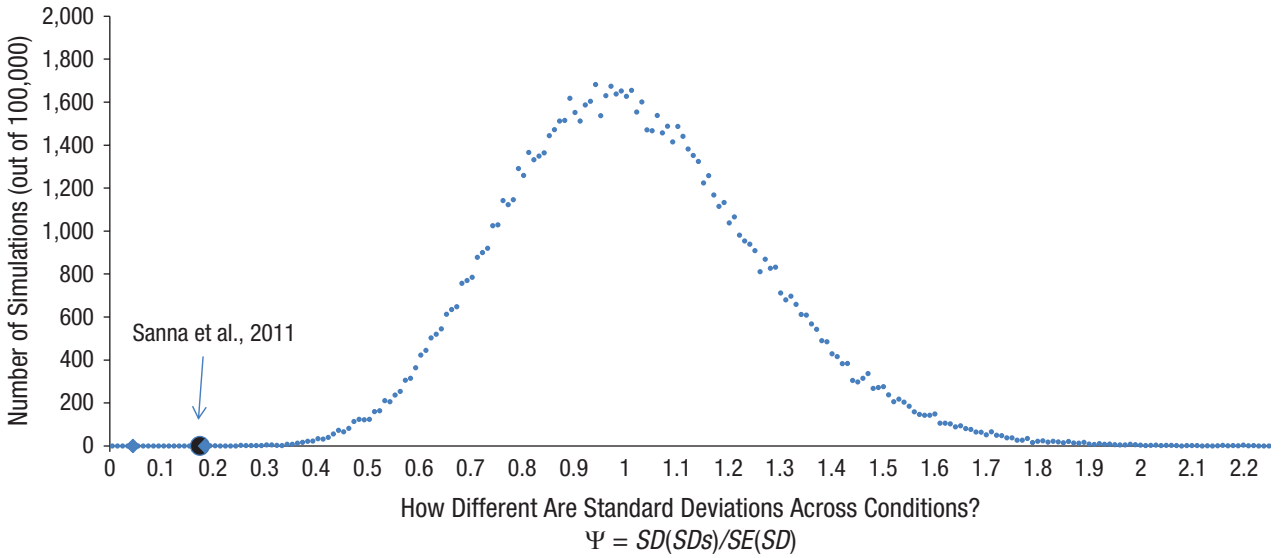

Fig. 2. Illustration of the extreme improbability of the similarity of the standard deviations reported for Sanna's three experiments on the embodiment of morality (see the text). Each condition in each experiment was simulated by drawing from a normal distribution with a mean equal to the sample mean for that condition and a standard deviation equal to the pooled standard deviation across all conditions. The standard deviation of the standard deviations for each simulated experiment was divided by the standard error of the pooled standard deviation. The graph shows the average for this value (Ψ) across the three experiments for both Sanna's reported data and 100,000 simulations of the experiments. Only 15 of the simulations yielded values that were as extreme as those for the published results.

simulations based on normal distributions. The fact that the three studies involved 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 standard deviations from reports by other authors using similar paradigms. Figure 3 shows that they obtained dramatically more varied standard deviations.

Same authors, other publications, same anomaly

Another concern regarding the simulations is their post hoc nature. I performed them *because* the standard deviations struck me as too similar. Although the low p value of .00015 provides some protection, this concern was important enough to warrant being addressed more explicitly through replications.

When looking for articles for the comparison presented in Figure 3, I searched for those citing the same articles cited by Sanna et al. (2011). Among the articles I found, three were also authored by Sanna and his colleagues, and two of these reported standard deviations. One involved a single three-condition experiment (Sanna, Chang, Parks, & Kennedy, 2009) and therefore provided limited statistical power to detect anomalous results. Nevertheless, the standard deviations were improbably similar to each other, $\Psi = 0.23$, $p = .056$. The other reported three relevant experiments, each of which included between six and nine conditions (Sanna, Parks, & Chang, 2003). The overall Ψ of 0.168 for this article is virtually impossible to obtain from random samples, $p <$ 1 in 58 billion. For details, see Section 4 in the Supplemental Material.

Convenience sample of publications

For one final comparison, I obtained the standard deviations from the first few articles in what was at the time the most recent issue of the *Journal of Experimental Social Psychology* (September 2011). To maximize comparability, I computed Ψs based on sets of two or three conditions, rather than entire articles or experiments, winding up with 39 estimates to use as a control set against 17 estimates for studies reported by Sanna and his colleagues. All studies by Sanna et al. had Ψs of 0.25 or less, but only 4 in the control set did. This difference was statistically significant by a proportions test, $p <$ 1 in 5 billion.

Analyses based on raw data

I shared all these analyses with the authors and requested the raw data behind the studies on embodiment of morality (Sanna et al., 2011). This allowed a series of additional analyses that further suggest the data did not originate in random sampling.

Simulations redone. On verifying that there were no reporting errors (i.e., ruling out the most benign of explanations), I reran all the simulations, this time drawing cards from decks representing the raw data rather than normal distributions. For each experiment, I created one card per observation and placed all the cards in a single deck (e.g., one deck with 45 cards for Study 3).[5] I then drew a card, took a note of its value, returned it to the deck, reshuffled the deck, and drew another card, until an entire simulated sample was created. I then did the same for the other two samples in each experiment. Data simulated this way, despite originating in the raw data, also almost never led to standard deviations as similar as those reported by Sanna et al., $p = .00011$, so the null hypothesis of random sampling was again rejected.[6]

Not only the standard deviations are too similar. Finally, the raw data revealed that the ranges of values the dependent variables took were also too similar. The differences between the maximum and minimum in the three conditions were as follows—Study 2: (10, 10, 10); Study 3: (76, 76, 74); and Study 4: (28, 26, 28). Maxima and minima are extremely volatile statistics when obtained from random samples.

Case Study 2: Colored Folders and Similar Means

The first anomalous finding

On the basis of the notion that the color red leads to avoidance and the color blue to approach, Smeesters and Liu (2011; retraction: Smeesters & Liu, 2013) predicted that priming participants with these colors could switch on and off contrast and assimilation effects, respectively. They reported a single 3 × 4 between-subjects experiment ($N = 169$). Instructions were given in a red, blue, or white folder and invited participants to write about one of four possible targets: Kate Moss (a fashion model), Albert Einstein, a model, or a professor. This task was followed by 20 multiple-choice general-knowledge questions. The number of correct answers was the dependent variable.

The authors expected that blue folders would generate assimilation for exemplars and stereotypes, red folders would generate contrast for exemplars and stereotypes, and white folders would generate contrast for exemplars and assimilation for stereotypes (a replication of previous findings). This hypothesis led them to predict high performance in six of the conditions and low performance

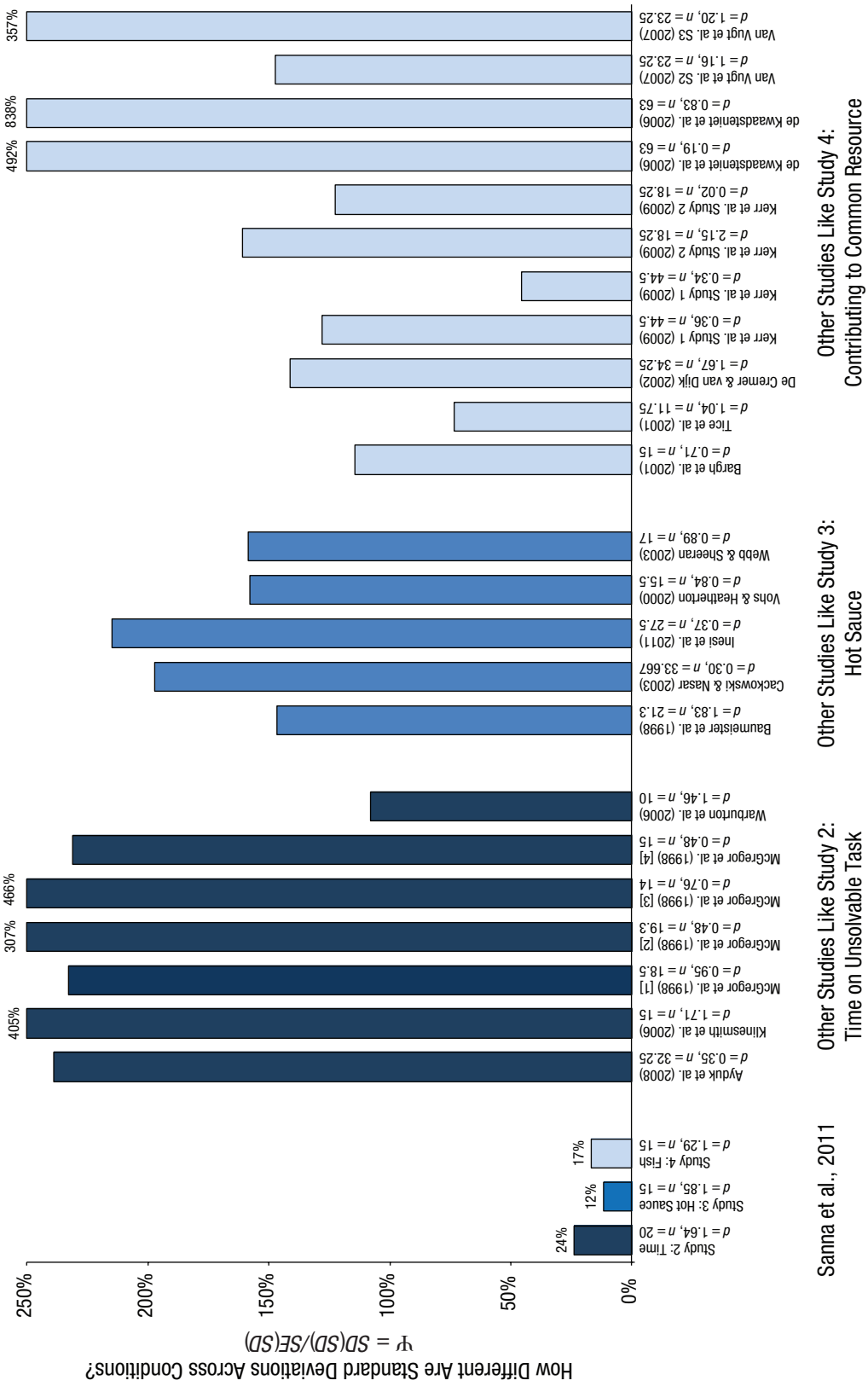

Fig. 3. Comparison of the similarity of the standard deviations in Sanna's article on the embodiment of morality (see the text) and in studies in which other authors used similar paradigms. Each bar indicates the degree of similarity in standard deviations in the study (measured in standard errors of the pooled standard deviation, Ψ). n = average number of observations per condition; d = effect size (Cohen's d).

Table 1. Means and Standard Deviations (in Parentheses) for the 12 Conditions in Smeesters and Liu (2011)

Low mean predicted	High mean predicted
9.07 (2.55)	11.43 (2.79)
9.43 (2.82)	11.71 (2.87)
9.43 (3.06)	11.77 (3.03)
9.56 (2.83)	11.85 (2.66)
9.64 (3.03)	12.00 (3.37)
9.78 (2.66)	12.07 (2.78)

Note: The table presents summary statistics for the number of correct answers (out of 20) in a general-knowledge test taken by 169 participants assigned to 12 conditions, 6 of which were predicted to have high means and 6 of which were predicted to have low means. The n in each condition was 14, except in the following cases: $n = 16$ for the fourth condition with a predicted low mean, and $n = 13$ for the third condition with a predicted high mean.

in the remaining six. The results are provided in Table 1. It shows data consistent with the predictions, but also a troubling anomaly: Means predicted to be similar are exceedingly so.[7]

Simulating the colored-folders experiment

To quantify how similar the means predicted to be similar were, I computed their standard deviation. For the high conditions, the standard deviation of the means (11.43, 11.71, . . . 12.07) was 0.23. For the low conditions, the standard deviation of the means was 0.24. As before, I divided by the standard error to obtain a scale-free metric.[8] Averaging across the high and low conditions, I arrived at a Ψ of 0.308; means predicted to be similar were 0.308 of a standard error apart from one another.

I then performed simulations to assess the likelihood that Ψ would be 0.308 or less if the data originated in random sampling. I drew samples of the reported sizes from normal distributions with μ equal to the pooled mean of the six predicted-to-be-similar conditions and σ equal to each sample's standard deviation. For example, the average number of correct answers in the high conditions was 11.81, and the standard deviation in one of those conditions was 2.79. I simulated this condition drawing from $N(11.81, 2.79)$.

Note 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 would lead to exactly the same average performance as the combination of a white folder with an Einstein prime. Only 21 of 100,000 simulations had a Ψ of 0.308 or less (see Fig. 4).

Same authors, other publications, same anomaly

Concerned that the previous analysis was at least somewhat post hoc (i.e., I examined mean similarity *because* means seemed too similar), I analyzed two other articles by Smeesters.[9] I selected them because they had been completed recently and had multiple conditions predicted to be similar.

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

For each article, I identified a dependent variable employed across all studies, selected all conditions in which that variable was predicted to show no effect or be at baseline, and analyzed the degree of similarity for those means, using simulations analogous to those presented earlier. The null hypothesis of random sampling was rejected for both articles, $p = .00011$ and $p = .0023$ (see Section 5 in the Supplemental Material).

Analyses based on raw data of the colored-folders study

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

Simulations redone. On verifying that there were no reporting errors, I reran all the simulations, this time drawing cards from the raw data instead of normal distributions. I performed two versions of the simulations. In both, I created two decks of cards, one with the 86 observations from the low conditions and one with the 83 observations from the high conditions, and generated high and low samples by drawing from the respective decks. One version of the simulations was with replacement: As before, I drew one card, made a note of it, returned it to the deck, reshuffled the deck, drew another card, and so on, until all conditions had been simulated. The other version of the simulations was without replacement: I shuffled the cards and dealt them all at once into the six corresponding conditions, thereby keeping the exact set of observations constant; the only thing that varied was which high condition got which high card and which low condition got which low card. The estimated p values for the null hypothesis of random sampling were .00030 and .00018 for the with- and without-replacement simulations, respectively—values

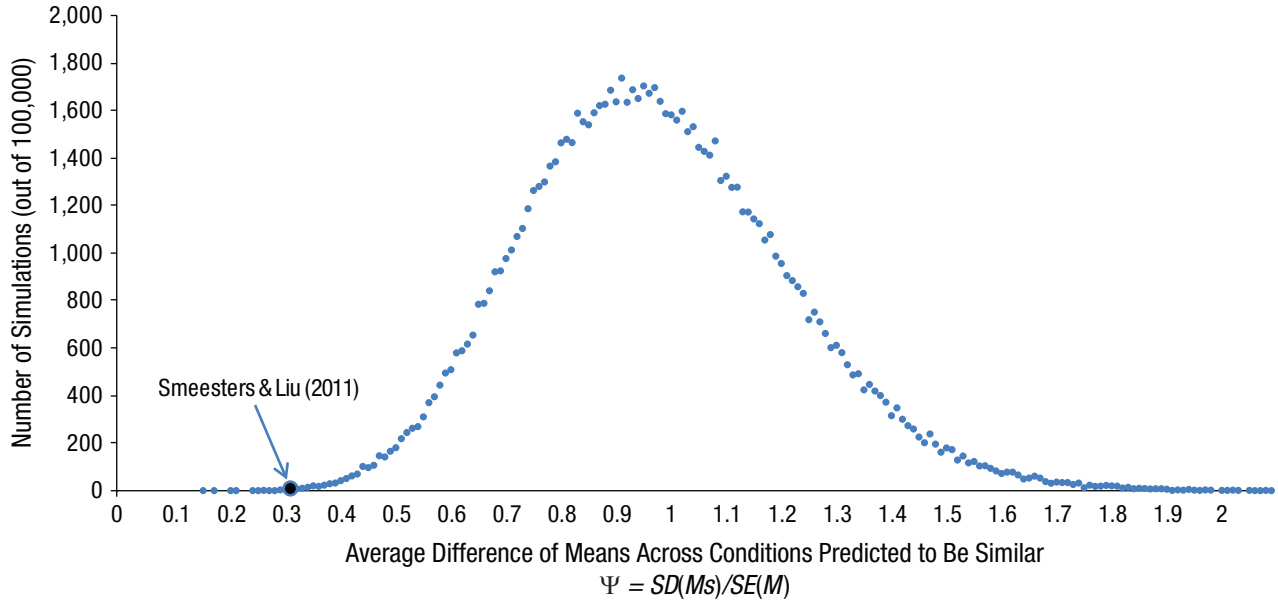

Fig. 4. Illustration of the extreme improbability of the similarity of the means reported for Smeesters's colored-folders experiment (Smeesters & Liu, 2011). Each of the six high conditions was simulated by drawing from the normal distribution with the mean equal to the pooled mean for those conditions and the standard deviation of the respective sample; simulation of the six low conditions proceeded analogously. The difference of means predicted to be similar (Ψ) was 0.308 for Smeesters's reported data; that is, means predicted to be similar differed by just 0.308 of a standard error from each other. Only 21 of the 100,000 simulations yielded such a low Ψ.

comparable to the *p* value obtained in the simulations assuming normality (i.e., .00021).

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)—the belief that even small samples 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. Although four identical coin tosses is an unlikely event (12.5%), it is more likely than the nearly 0% probability observed in people's imagined coin tosses.

If a person believing in the law of small numbers were asked to generate scores for the 12 conditions in the colored-folders study, we might analogously expect him or her to avoid too many of the same scores in a given condition, that is, to generate sets of scores that are too evenly distributed. 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 14 scores for 1 of the 12 conditions were [6, 7, 7, 8, 8, 9, 9, 10, 10, 10, 12, 12, 14, 15]. The mode was 10, and it appeared three times. Nine of the 12 conditions had the mode appearing three times, and 3 had it appearing just two times. Hence, the sum of mode frequencies, **F**, was 33 (9 × 3 + 3 × 2).

How unlikely is it for **F** to be 33 or less? It occurred just 21 times in the 100,000 bootstrap simulations with replacement (see Fig. 5), and 93 times in the 100,000 bootstrap simulations without replacement. This test of the data having originated in random sampling, then, also rejected the null hypothesis of random sampling. It is interesting to establish how independent the excessive similarity of means (Ψ) and the excessive evenness of scores (**F**) were. Was this a matter of observing the same red flag twice or of seeing two red flags? For each of 100,000 simulations, I had values for Ψ and **F**. The correlation of these two metrics, it turned out, was quite low, $r = .16$ for the simulations with replacement and $r = .18$ for the simulations without replacement. Thus, it was more a matter of observing two flags than of observing the same flag twice. Not a single simulation had both **F** of 33 or less and Ψ of 0.308 or less; random sampling would seem to never lead to the observed data.[10]

Nonexplanations

When interviewed by the Erasmus committee examining possible misconduct on his part (Zwaan, Groenen, van der Heijden, & te Lindert, 2012), Smeesters said he "possibly made a coding mistake in two questions" (p. 4) and that it is possible that "people who answered a difficult question correctly always answered another difficult question correctly" (p. 3). Although both of these things

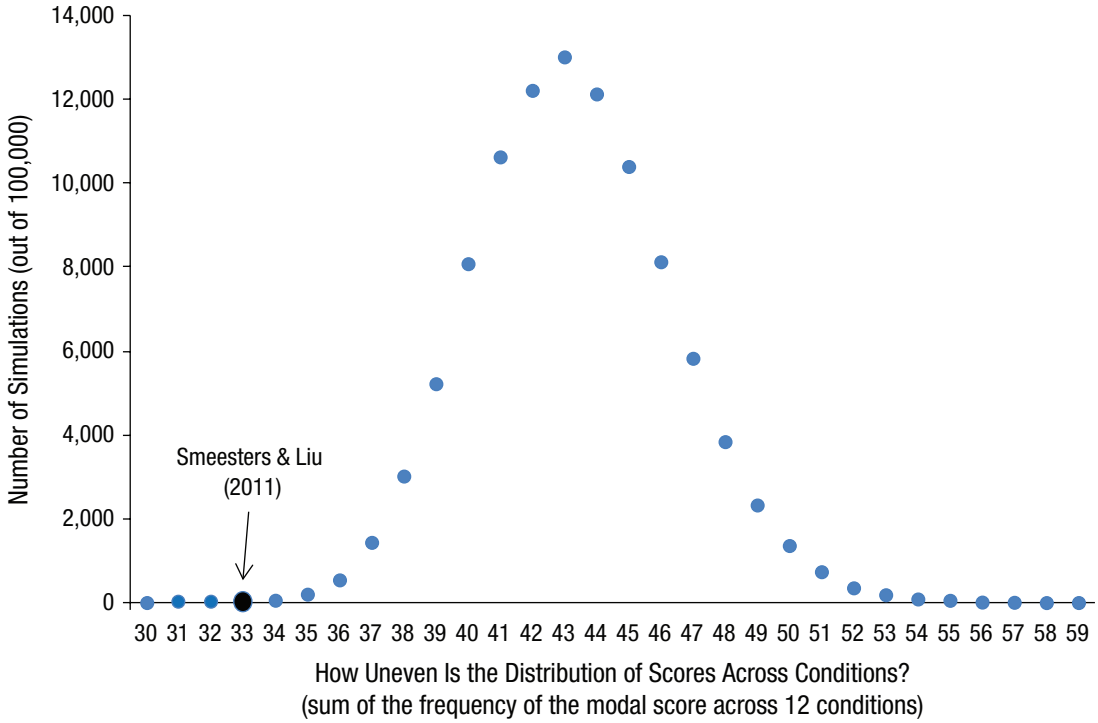

Fig. 5. Illustration that general-knowledge scores were more evenly distributed across the 12 conditions in Smeesters and Liu's (2011) colored-folders study than in 99.9% of its 100,000 simulations. The single study reported by Smeesters was simulated by drawing cards with replacement from the raw data. The *x*-axis corresponds to the sum of the frequencies of the indicated mode across the 12 conditions. For example, if the mode appeared four times in each of the conditions, the sum was 48; if there were no repeated scores in any conditions, the sum was 12. In Smeesters's raw data, the sum was 33. Only 21 of the 100,000 simulations had such a low sum (or lower).

may be true, neither could account for the excessive similarity of means or the 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 were 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 the likelihood of that person getting another question right, then we would see more rather than fewer people with the same scores within conditions. In the extreme, if participants knowing one answer knew all of them, we would see scores of 20 and 0 only. Again, we are trying to explain the opposite, that there were too few, not too many, people with the same scores in each condition.

Finally, and just as important, neither sloppy coding nor correlated answers to the questions can possibly account for the simulations showing that the data are incompatible with random samples, because these simulations drew from those raw data. The simulations already took into account all such idiosyncrasies. If 6 people got higher scores than they should have because of sloppy grading, for example, then there would have been six cards with higher scores than there should have been, and the simulations would have drawn from them. Similarly, if everyone getting Question 3 right also got Question 6 right, all cards for a person with Question 3 right would have had Question 6 right.

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

One of Smeesters's two additional publications analyzed to replicate the similarity-of-means analysis contains studies in which participants were asked to indicate their maximum willingness to pay (WTP) for each of two black T-shirts with very similar designs (see Fig. 6). I obtained the raw data for one of these studies (Study 3) and compared it with data from several other studies eliciting WTP: two published in journals that post data (Fudenberg, Levine, & Maniadis, 2012; Newman & Mochon, 2012), four by colleagues who at some point had e-mailed me

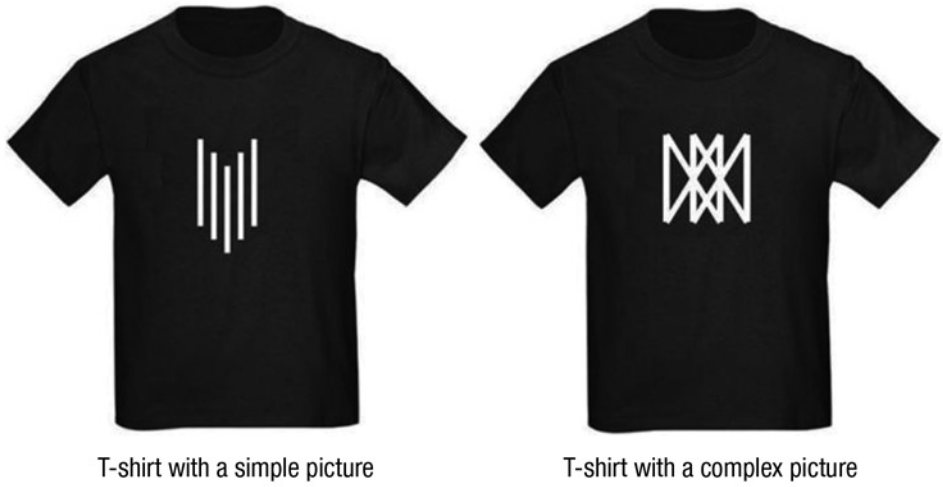

Fig. 6. T-shirt images used in the willingness-to-pay study by Smeesters (see note 9) and my replication. Participants saw one T-shirt image at a time.

their WTP data (Frederick, 2012; Keren & Willemsen, 2009; Simonson & Drolet, 2004; Yang, Vosgerau, & Loewenstein, 2012), a previous publication of my own (Simonsohn, 2009), and a new study I ran just for this analysis, asking participants to indicate their WTP for the same two T-shirts from Figure 6 (see Section 6 in the Supplemental Material). The data from Smeesters's WTP study stand out markedly from all of these benchmark studies in a variety of ways. In what follows, I discuss three examples.

Multiples of $5 in the willingness-to-pay study. A striking pattern in Smeesters's WTP data 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 expected in the absence of such a practice (20%). Figure 7 shows that this is the case for all the benchmark studies, including two also run in The Netherlands and employing euros as the currency and three in which participants were excluded, as Smeesters claims to have done, on the basis of the instructional manipulation check (Oppenheimer, Meyvis, & Davidenko, 2009; see the Another Nonexplanation section). In Smeesters's study, however, the rate of multiples of $5 is at the 20% baseline. That article, recall, also exhibits excessive similarity of means, an orthogonal anomaly. The upward trend evident in the figure indicates that multiples of $5 are more common among higher valuations than among lower valuations.

Correlation of valuations. Recall that every participant indicated his or her WTP for both T-shirts in short succession. Individual differences in income, liking of T-shirts, attentiveness, and other variables should lead these WTPs to be correlated. The benchmark studies showed strong correlations in WTP for even completely unrelated items that were valued back-to-back by the same respondents. The Cronbach's αs for overall WTP correlations were .48 for an air purifier, a DVD box, a chocolate bar, and candles (Frederick, 2012); .62 for a toaster, a telephone, a backpack, and headphones (Simonson & Drolet, 2004); and .52 for a planner, a keyboard, a calculator, a book, chocolates, and a computer mouse (Fudenberg et al., 2012).

In contrast, the correlation between the valuations of the two nearly identical T-shirts in the suspicious study was negative, $r = -.67$.[11] 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 was highly significant, $Z = 13.82$, with a p of effectively 0.

Correlation in use of $5 multiples. Combining the ideas behind the previous two analyses, I examined if the use of multiples of $5 was correlated within subjects. There are again many reasons to expect that this would be the case in a sample from real valuations (e.g., respondents may differ in their tendency to use round numbers or their uncertainty in valuating T-shirts). Such a tendency was indeed correlated in the benchmark studies, αs = .64, .58, and .57, respectively. The correlation in the suspected study, however, was negligible, $r = -.04$. In the replication with the same T-shirts, r was .62. The difference between these two correlations was again highly significant, $Z = 5.52$, $p < 1$ in 58 million.

Another nonexplanation

Smeesters also told the Erasmus committee that dropping participants failing an instructional manipulation check (Oppenheimer et al., 2009) may explain the excessive

Fig. 7. Illustration that participants in all studies included in the comparison of willingness-to-pay data, except for Smeesters's study (see note 9), disproportionately employed multiples of $5 in their valuations. Each dot indicates the percentage of valuations in a given study that were multiples of $5. Items are sorted by average valuation (indicated by the mean after each item's description). The first two items are from a replication study in which participants valued the same two T-shirts used in Smeesters's study. The remaining valuations come from various studies for which raw willingness-to-pay data were available (see the text). The horizontal line represents expectations in the absence of any preference for multiples of $5. Items from studies in which participants were eliminated for failing an instructional manipulation check are marked with "#."

similarity of means in his data. This does not make sense. Excluding noisy answers lowers standard deviations, making means seem more *different* from each other, because their difference is benchmarked against the standard deviations (see, e.g., the literature examining the impact of trimming means on standard errors; Keselman, Othman, Wilcox, & Fradette, 2004; Yuen, 1974).

Even if a researcher were to delete the highest few observations in conditions predicted to have low values, and the lowest few in conditions predicted to have high values, the result would be that means predicted to be similar would be too different, not too similar, to each other. In any case, I examined whether means in the article first proposing this technique (Oppenheimer et al., 2009) and in a recent study using it across seven experiments (Yang et al., 2012) were too similar; they were not (see Section 7 in the Supplemental Material).

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 on, by anyone conducting these types of analyses in the future: First, replicate analyses across multiple studies before suspecting foul play by a given author.[12] Second, compare suspected studies with similar ones by other authors. Third, extend analyses to raw data. Fourth, contact authors privately and transparently, and give them ample time to consider your concerns. Fifth, offer to discuss matters with a trusted statistically savvy advisor. Sixth, give the authors more time. Finally, if suspicions remain, convey them only to entities tasked with investigating such matters, and do so as discreetly as possible.

Author Contributions

U. Simonsohn is the sole author of this article and is responsible for its content.

Acknowledgments

The data and SAS code behind all results in this article are available from http://opensciencefra mework.org/project/HwG3W/files. I thank Leif Nelson and Joe Simmons for invaluable support and contributions to this project from beginning to end. Nick Epley, Christophe Van den Bulte, Hal Pashler, and Henry L. Roediger, III, helped dramatically improve the writing and positioning of this article. Conversations with Rolf Zwaan, who headed the committee investigating Smeesters, proved instrumental for some of the analyses presented here. The coauthors of both Sanna and Smeesters were responsive and collaborative throughout, despite the obvious difficulty of confronting the possibility that a trusted friend, colleague, and advisor had betrayed their trust and endangered their reputations for years. This project did not benefit from the refusal of the University of Michigan and the University of North Carolina to share the outcome of their investigation into possible misconduct by Lawrence Sanna. I am exclusively responsible for all errors that remain.

Declaration of Conflicting Interests

The author declared that he had no conflicts of interest with respect to his authorship or the publication of this article.

Supplemental Material

Additional supporting information may be found at http://pss.sagepub.com/content/by/supplemental-data

Notes

1. In Section 1 of the Supplemental Material available online, I contrast the present case with the famous cases of Gregor Mendel and Cyril Burt. In Section 2, I discuss why Benford's (1938) law, a regularity involving the frequency of digits across numbers that can be used to detect irregularities in data, is not generally applicable to psychological data.
2. The retraction is by Johnson, Smeesters, and Wheeler (2012).
3. Hypothesis testing traditionally examines whether data differ too much from a null hypothesis. Excessive-similarity tests, of course, examine the opposite. A reviewer pointed out that this may mean one ought to correct α levels to include rejection of the null when data are too similar in addition to too dissimilar to the null, in effect making two-sided tests three-sided, and one-sided tests two-sided. This reviewer wrote, "*F* tests should be two-tailed instead of one-tailed, and therefore the critical *p* value for significance should be 2.5% for high *F* values and 2.5% for small *F* values." Critical α levels ought to be set taking into account the costs and benefits of false positives versus false negatives. For fraud, the costs of a false positive are so high that an α of 2.5% seems unacceptably liberal. A more reasonable α may be 1/1,000, or 1/10,000, or lower still. Hence, we would need to reduce α for traditional too-much-deviation tests from 5% merely to 4.99% to accommodate the 0.01% of the too-little-deviation test. At that point, we might as well maintain the (anyway arbitrary) α of 5% overall (close enough to 4.99%) and remember to use very small αs when considering fraud.
4. The standard error of the standard deviation, assuming normality, is calculated as $SD/\sqrt{2n}$ (Yule, 1922). For the hot-sauce study, for example, the standard deviation was 25.11, and n was 15, so the standard error of the standard deviation was 4.58 ($25.11/\sqrt{30}$). Note that the standard error is used merely for scaling, so although it is only approximate, the results do not hinge on possible deviations from such approximation. For more details, see Section 3 in the Supplemental Material.

5. To simulate from the raw data under the null hypothesis of equal variances, one first must subtract the condition's mean from each observation (for more details, see Boos & Brownie, 1989).
6. These simulations included both number of fish and mood from Study 4. Without the latter dependent variable, the *p* value was .00083
7. Some of these were not reported in the original article and were obtained via e-mail from Smeesters.
8. In the first case study, I analyzed similarity of standard deviations, so I used the standard error of the standard deviation, approximately $SD/\sqrt{2n}$. For this case study, I analyzed similarity of means, so I used the standard error of the mean, $SD/\sqrt{n}$.
9. These two additional articles are Liu, Smeesters, and Trampe (2012) and Smeesters, Wheeler, and Kay (2009). The former article reported the study referred to here as Smeesters's willingness-to-pay study.
10. A student of Smeesters failed to replicate this study. His data did not show the excessively evenly distributed pattern (Zwaan et al., 2012, p. 27).
11. The correlation was also negative within the four conditions in which there was neither a predicted nor an observed difference in valuations of the T-shirts ($r = -.64$).
12. As a reviewer pointed out, although the probability that any given odd pattern is present in a data set is low, the probability that some odd pattern is present is obviously much greater. Replicating the same analyses across multiple studies by the same authors is a way to prevent the possibility of erroneously concluding that a given pattern is inconsistent with proper sampling merely because one considered too many patterns and failed to correct for the multiple comparisons that were carried out. Another way to prevent this problem is to use very small αs (see note 2). For example, although the WTP abnormalities observed in Smeesters's data were tested only in a single publication of his, the *p* values associated with the null hypothesis of random sampling are low enough (< 1 in several million) to protect against the concern of a higher family-wise error rate due to multiple comparisons.

References

Ayduk, O., Gyurak, A., & Luerssen, A. (2008). Individual differences in the rejection-aggression link in the hot sauce paradigm: The case of rejection sensitivity. *Journal of Experimental Social Psychology*, *44*, 775–782.

Bargh, J. A., Gollwitzer, P. M., Lee-Chai, A., Barndollar, K., & Trotschel, R. (2001). The automated will: Nonconscious activation and pursuit of behavioral goals. *Journal of Personality and Social Psychology*, *81*, 1014–1027.

Baumeister, R. F., Bratslavsky, E., Muraven, M., & Tice, D. M. (1998). Ego depletion: Is the active self a limited resource? *Journal of Personality and Social Psychology*, *74*, 1252–1265.

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

Boos, D. D. (2003). Introduction to the bootstrap world. *Statistical Science*, *18*, 168–174.

Boos, D. D., & Brownie, C. (1989). Bootstrap methods for testing homogeneity of variances. *Technometrics*, *31*, 69–82.

Cackowski, J. M., & Nasar, J. L. (2003). The restorative effects of roadside vegetation. *Environment and Behavior*, *35*, 736–751.

Carlisle, J. B. (2012). The analysis of 169 randomised controlled trials to test data integrity. *Anaesthesia*, *67*, 521–537.

De Cremer, D., & van Dijk, E. (2002). Reactions to group success and failure as a function of identification level: A test of the goal-transformation hypothesis in social dilemmas. *Journal of Experimental Social Psychology*, *38*, 435–442.

de Kwaadsteniet, E. W., van Dijk, E., Wit, A., & de Cremer, D. (2006). Social dilemmas as strong versus weak situations: Social value orientations and tacit coordination under resource size uncertainty. *Journal of Experimental Social Psychology*, *42*, 509–516.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. Boca Raton, FL: Chapman & Hall.

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*, 1–21.

Fudenberg, D., Levine, D. K., & Maniadis, Z. (2012). On the robustness of anchoring effects in WTP and WTA experiments. *American Economic Journal: Microeconomics*, *4*, 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*, 298–301.

Inesi, M. E., Botti, S., Dubois, D., Rucker, D. D., & Galinsky, A. D. (2011). Power and choice: Their dynamic interplay in quenching the thirst for personal control. *Psychological Science*, *22*, 1042–1048.

Johnson, C. S., Smeesters, D., & Wheeler, S. C. (2012). Retraction of Johnson, Smeesters, and Wheeler (2012). *Journal of Personality and Social Psychology*, *103*, 605.

Kalai, G., McKay, B., & Bar-Hillel, M. (1998). *The two famous rabbis experiments: How similar is too similar (Discussion Paper 182)*. Jerusalem, Israel: Center for the Study of Rationality.

Keren, G., & Willemsen, M. C. (2009). Decision anomalies, experimenter assumptions, and participants' comprehension: Revaluating the uncertainty effect. *Journal of Behavioral Decision Making*, *22*, 301–317. doi:10.1002/bdm.628

Kerr, N. L., Rumble, A. C., Park, E. S., Ouwerkerk, J. W., Parks, C. D., Gallucci, M., & Van Lange, P. A. M. (2009). "How many bad apples does it take to spoil the whole barrel?": Social exclusion and toleration for bad apples. *Journal of Experimental Social Psychology*, *45*, 603–613.

Keselman, H., Othman, A. R., Wilcox, R. R., & Fradette, K. (2004). The new and improved two-sample *t* test. *Psychological Science*, *15*, 47–51.

Klinesmith, J., Kasser, T., & McAndrew, F. T. (2006). Guns, testosterone, and aggression: An experimental test of a mediational hypothesis. *Psychological Science*, *17*, 568–571.

Liu, J., Smeesters, D., & Trampe, D. (2012). Effects of messiness on preferences for simplicity. *Journal of Consumer Research*, *39*, 199–214.

McGregor, H. A., Lieberman, J. D., Greenberg, J., Solomon, S., Arndt, J., Simon, L., & Pyszczynski, T. (1998). Terror management and aggression: Evidence that mortality salience motivates aggression against worldview-threatening others. *Journal of Personality and Social Psychology*, *74*, 590–605.

Newman, G. E., & Mochon, D. (2012). Why are lotteries valued less? Multiple tests of a direct risk-aversion mechanism. *Judgment and Decision Making*, *7*, 19–24.

Oppenheimer, D., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*, 867–872.

Roberts, S. (1987). Less-than-expected variability in evidence for three stages in memory formation. *Behavioral Neuroscience*, *101*, 120–130.

Rubin, D. B., & Stigler, S. M. (1979). Dorfman's data analysis. *Science*, *205*, 1204–1206.

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 [Retracted article]. *Journal of Experimental Social Psychology*, *47*, 472–476.

Sanna, L. J., Chang, E. C., Miceli, P. M., & Lundberg, K. B. (2013). Retraction notice to "Rising up to higher virtues: Experiencing elevated physical height uplifts prosocial actions" [Journal of Experimental Social Psychology 47 (2010) 472–476]. *Journal of Experimental Social Psychology*, *49*, 316.

Sanna, L. J., Chang, E. C., Parks, C. D., & Kennedy, L. A. (2009). Construing collective concerns: Increasing cooperation by broadening construals in social dilemmas. *Psychological Science*, *20*, 1319–1321.

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*, 26–40.

Simonsohn, U. (2009). Direct risk aversion: Evidence from risky prospects valued below their worst outcome. *Psychological Science*, *20*, 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*, 681–690.

Smeesters, D., & Liu, J. E. (2011). The effect of color (red versus blue) on assimilation versus contrast in prime-to-behavior effects [Retracted article]. *Journal of Experimental Social Psychology*, *47*, 653–656.

Smeesters, D., & Liu, J. E. (2013). Retraction notice to "The effect of color (red versus blue) on assimilation versus contrast in prime-to-behavior effects" [Journal of Experimental Social Psychology 47 (2011) 653–656]. *Journal of Experimental Social Psychology*, *49*, 315.

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*, 395–414.

Sternberg, S., & Roberts, S. (2006). Nutritional supplements and infection in the elderly: Why do the findings conflict? *Nutrition Journal*, *5*(1), 30. Retrieved from http://www.nutritionj.com/content/5/1/30

Tice, D. M., Bratslavsky, E., & Baumeister, R. F. (2001). Emotional distress regulation takes precedence over impulse control: If you feel bad, do it! *Journal of Personality and Social Psychology*, *80*, 53–67.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*, 105–110.

Van Vugt, M., De Cremer, D., & Janssen, D. P. (2007). Gender differences in cooperation and competition. *Psychological Science*, *18*, 19–23.

Vohs, K. D., & Heatherton, T. F. (2000). Self-regulatory failure: A resource-depletion approach. *Psychological Science*, *11*, 249–254.

Warburton, W. A., Williams, K. D., & Cairns, D. R. (2006). When ostracism leads to aggression: The moderating effects of control deprivation. *Journal of Experimental Social Psychology*, *42*, 213–220.

Webb, T. L., & Sheeran, P. (2003). Can implementation intentions help to overcome ego-depletion? *Journal of Experimental Social Psychology*, *39*, 279–286.

Wicherts, J. M. (2011). Psychology must learn a lesson from fraud case. *Nature*, *480*, 7.

Wicherts, J. M., & Bakker, M. (2012). Publish (your data) or (let the data) perish! Why not publish your data too? *Intelligence*, *40*, 73–76.

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, *61*, 726–728.

Yang, Y., Vosgerau, J., & Loewenstein, G. F. (2012). *The influence of framing on willingness to pay as an explanation of the uncertainty effects*. Unpublished manuscript, Carnegie Mellon University.

Yuen, K. K. (1974). The two-sample trimmed *t* for unequal population variances. *Biometrika*, *61*, 165–170.

Yule, G. U. (1922). *An introduction to the theory of statistics*. London, England: Charles Griffen.

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*. Retrieved from http://www.eur.nl/fileadmin/ASSETS/press/2012/Juli/report_Committee_for_inquiry_prof._Smeesters.publicversion.28_6_2012.pdf