BASIC AND APPLIED SOCIAL PSYCHOLOGY
2016, VOL. 38, NO. 1, 3–18
http://dx.doi.org/10.1080/01973533.2015.1124767

Routledge
Taylor & Francis Group

A Social Priming Data Set With Troubling Oddities

Harold Pashler[a], Doug Rohrer[b], Ian Abramson[a], Tanya Wolfson[c], and Christine R. Harris[a]

[a]University of California, San Diego; [b]University of South Florida; [c]UCSD Supercomputer Center

ABSTRACT

A recent paper by Chatterjee, Rose, and Sinha (2013) reported impressively large "money priming" effects: incidental exposure to concepts relating to cash or credit cards made participants much less generous with their time and money (after cash primes) or much more generous (after credit card primes). Primes also altered participants' choices in a word-stem completion task. To explore these effects, we carried out re-analyses of the raw data. A number of strange oddities were brought to light, including a dramatic similarity of the filler word-stem completion responses produced by the 20 subjects who contributed most to the priming effects. We suggest that these oddities undermine the credibility of the paper and require further investigation.

Chatterjee, Rose, and Sinha (2013) presented results from three experiments investigating social priming—specifically, priming effects induced by incidental exposure to concepts relating to cash or credit cards. They reported that exposing people to cash concepts made them less generous with their time and money, whereas exposing them to credit card concepts made them more generous.

The article by Chatterjee et al. (2013) was brought to the attention of the present authors by an investigator working on social priming. Struck by the large effect sizes in the first two studies, and hoping to better understand how such large effects had emerged, we requested the raw data. The authors provided us with three Microsoft Excel files containing summary data for each experiment, as well as the materials used in the studies. Later, they also provided an Excel file with additional data for the third study: the subjects' word-stem completion responses. Examining these data brought to light a number of oddities.

These peculiarities—described in detail next—included extraordinary similarity of word-stem completion responses attributed to the subset of subjects who contributed the most to the large effects reported in the original article. As we show, this high rate of reduplication appears in a number of distinct aspects of the data, and the likelihood that it would have occurred by chance are found to be infinitesimally small. Furthermore, as pointed out to us by a reviewer of a previous version of our paper, there were puzzling inconsistencies between the materials used in data collection and the subjects' reported responses. We argue here that these findings make it extremely unlikely that the data were produced by the process described in the Method section of the original article.

Prior to the preparation of the current article, we brought the concerns described here to the attention of the original authors via extensive e-mail correspondence. Over a long period, these authors argued that our findings did not undermine the credibility of their data, and in a review of an earlier version of this paper and in a lengthy e-mail correspondence, they sought to rebut our concerns. We discuss some of their reactions next.

Overview of the experiments

In all three studies reported by Chatterjee et al. (2013), the priming manipulation involved asking subjects to solve sentence descrambling tasks. Here, subjects see five words and try to construct a valid sentence using four of the words. Subjects were randomly assigned to descramble sentences related to cash (*cash condition*) or credit cards (*credit condition*) or (in some studies) neither one (*neutral condition*). For example, in the credit condition a subject might be asked to create a sentence with four of the five words in "TV shall watch we Visa," for which a solution might be "we shall watch TV." The authors contended that their data demonstrated that "priming cash concepts reduces

CONTACT Hal Pashler hpashler@ucsd.edu Department of Psychology, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/hbas.

willingness to help others, while activating credit card concepts reverses these effects" (p. 109).

The choice of dependent variables varied slightly across studies. In Experiment 1, subjects were given four 25-cent coins ("quarters") in return for completing the study. Then, as each subject left the lab, the experimenter mentioned that the lab was accepting donations for the "University Student Fund" and that, if they wished, subjects could put some of the quarters in a box to support this charitable effort. The article reported that, on average, subjects in the credit condition donated more money ($M = 73$ cents) than did subjects in the neutral condition (41 cents), who in turn donated more money than did subjects in the credit condition (27 cents). Experiment 2 was similar to Experiment 1, except that subjects were invited to volunteer time rather than money for a charitable cause. Experiment 3 focused on cash and credit priming effects upon subjects' completions of word stems (although donation choices were again reported). The specific procedures of Experiment 3 are discussed in more detail further below.

Unusually large effect sizes

Although Chatterjee et al. (2013) did not report effect sizes, before requesting the raw data we calculated effect sizes in the form of Cohen's *d* (*d* represents the difference between the conditions divided by the pooled standard deviation). In Experiment 1, for the comparison between credit and cash conditions, the effect size for amount of money donated was $d = 2.19$. In Experiment 2, subjects indicated how many hours per month they would be willing to volunteer to help an organization described as "University Student Welfare." The effect sizes for the differences between the control condition and the cash and credit conditions were each about $d = 1.6$ in magnitude, and the difference between the latter two conditions measured $d = 2.98$. These are all very large effects for any behavioral science experiment—particularly surprising, it seemed to us, to find in the context of a study with such a subtle and indirect manipulation.

To place these effect sizes in context, it may be illuminating to compare them with reasonable priors for effect sizes in general. Large-scale surveys of the social-psychological literature show that *published* effects average $d = .45$ (Richard, Bond, & Stokes-Zoota, 2003; Westfall, 2015), but of course these are undoubtedly inflated due to publication bias. To provide a more accurate estimate of typical effect sizes, Simmons, Nelson, and Simonsohn (2013) studied a large sample (697 participants) and deliberately chose a number of glaringly "obvious" effects to measure, reporting all of their measures. For example, they tested the hypothesis that men weigh more than women and the hypothesis that people who like eggs tend to eat more egg salad than do people who do not like eggs. The effect sizes for these two effects were $d = 0.59$ and 1.07, respectively. As another point of comparison involving an experimental manipulation, Lobbestael, Arntz, and Wiers (2008) compared four different methods used by emotion researchers to elicit anger in subjects. The manipulations were far from subtle, for example, repeatedly verbally harassing subjects produced effects. Yet the observed effects on subjects' self-reported anger averaged in the range from $d = 0.64$ to $d = 0.74$. As Simmons et al. pointed out, a *d* of 1.0 tends to occur only with effects so potent that they are easily detected through casual observation. Thus, to return to the Chatterjee et al. (2013) article, finding that subtle priming manipulations yield effects in the range of 1.5–3 seemed to us to be quite extraordinary.

Of course, the very large effect sizes reported by Chatterjee et al. (2013) could have been inflated by sampling error. Even if that were the case, however, the occurrence of multiple unusually large effects within a single article seemed odd to us. Seeking to explore the data set in greater detail, we requested the raw data to see if the data sets had other unusual features. The authors were kind enough to provide their data (which can be downloaded at http://laplab.ucsd.edu/Chattdata/). Our examination of these data is the main focus of the remainder of this article.

Reduplication of nontarget responses in Study 3

The most remarkable oddities that were turned up by our examination of the data related to Study 3, and we focus on that study throughout this article. In Study 3, subjects were assigned to one of two priming conditions (cash or credit) and performed the corresponding sentence unscrambling task. They were then given information about a nonprofit organization (the Nature Conservancy), reading a page about the potential benefits and costs of volunteering to work with this organization (see Appendix A).

Next, subjects performed a word-stem completion task, which required that they complete 25 word stems (e.g., TI___ or BR____). These stems are listed in Appendix B. Eight of the stems could be completed with a "benefit-related word" that had been mentioned in the immediately preceding reading about the Nature Conservancy (e.g., the stem OP_ could be completed by

OPPORTUNITIES). Another eight of the stems could potentially be completed with a "cost-related" word (e.g., TI___ could be completed by TIME).

Each individual was scored on the number of Benefit and Cost words that they filled in. Thus, each subject will be described as having a score on the measure "benefit word completions" and a score on the measure "cost word completions," with each of these scores ranging from 0 to 8. When we refer to a subject as being in "the (3, 4) group," we are referring to a subject who produced three of the eight possible benefit target words and four of the eight possible cost target words. In addition to the 16 stems that could be completed by a benefit or cost word, the list of 25 stems also contained an additional nine *filler word* stems chosen to be unrelated to any of the cost words, benefit words, or anything else involved in the study. (In the original article by Chatterjee et al., 2013, these filler words are referred to as *neutral words*.)

The key finding that the authors reported from this experiment was that the cash versus credit priming manipulation affected the number of benefit word completions and cost word completions that the subjects produced (cash priming making them more likely to produce cost words, credit priming making them more likely to produce benefit words).

We would note that the priming effect reported by the authors in Study 3 is not as simple or straightforward as it may seem. It would seem very plausible that exposure to the concept of cash might lead people to think more about cash or to pick word-stem completions related to cash (a commonsense sort of priming effect confirmed in many laboratory studies). That was not the effect reported in the study, nor was the effect anything that directly paralleled the findings of Study 1 and 2 (e.g., thinking about cash making people complete word stems with fewer altruism-related completions, and thinking about credit having the opposite effect). Rather, it was reported that exposure to cash made people produce completions focused on the costs of volunteering, and this in turn selectively amplified the priming effect produced by exposure to words representing costs of volunteering as revealed in a subsequent word-stem completion test. Exposure to the notion of credit, on the other hand, had a corresponding effect on word described earlier as representing potential benefits of volunteering.

We emphasize that words are categorized as cost or benefit words based on the materials in Appendix A, because if one inspects the actual lists of cost and benefit word completions, one notices that many of them would probably not strike most people as a cost or a benefit, respectively. For example, the cost word list included *Travel, Food, Supplies,* and *Pocket*—so classified because they appeared in a paragraph the subject read listing "Cons of volunteer work at the Nature Conservancy." What is odd about this is that many people undoubtedly have positive feelings about travel, food, and perhaps even supplies and pockets. So the effect reported in the experiment would have to reflect the content of the material the subjects read: A brief, incidental exposure to the concept of cash would have to be changing people's mental state in such a way as to strongly prime some other (often intrinsically neutral or positive) words that had been listed as "costs of volunteering for the Nature Conservancy" within text that subjects had seen a few minutes earlier (the content seen in Appendix A).

Our examination revealed numerous small material-preparation and scoring errors in the execution of the study, which we would have expected to have attenuated such priming effects. A complete list of all the errors that were uncovered in our examination of the study is provided in Appendix C.

Although the errors described in Appendix C suggest that there may have been a great deal of sloppiness in the execution of the study, making the appearance of very strong priming effects all the more surprising, these defects in execution are not the main point of the present article. We turn now to what we see as much more troubling discoveries relating to the Study 3 data set, which in our view cannot so readily be attributed to sloppiness or oversight. We begin our analysis by asking about the distribution of key dependent variables. Figure 1 plots each subject's benefit word completion score against his or her cost word completion score, generating a lattice-like structure. There is a strong negative correlation between these two dependent variables ($\rho = -.37$, $p = .0002$, $n = 94$). In addition, the figure exhibits a striking (and, to us, quite odd) pattern going beyond this negative correlation. There were nine subjects who provided five target benefit words and no target cost words—what we call the (5,0) group or "node"—and 11 subjects who provided no target benefit words and five target cost words—the (0,5) group. Yet only four subjects provided more than a total of five target words, yielding the triangular-shaped lattice. The results look rather different than the sort of cloudlike bivariate and (roughly) Gaussian distributions that one typically expects to see with a pair of behavioral dependent variables. Instead of a cloud, there is a remarkably triangular distribution with peaks at all three vertices—corresponding to the (5,0), (0,5) and (0,0) points.

Of interest, the priming effect, which was the main point of the original article, was largely driven by a

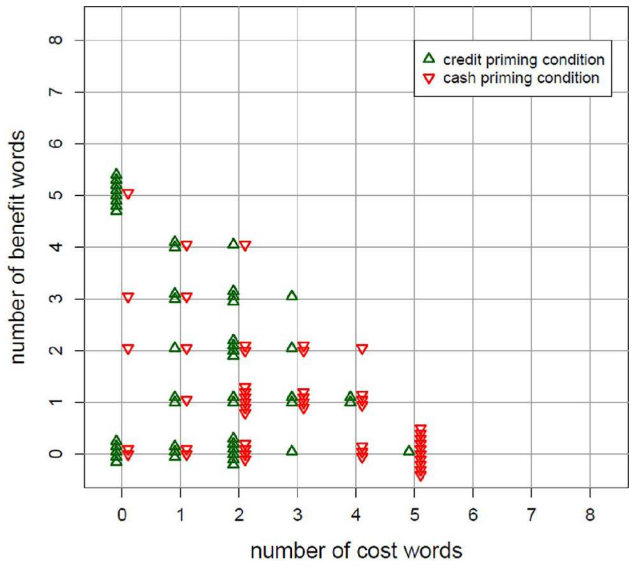

Figure 1. Results of Study 3 (Chatterjee et al., 2013). *Note.* Each triangle represents the data from one subject. Jitter in both coordinates was added as needed to make overlapping data points visible. Each subject completed word stems. Eight of the stems could be completed by words representing a cost of volunteering, and eight could be completed by a word representing a benefit of volunteering. The *x* and *y* axes reflect the subject's score on these two variables, respectively. Red data points represent subjects in the cash condition, and green data points represent subjects in the credit condition.

high concentration of (5,0) subjects in the credit condition (the eight green and one red data points in the upper leftmost node of the lattice) and a high concentration of (0,5) subjects in the cash condition (the 10 red and one green data points in the lower rightmost node).

Surprised by the character of this distribution—that is, numerous datapoints at the three corners of the lattice and the fact that data seem almost entirely confined by an imaginary diagonal line running from (0,5) to (5,0)—we asked the original authors for additional data. Specifically, we asked for the raw data showing each subject's specific word completion choices for each of the 25 word stems. The authors were kind enough to provide these data (R. Rose, personal communication, February 4, 2014). (The reader can inspect the data at http://laplab.ucsd.edu/Chattdata/)

In examining the raw data, we began by focusing on the (5,0) and (0,5) subjects, for two reasons. First, the occurrence of so many values at these extreme points in the distribution had seemed surprising to us, as just mentioned. Second, these points interested us because the concentration of these extreme values in the two respective conditions was largely driving the overall priming effects. In Figure 1, this is seen in the preponderance of green symbols in the upper left-hand corner of the lattice and the preponderance of red symbols in the lower right-hand corner. If one excludes these two clumps of subjects, the priming effect the authors reported would have been nothing more than an insignificant trend.

Exploring the raw data, we quickly noticed something unusual about the nine filler word stem completions. (The filler stems were chosen to be unrelated to anything else in the study and were included, we presume, to draw the subject's attention away from the link between the stems and the Nature Conservancy materials.) Consider first the nine subjects in the (5,0) group. The particular filler word completions chosen by these subjects are all shown in Table 1 (each subject is a row and each filler word is a column). Naturally, because these are filler words, the subjects' choices for completing these words did not play a role in getting them included in the (5,0) group to start with. Nonetheless, we observed what struck us as a strangely high degree of commonality in these subjects' filler responses, especially when compared against the rest of the data. For example, the stem BR_ was completed as BRAIN by eight of the nine subjects in the (5,0) group. But of the other 74 subjects not in either the (5,0) or (0.5) groups, only seven chose BRAIN. A similarly high degree of within-group similarity in filler word completions was shown by the 11 subjects in the (0,5) group. For example, nine of the 11 subjects in that group completed BR___ as BRIBE (a choice favored by only four of the remaining 74). Similarly, 10 of the 11 people in the (5,0) group chose NAP, whereas only 24 out of the remaining 74 chose that particular completion. These are but a few of many such examples.

Do subjects within other "nodes" of Figure 1 also show reduplication?

Thus far, we have discussed similarity of filler word completion choices within the (5,0) and (0,5) groups. We had focused on these groups from the start because they drove the overall priming effects and because the existence of such a large number of subjects in these extreme points of the bivariate distribution (as seen in Figure 1) had seemed strange to us. But is the reduplication of filler word stem completion choices actually specific to these two groups? Is it possible that within *every* node in the lattice shown in Figure 1, subjects exhibit a high degree of similarity in their choice of filler stem completions?

To assess this, we took all possible pairs of different subjects in Experiment 3 and, for each pair, computed

Table 1. Filler word stem completions chosen by the subjects in the (5,0) and (0,5) groups.

CHA	LA	BR	TAB	BO	DU	FO	NA	SPO
(5,0) Group								
CHAIR	LAP	BRIGGLE	TABLE	BOOK	DUE	FOOL	NAP	SPOOK
CHAIR	LAP	BRAIN	TABLE	BOULDER	DUST	FOOT	NAP	SPOOK
CHAIR	LAP	BRAIN	TABLE	BOOK	DUMB	FOND	NAP	SPOOK
CHAIR	LAP	BRAIN	TABLE	BOOK	DUMP	FOOT	NAP	SPOOK
CHAIR	LAP	BRAIN	TABLE	BOOK	DUG	FORD	NATE	SPOOK
CHART	LAP	BRAIN	TABLE	BOOK	DUNK	FOOT	NAP	SPOOK
CHAIR	LAP	BRAIN	TABLE	BOSS	DURATION	FOOT	NAP	SPOOK
CHAIR	LAP	BRAIN	TABLE	BOOK	DULL	FOOT	NAW	SPOOK
CHAIR	LAP	BRAIN	TABLE	BOOK	DUE	FOOT	NAP	SPOOK
(0,5) Group								
CHA	LA	BR	TAB	BO	DU	FO	NA	SPO
CHAIR	LAY	BRIBE	TABLE	BOOK	DUNE	FOOT	NAP	SPOOK
CHAIR	LAVA	BRIBE	TABLE	BOOK	DUCK	FOOT	NAP	SPOOK
CHAP	LAKE	BROWN	TABLE	BOG	DUKE	FOG	NAG	SPOKEN
CHAIR	LAVA	BRIDE	TABLE	BOAT	DUC	FOND	NAP	SPOOK
CHAIR	LAVA	BRIBE	TABLE	BOOK	DUD	FOOT	NAP	SPOOK
CHAIR	LAVA	BRIBE	TABLE	BOOK	DUMB	FOOT	NAP	SPOOK
CHAIR	LAVA	BRIBE	TABLE	BOOK	DUMB	FOOT	NAP	SPOOK
CHAIR	LAY	BRIBE	TABLE	BOSS	DUMB	FOOT	NAP	SPOOK
CHAIR	LAVA	BRIBE	TABLE	BOOK	DUNK	FOOL	NAP	SPOOK
CHAT	LAMP	BRING	TABLE	BOAT	DUE	FORT	NAP	SPOT
CHAIR	LAVA	BRIBE	TABBY	BOOK	DUMB	FOOL	NAP	SPOOK

Note. The stem is shown on the first line of each set, followed by the complete set of subject responses within the group.

the city-block "distance" between their filler word choices—defined as the number of filler word completions they differed on (yielding a number between 0 and 9). For example, if two subjects produced the same filler words for three of the nine filler word stems, the distance between them would be six. This distance was calculated for each of the 4,371 possible pairings of different subjects. (There are 4,371 because there are (94*93/2) combinations of 94 things taken two at a time.) Figure 2 shows four histograms showing the resulting frequency distribution for four distinct subsets of the population of subject pairings.

Figure 2, top panel, shows the first subset: the 91 cases where both subjects lie within the (0,5) group *or* both subjects lie within the (5,0) group (i.e., this shows the two groups just discussed, representing 36 and 55 pairings, respectively). As expected given the reduplication just discussed, these distances tended to be extremely low, with a mode of 2, showing a high frequency of overlapping word choices. Figure 2 (second panel) shows the 164 cases where both subjects are within some *other* single node in the Figure 1 lattice—that is, a node other than (5,0) or (0,5). The distribution of filler word distances is shifted very far to the right with respect to the top panel, and it has a mode of 7. Comparing the top two panels, one can see that the extraordinary similarity in the filler word stem completions found within the (5,0) and (0,5) groups is *not* a general property of the nodes displayed in Figure 1. It is primarily a property of the two groups of subjects who were extreme on the main dependent variable and whose data drove the primary finding of the original article.

Figure 2 (third panel) shows the 4,017 pairs where the two subjects within the pair are not in the same node at all—but after excluding all subjects in either the (5,0) group or the (0,5) group. This distribution for internode pairings looks very similar to the distribution in the second panel, suggesting that leaving out the (5,0) and (0,5) groups, being in the same node of the lattice (i.e., having the same number of benefit and cost target words) or not being in the same node *does not generally make any difference for the similarity of filler word completions.* From a commonsense standpoint, what is seen in the second and third panels is completely unsurprising, of course, because the filler words were chosen by design to have no obvious semantic or other relationship to the target words.

Another rather interesting result of this analysis is seen in the bottom panel of Figure 2. This shows the 99 subject pairs where one subject is in the (5,0) group and the other is in the (0,5) group. Oddly, these 99 subjects tend to show a strikingly *smaller* distance than the 4,017 subjects in the final group (bottom panel of Figure 2), which shows between-group pairs.

If one attributed the overall similarity of filler completions within the (0,5) and (5,0) groups to the notion that people who tend to pick similar targets words also choose similar filler word completions, then the high similarity between (0,5) and (5,0) people's choices is the opposite of what we would have expected. The (5,0) and the (0,5) groups are maximally different in their target word choices—indeed, they chose zero target words in common. But rather than being

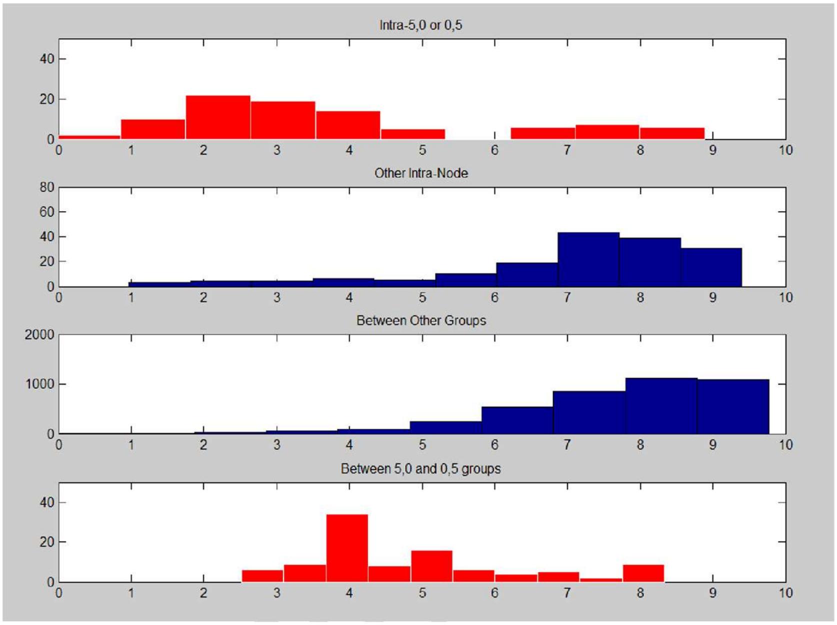

Figure 2. Histograms showing the frequency of different filler word distance values within four different subsets of distinct subject pairs (higher values on the *x* axis mean more differences in choice of filler word stem completions). *Note.* Red bars show data involving (5,0) and (0,5) groups, whereas blue bars show data not involving these groups. Top panel: Pairings within either the (5,0) or (0,5) groups shown in Figure 1. Second panel: Pairings of subjects from within any of the other "nodes" shown in Figure 1. Third panel: Pairings in which one subject is in one of the "nodes" besides the (5,0) and (0,5) nodes and the other subject is in another of the nodes besides the (5,0) and (0,5) nodes. Bottom panel: Pairings consisting of one (5,0) subject and one (0,5) subject. The results show that the small distance within the (0,5) and (5,0) groups is quite different from what is seen within or across other groups of subject.

dissimilar in their choices of filler words, they are unusually similar.

How big is the reduplication effect?

From casual inspection, the reduplication effect involving filler words (Table 1) struck us as very large in magnitude. To see whether this is the case, we need a measure of the size of the effect. Effect sizes for differences in frequency values are usually represented as odds ratios (this statistic represents the difference between two probabilities p1 and p2 as $((p1/(1-p1)/(p2/(1-p2)))$. In a recent publication, Chen, Cohen, and Chen (2010) suggested that odds ratios of 1.7, 3.5, and 6.7 are roughly comparable to Cohen's well-known guidelines for "small," "medium," and "large" effects in the realm of Cohen's *d*.

By that metric, is the tendency of (5,0) and (0,5) subjects to pick the same modal word stem completion for filler words really a large effect? For each of the nine filler words, we computed the odds of the (5,0) group producing the modal response for the group, and we computed the same quantity for the (0,5) group. The average of these 18 values equaled 4.87 (in three cases, the observed odds were 9:0, making the odds undefined; we replaced these undefined values with 8:1, which is a conservative decision). We then computed the nine odds values for the other subjects not in either group; the average of these values was 0.45. Dividing the former by the latter yields an overall odds ratio of 10.92.

This measure of the tendency of the (5,0) and (0,5) groups to produce the same response to the filler words (odds ratio = 10.92) is far in excess of the boundary for what Chen et al. (2010) would view as a "large" effect.

To put this value in perspective, we thought it might be helpful to compare the 10.92 odds ratio with the effect size (again measured with odds ratio) for a basic and commonsense priming effect: the tendency for the reading of a word to prime the production of the *very same word* as a stem completion. So, for example, if someone has read the word *quinine*, are they more inclined than they would otherwise be to complete the stem QUI_ as QUININE? Indeed they are. And how big is *that* effect? To answer this question, we picked four highly cited word-stem completion priming articles from the cognitive psychology literature.

Rajaram and Roediger (1993, Table 1, p. 769) reported that when subjects had read the prime word, 44% of them completed the word stem with the primed word, as compared to 30% for unstudied words. This yields an odds ratio of 1.17. Roediger and Blaxton (1987, Table 1, p. 382) reported a similar study, where 51% of subjects completed the word fragment with the primed word, as compared to 27% for unprimed (using the data for typing condition in the study).

The odds ratio here was 1.26. MacLeod (1989, Table 1, p. 400) reported a similar study, where 34% of subjects completed the word fragment with the primed word, as compared to 20% for unstudied (data from “crossed out” condition). The odds ratio here was 1.40. (Of interest, the article also showed that the priming effect was greatly reduced when the primes were read in the context of a sentence, suggesting the fragility of word fragment completion priming.) MacLeod and Kampe (1996) reported a similar study, where 42% of subjects completed the word fragment with the primed word, as compared to 20% for unstudied (averaging over word frequency). The odds ratio here was 1.55. The basic priming effect sizes in these articles seem quite consistent, and this most obvious of word-stem priming effects is actually a very modest effect.

To summarize, then, within each of the two groups that drove the reported priming effects of Chatterjee et al. (2013), the tendency to pick the same *filler* word responses as other subjects in the same group is about 5 to 10 times as large as the most basic and commonsense priming effect one can find in the word-stem priming literature. Or to put it in simpler terms, reading a word increases the odds of completing a fragment with the same word only by about 20% to 50%, but in the raw data provided by Chatterjee et al. (2013), being a subject in the key (5,0) or (0,5) groups purportedly raised the odds of completing filler words in the common fashion by a factor of approximately 10.

Is reduplication due to sampling error?

The results thus far can be summarized simply: The data from the two groups of subjects who were driving the key priming effects reported in the original article, namely, the (5,0) and the (0,5) “nodes” of Figure 1, show a powerful resemblance to each other in what one would have expected to be the completely unrelated word-completion choices these people made. This cannot be explained in terms of any simple selection bias: Their choice of filler words nontarget words is not what caused these subjects to be included in the (5,0) or (0,5) groups in the first place. Moreover, we have seen that (a) pairs of subjects inhabiting other particular “nodes” in Figure 1 do not show any notable resemblance to each other in their choice of filler word completions, and (b) the tendency for (0,5) and (5,0) subjects to choose the same filler word completions is actually a far stronger effect (in terms of odds ratio) than the obvious priming effect whereby a person who reads, for example, the word *quinine* tends to complete the word stem QUI as *quinine.* Even more odd, we have seen that the (0,5) group shows a moderate resemblance to the (5,0) group in terms of filler word completions, despite their maximal dissimilarity in likelihood of choosing the two types of target words.

Could all of this reduplication be attributed to chance? As a preliminary test of the statistical significance of the reduplication, a simple and crude resampling test was carried out. We tested a null hypothesis stating that these subjects were not systematically different from the other subjects in the study in their propensity to pick particular filler stem completions and that every subject chose words independently of what other subjects did, and independently of their other word choices. Based upon those (admittedly rather simplified) assumptions, the chance of this reduplication occurring by chance would be much less than 1 in 100,000 for the filler words for each group. Given the nulls just described, the total pattern for both nodes would have been expected to occur less than one time in 10^{10}. In our judgment, the infinitesimal value here would overwhelm any legitimate concerns one might have about the arguably post hoc nature of the observation being measured.

The assumptions made for this test are admittedly simplistic, however. One can imagine how the choice of a completion for one filler word stem might not be completely stochastically independent of the completion of another filler word stem. Perhaps people who happen to be inspired to complete TA as TABLE might also, for some quite inscrutable (but nonetheless natural) reasons, be prone to complete BR as BRAIN, and so forth. Given that the filler words have no obvious semantic or other relationship to each other or to the cost and benefit target words, the mechanism for such a statistical dependency seems hard to imagine. Nonetheless, it is always possible that this could happen due to some not yet understood psychological process.

This led us to create a more refined and conservative test for the idea that the reduplication far exceeds anything to be expected based upon the methods described in the article, we performed a second set of item-by-item permutation tests for each of a subset of the reduplicated filler words. The detailed methods used there are described in Appendix D. Figures 4 and 5 show the results of the resampling test for the extreme node with five benefit words but no cost words (5,0) group ($n = 9$) and the (0,5) group ($n = 11$).

As seen in the figures, even with this conservative test, seven of the eight resampling tests showed a dramatic deviation from expectation in the filler word completions. This would imply that for each of these words, whatever slight differences exist between the key groups and their immediate neighbors in the lattice shown in Figure 1 cannot account for the homogeneity in filler word completions chosen by this group.

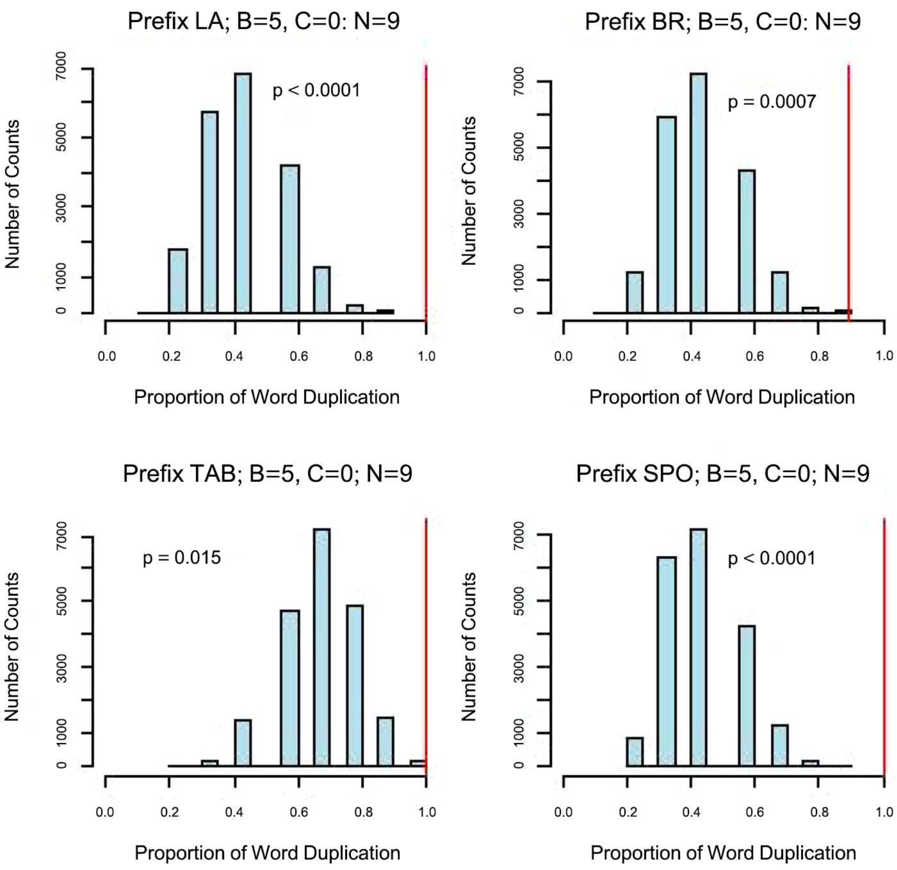

Figure 3. Resampling test for the filler completions chosen by the (5,0) group. *Note.* Each histogram shows the permutation distribution of the proportion of maximum word duplication for a specific stem in the extreme node. The *y* axis shows the number of counts (in each bin, out of 20,000 total). The red line marks the original level of duplication in the (5,0) group. A two-tailed permutation *p* value is shown on each plot. B = number of benefit words; C = number of cost words; N = number of subjects' data within the critical cell.

Something made just the people within the particular (0,5) and (5,0) nodes make extremely similar choices in how they complete each of these task-unrelated filler word fragments. The conclusion of these various analyses can be simply summarized: The reduplication of responses within the (5,0) and (0,5) groups goes very far beyond what can reasonably be attributed to chance.

Does the reduplication have a natural causal explanation?

In responding to an earlier version of the current article, a review submitted by one or more of the original authors argued that even if the reduplication was far too great to be attributed to sampling error, our analysis unreasonably underestimated the likelihood that there might be a causal explanation for the reduplication. After all, they pointed out, the subjects in the (0,5) and (5,0) groups are not a random subset of subjects. These are the subjects who showed the greatest amount of the reported priming effect. The authors argued that

> an alternative explanation is that the primes affected these extreme participants in a similar way. If the participants' susceptibility to money primes is high (as is evident from the responses), because money is likely associated with a rich network of concepts, it is plausible that they could respond by accessing a similar constellation of words, even words that are not directly related to money, at first glance. (Anonymous reviewer)

Fortunately, the existence of a large literature on various kinds of priming allows us to assess empirically whether it is indeed true that the people most highly primed by a given theme tend to produce filler stem completions that overlap with the filler stem completions produced by others who are also highly primed by the same theme. We were able to locate an article by Kemps, Tiggemann, and Hollitt (2014) in which 160 female undergraduates had watched a series of TV commercials focused on either food or nonfood products. The subjects then completed word stems. The list of stems included 15 filler word stems that were selected to have no food-related completions, as well as 45 stems that allowed food-related completions.

Eva Kemps (personal communication, May 25, 2015) was kind enough to share her raw data in the form of hand-completed stem completion forms. The filler

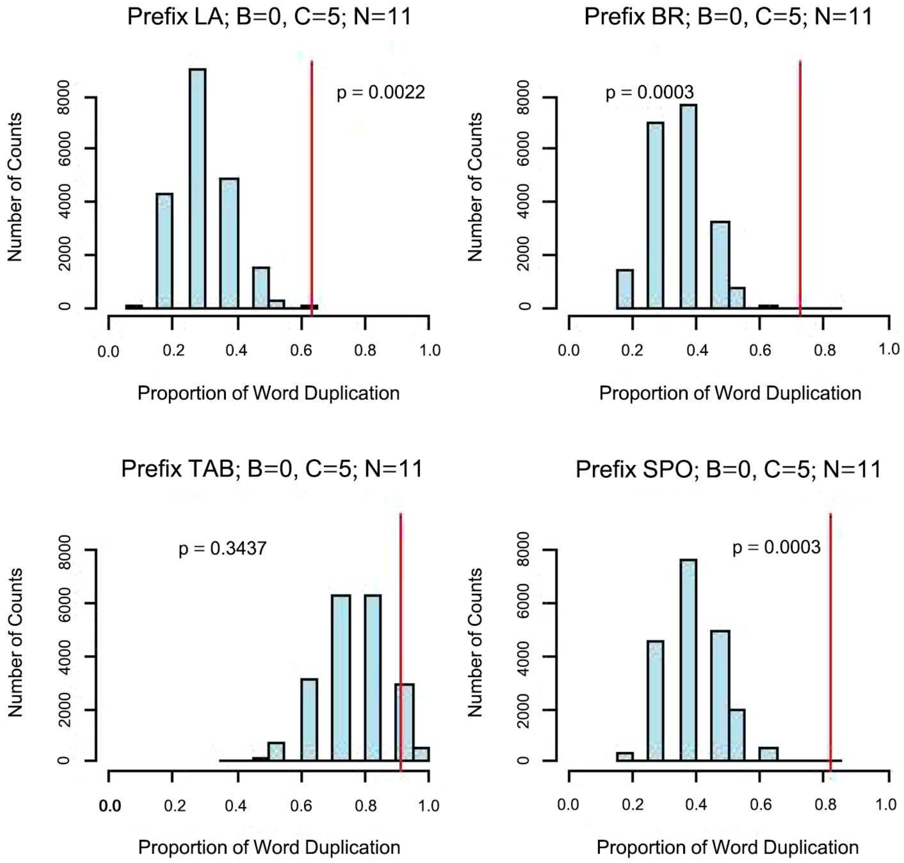

Figure 4. Resampling test for the filler completions chosen by the (0,5) group. *Note.* Each histogram shows the permutation distribution of the proportion of maximum word duplication for a specific prefix in the extreme node. The *y* axis shows the number of counts (in each bin, out of 20,000 total). The red line marks the original level of duplication in the (0,5) group. A two-tailed permutation *p* value is shown on each plot. B = number of benefit words; C = number of cost words; N = number of subjects' data within the critical cell.

responses written on the forms had not been analyzed before. We copied them into a computer file and analyzed them. For each subject, this yielded the following information: the subject's condition (food priming or control), the number of food-related stems that were completed with food items, and their completion responses to the 15 filler stems.

Did the subjects in Kemps et al. (2014) who showed the strongest responses to the food primes tend to produce filler word completions that were similar to other people showing strong food priming? To answer this, we carried out word-by-word resampling analyses mirroring our conservative bootstrap analyses of Chatterjee et al. Experiment 3 described previously. For each filler word, two groups were compared. Group 1 consisted of the nine food-primed subjects who produced the highest number of food responses (eight or more). Group 2 consisted of the 14 "near-neighbor" food-primed subjects for whom the number of food-related responses was almost as high (six or seven). Resampling tests were used to ask whether the responses chosen by the most primable group show a greater resemblance to each other than did the responses in the other group, relative to a resampled distribution. As seen in Figure 5, there was no systematic tendency for these two groups to differ in response reduplication rates. To see the point here, the reader may wish to compare Figure 5 with Figures 3 and 4 (Chatterjee et al. data), a comparison that reveals the power of the reduplication pattern in the Chatterjee et al. data, a pattern quite absent from the Kemps et al. data.

We also compared the filler stem completion response homogeneity of the nine maximally food-primed subjects who had the largest number of food-related completions against that of the 18 *least* food-primed subjects, those who chose only one or two food-related responses. Again, the results looked much like Figure 5.

Another puzzling reduplication: Nontarget wordstem completions

Thus far we have focused on the puzzling similarity of the filler word completions chosen by subjects within the two clusters of subjects that drove the effects

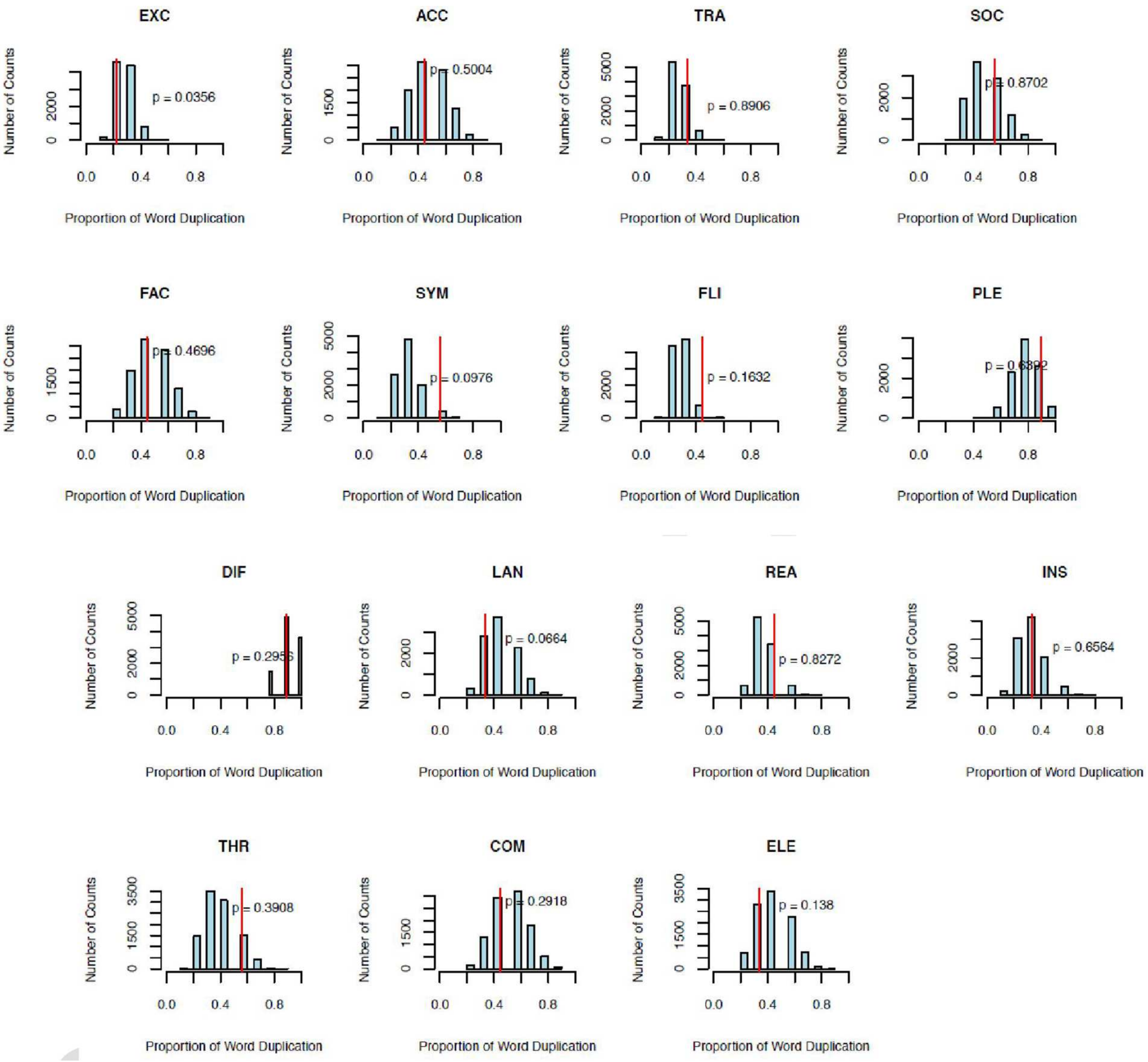

Figure 5. Resampling test for the proportion of word duplication in the filler stem completions in Kemps et al. (2014). *Note.* Each histogram shows the permutation distribution of the proportion of maximum word duplication for a specific stem in the most-food-primed and second-most-food-primed groups. The *y* axis shows the number of counts (in each bin, out of 20,000 total). The red line marks the original level of duplication in the most-food-primed group. A two-tailed permutation *p* value is shown on each plot. The Kemps et al. data set shows no sign of the strange reduplication pattern seen in the results of Chatterjee et al. (2013).

reported in the article. We also looked at another category of words: the *nontarget word completions.* We use this phrase to refer to word stems that could potentially have been completed with cost or benefit words but instead were completed with some other word. (Just to remind the reader, if a subject is in, say, the (0,5) group, this means the subject came up with zero benefit words and five cost words in response to the target stems. Because the subject selected zero benefit targets, this means that for each stem that had a potential benefit word completion, the subject must have put down something else besides the benefit word. So we can ask whether the particular nonbenefit words the subject chose were abnormally similar to the choices provided by the other subjects in the very same (0,5) group.) Sure enough, the raw data again showed an extraordinarily high level of reduplication for the two groups of subjects who drove the authors' reported priming effect. For example, the stem TELE, which was associated with the target word *telephone*, was completed as *telepathy* by eight of the nine subjects (89%) in the (5,0) group but by only three of the 25 (12%) other subjects in the study who produced a nontarget response to this stem. The other subjects produced a very wide range of completions, such as *teleport, television, teller,* and *telemarketer*. Similarly, nine of the nine subjects in the (5,0) condition produced the word TIE in response to the stem TI_. However, only two of the

51 other comparison subjects chose TIE; instead, they produced a long and varied list of words including TIGER, TICK, TIN, TINY, TIED, TIP, and TIRE, among others.

Looking further at the data, one sees that seven of the nine subjects in the (5,0) condition completed PO_ with POKE. But of the 80 other subjects who produced nontarget responses to PO_, only 10 produced POKE; the others made varied choices including POKER, POLE, POLO, POLISH, POND, POEM, PODCAST, and others. Of the (0,5) subjects (i.e., subjects who produced no target benefit words), nine of the 11 subjects produced RECOVERY to the cue RECO_, whereas of the other 83 people who gave a response other than the target (RECOGNITION*)* in response to this stem, only eight produced RECOVERY. A crude resampling test similar to the one just described was performed on these two sets of resemblances, and again the simplest null hypothesis was rejected at $p < 10^{-10}$.

Incompatibility of data with reported methods in Study 3

Thus far we have focused on the similarity of word-stem completions produced by the subset of the subjects who had extreme scores, driving the effects reported in the article. A reviewer of an earlier version of this article (Jelte Wicherts) personally examined the Chatterjee et al. (2013) data set and made several additional observations bearing on the validity of the data from a completely different perspective. The reviewer pointed out that the stem SUPP_ was completed as SURGERY by six subjects. Of course, SURGERY fails to match two of the four letters in SUPP_. As Wicherts pointed out in his review, this pattern could not arise from the data generation process described in the article, even given the tendency of human beings to make errors. Although it seems conceivable (albeit strange) that even one person in a group of 94 students might complete SUPP_ with SURGERY, the idea that six people would produce the same low-frequency word mismatching the stem in two letter positions seems to us patently absurd. (One might hypothesize that perhaps there was an error in the list of stems and that the subjects were actually given SU_ rather than SUPP_. But then one would have to explain how six people would come up with SURGERY when zero came up with SUN, SUGAR, SUIT, SUSHI, SURE, and any of the other frequent completions for SU_.)

Equally strikingly, Professor Wicherts noted, for the stem CE_, 17 of the 94 completions in the file began with CE but 77 began with CA. Again, it is not clear how any data collection process remotely like that described in the article could have resulted in this outcome. Additional mismatches between materials and responses were noted with smaller numbers of cases.

What generated the data?

In our opinion, it is clear that the strangely similar word choices provided by the (5,0) and (0,5) subjects—the data points that basically drove the primary findings reported in the article—could not realistically have been created by the process described in the Method section of Chatterjee et al. (2013). The same is true for the repeated occurrences of specific word choices that fell well short of completing the corresponding stems (e.g., the six occurrences of SURGERY) as noted by reviewer Jelte Wicherts. If there is a reasonable and innocuous explanation of what produced these strange patterns, the original authors have not shared it with us despite repeated attempts to elicit such an understanding.

Naturally, we are not in a position to determine exactly what series of actions and events could have resulted in this pattern of seemingly corrupted data. In our view, given the results just described, possibilities that would need to be considered would include (a) human error, (b) computer error, and (c) deliberate data fabrication. In our opinion based solely on the analyses just described, the findings do seem potentially consistent with the disturbing third possibility: that the data records that contributed most to the priming effect were injected into the data set by means of copy-and-paste steps followed by some alteration of the pasted strings in order to mask the abnormal provenance of these data records that were driving the key effect. Of course, as we have seen, the added noise was, in the end, quite insufficient to cover up the reduplication pattern. In our opinion, it may be an interesting clue about what happened here that, as just noted, the 20 subjects in the (5,0) and (0,5) groups—although they are ostensibly the choices of people maximally *dissimilar* to each other in terms of the key priming variables reported in the study—nonetheless resemble each other markedly in their filler word stem completions. In our opinion, this makes any innocuous causal explanation for the reduplication even less plausible than it would otherwise have been and suggests that all 20 data points may have been copied into the data set from a common source, with slight variations introduced to mask this fact.

The same data-fabrication interpretation might potentially explain how the data set could have come to include words (such as SURGERY) that did not come close to matching the stem SUPP_. The data file we received did not include the stems as column labels. If someone hastily added fabricated words to the data file, they could easily have gotten confused from time to time about what the stem was. Most of the stems consisted of just two letters, and thus someone glancing at a column of data containing entries like SUPPORT and SUPPLIES one might erroneously assume that the stem was SU_. Based on that misimpression, a fabricator might have gone on to pick SURGERY as a completion, pasting this word into a number of cells. As reviewer Wicherts also noted, strange mismatches between stimulus materials and responses in raw data files have proven crucial in past inquiries into scientific irregularities, for example, the data fabrication eventually acknowledged by Diederik Stapel in the Netherlands (explored in Levelt, Drenth, & Noort, 2012).

Although, based on the facts apparent to us, the hypothesis of data fabrication seems possible, we would nonetheless appeal to readers to hold open another possibility: that some or all of the oddities discussed here might have arisen as a consequence of some sort of human or machine error that we currently do not understand.

Statistical issues and broader implications

The analyses described here have generally followed the approach common to conventional "null hypothesis statistical testing," in which the extremity of a test statistic is compared to what would be expected assuming some null hypothesis. When this conditional probability is low enough, the credence given to the null hypothesis is reduced. As many authors have pointed out, there is no rule of logic or probability theory that directly warrants an inference from "low *p* value" to the conclusion that the null hypothesis is false. This has led many researchers to embrace a Bayesian approach, which begins with assumptions about the a priori probability of different hypotheses and then explicitly updates these probabilities in light of the data. Can the statistical inferences just described be recast in Bayesian terms? Strictly speaking, they cannot: A full-fledged Bayesian analysis is impossible, because we lack a basis for firmly specifying prior likelihoods of "the data collection proceeded as described in the original article" or "the data collection seriously deviated from that described in the original article." The conditional probability of the data given each of these hypotheses is also extremely hard to quantify.

However, an informal and qualitative sort of Bayesian reconstruction can be envisioned if one is willing to make rough and intuitive estimates of the various quantities required. What would be reasonable to assume about the prior probability of different hypotheses regarding data integrity? Recent evidence makes it clear that corrupt research practices are unfortunately not nearly as rare as sometimes hoped or imagined. For example, John, Loewenstein, and Prelec (2012) estimated a "surprisingly high" rate of 1.7% for "falsifying data" based on a survey of 2,000 psychologists. Other kinds of errors presumably occur at even higher rates. Thus, in examining data sets with troubling oddities, one should probably assume that although corrupted data collection is uncommon, it is not extremely rare. To employ Bayesian reasoning in the present case, one would also need to estimate the conditional probability of the observed abnormalities arising given each of the two hypotheses (i.e., that the data were generated by the methods described, and that they were not). As noted earlier, if the data were deliberately fabricated, the reduplication of filler word stem completions seems potentially comprehensible, because it might reflect a convenient strategy of goal-directed data alteration (insertion of extreme-valued records into the data set, engineered to produce a desired effect). The conditional probability of these abnormalities appearing given other, more benign, forms of corrupted data collection—for example, human or machine error—are far harder to estimate. Indeed, as just mentioned, we are unable to think of any concrete scenario for how the oddities described in this article could have resulted purely from human or machine error, but of course that does not rule out this possibility.

A reviewer of an earlier version of this article observed that in several recent discussions of possible data falsification, discussion revolved almost entirely around *p* values derived from the null hypothesis of non-corrupted data collection. For example, in one recent case occurring in the Netherlands, the average measurements reported in a published article for the second of three values of an independent variable repeatedly lay extremely close to the arithmetic mean of the measurements for the other two values (van Kolfschooten, 2014). Although investigators did very careful analysis of the conditional probability of this pattern given the methods described in the underlying article (e.g., Klaassen, 2015), to our knowledge there was very little discussion of how and why the data might have exhibited this pattern on the assumption that they were not honestly generated. This was also the case in some other recent findings related to suspect data sets (Simonsohn, 2013). By contrast, with the data set of Chatterjee et al. (2013), one can at least potentially envision how data fabrication might have resulted in the appearance of some of the key oddities discussed here.

For psychology as a whole, the results help to reinforce the growing belief across all fields of science that posting of raw data can have a powerful and useful benefit in promoting the integrity of the scientific process (Simonsohn, 2013; Wicherts & Bakker, 2012; Wicherts, Borsboom, Kats, & Molenaar, 2006). In agreement with these authors, we would argue that routine and obligatory sharing of raw data is likely to dramatically increase incentives for investigators to avoid questionable or corrupt research practices. The demonstration, here and in other recent cases, that a detailed probe of raw data is capable of revealing easily overlooked but converging indicators of serious problems may, we hope, help to deter improper scientific practices in the future.

Conclusions about Chatterjee et al. (2013)

To sum up, the data set underlying Chatterjee et al. (2013) is characterized by a number of disturbing oddities. A close examination of the data files turned up a mysterious pattern of reduplication of four different categories of specific word stem completion choices in Experiment 3. These reduplication patterns—present in the very subset of subjects whose extreme priming scores drove the primary statistical significant effects reported in the article—seem to us to lack any reasonable explanation that would be consistent with the data collection procedures described in the article. Our probe of these strange reduplication effects revealed them to be large in magnitude (e.g., when measured as an odds ratio, they were much stronger than the basic and commonsense sort of priming effect reported in the cognitive literature on word stem completion). The simplest and crudest resampling tests that we performed suggested that this pattern of reduplication might be expected to occur by chance less than one time in 100 million million million. On the other hand, more refined and conservative resampling tests suggested (through their recurring and dramatic rejection of the null hypothesis; see Figures 3 and 4) lower but still very extreme levels of statistical significance. We were unable to find any innocuous explanation for these patterns, which recurred within four orthogonal subsets of the data, all involving the subsets of subjects who did the most to contribute to the effects reported. Several features of these analyses seem particularly problematic for innocuous explanations for the reduplication, namely, (a) the fact that the extreme subsets actually resembled each other in the filler word completions, (b) the fact that other subsets of the data did not generally show much in the way of excessive reduplication, and (c) the fact that the effect size for the reduplication (odds ratio) far exceeded previous observations of the size of the most obvious form of priming (e.g., the effect wherein reading QUININE makes a person more likely to complete QUI___ as *quinine*). The examination of the raw data from Kemps et al. (2014) confirm, as we expected, that there is no general psychological force that leads highly primable people to make very similar choices in their completion of word stems unrelated to the priming intervention.

Another very troubling aspect of the data set was the occurrence (brought to light by a reviewer of an earlier version of this article) of many occurrences of specific word-stem completion responses that did not actually complete the stem or even come close to doing so (particularly the multiple instances of SURGERY being offered by subjects to complete the stem SUPP_). These observations also seem hard to square with any data collection procedure resembling the description provided by Chatterjee et al. (2013) but seem potentially consistent with data fabrication or perhaps some form of error.

Of course, it is not our job to draw firm conclusions about how the data files came to exhibit all of these odd features and what the consequences of that discovery should be. If the data were generated in the way described in the article, the co-occurrence of all these effects would be extraordinarily unlikely, in our opinion, based on the calculations we report. Given the strong and unexplained oddities enumerated here, we would suggest that the results of Chatterjee et al. (2013) should be assumed to lack scientific validity.

Acknowledgments

We thank Jelte Wicherts, Justin A. MacDonald, and several anonymous reviewers of earlier versions of this article.

Funding

This research was supported a collaborative activity grant from the James S. McDonnell Foundation and by the National Science Foundation (Grant SBE-582 0542013 to the University of California, San Diego, Temporal Dynamics of Learning Center), and by a grant from the Institute of Education Sciences, U.S. Department of Education (Grant R305A110517) to the University of South Florida. The opinions expressed are those of the authors and do not represent views of the funding organizations.

References

Chatterjee, P., Rose, R. L., & Sinha, J. (2013). Why money meanings matter in decisions to donate time and money. *Marketing Letters*, *21*(2), 1–10.

Chen, H., Cohen, P., & Chen, S. (2010). How big is a big odds ratio? Interpreting the magnitudes of odds ratios in epidemiological studies. *Communications in Statistics—Simulation and Computation*, *39*, 860–864. doi:10.1080/03610911003650383

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*, 524–532. doi:10.1037/e632032012-001.0956797611430953.

Kemps, E., Tiggemann, M., & Hollitt, S. (2014). Exposure to television food advertising primes food-related cognitions and triggers motivation to eat. *Psychology & Health*, *29*, 1192–1205. doi:10.1080/08870446.2014.918267

Klaassen, C. A. J. (2015). *Evidential value in ANOVA-regression results in scientific integrity studies*, 1–12. arXiv:1405.4540 [stat.ME]

Levelt, W. J., Drenth, P. J. D., & Noort, E. (2012). *Flawed science: The fraudulent research practices of social psychologist Diederik Stapel*. Retrieved from https://www.commissielevelt.nl/wp-content/uploads_per_blog/commissielevelt/2013/01/finalreportLevelt1.pdf

Lobbestael, J., Arntz, A., & Wiers, R. W. (2008). How to push someone's buttons: A comparison of four anger-induction methods. *Cognition & Emotion*, *22*, 353–373. doi:10.1080/02699930701438285

MacLeod, C. M. (1989). Word context during initial exposure influences degree of priming in word fragment completion. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(3), 398–406. doi:10.1037//0278-7393.15.3.398

MacLeod, C. M., & Kampe, K. E. (1996). Word frequency effects on recall, recognition, and word fragment completion tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 132–142. doi:10.1037//0278-7393.22.1.132

Rajaram, S., & Roediger, H. L. (1993). Direct comparison of four implicit memory tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 765–776. doi:10.1037//0278-7393.19.4.765

Richard, F. D., Bond, C. F., Jr., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, *7*, 331–363. doi:10.1037/1089-2680.7.4.331

Roediger, H. L., & Blaxton, T. A. (1987). Effects of varying modality, surface features, and retention interval on priming in word-fragment completion. *Memory & Cognition*, *15*(5), 379–388.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2013, January). *Life after P-Hacking*. Paper presented at the meeting of the Society for Personality and Social Psychology, New Orleans, LA. http://dx.doi.org/10.2139/ssrn.2205186. Available at SSRN: http://ssrn.com/abstract=2205186

Simonsohn, U. (2013). Just post it: The lesson from two cases of fabricated data detected by statistics alone. *Psychological Science*, *24*, 1875–1888. doi:10.1177/0956797613480366

van Kolfschooten, F. (2014). Fresh misconduct charges hit Dutch social psychology. *Science*, *344*, 566–567. doi:10.1126/science.344.6184.566

Westfall, J. (2015, June 16). Don't fight the power (analysis) [Web log post]. Retrieved from www.jakewestfall.org

Wicherts, J. M., & Bakker, M. (2012). Publish (your data) or (let the data) perish! Why not publish your data too? *Intelligence*, *40*, 73–76. doi:10.1016/j.intell.2012.01.004

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, *61*, 726–728. doi:10.1037/0003-066x.61.7.726

Appendix A: Stimulus materials for Study 3

Review of volunteer work at the Nature Conservancy

PROS	CONS
Personal:	Monetary:
Self-esteem and personal development. Increased skills. Recognition (community, awards, certificates)	Travel to/from worksite. Out-of-pocket expenses not covered by host, such as telephone calls related to volunteer duties, supplies, food etc.
Community:	Personal Time:
Increased services (meals served, trees planted, trash cleaned up). Social opportunities (meet new people), Chance to pay back community	Less time with family and for personal needs

Appendix B: Word stems used in Study 3

Word completion study

Please complete the unfinished words. Try and write down the first word that comes to your mind. For example:

LO___ can be completed as LOFT or LOTUS any such word. However if LOFT came to your mind first, please write LOFT. There is no limitation on the length of the word, however, the word should be a meaningful word.

1. Tra____
2. Cha___
3. Es____
4. Ca____
5. Awa___
6. Ski___
7. Op_____
8. Tele____
9. Ex____
10. La___
11. De___
12. Br___
13. Ce___
14. Ta____
15. Tr_____
16. Supp___
17. Ti___
18. Fo____
19. Bo___

20. Du____
21. Po____
22. Fo____
23. Na___
24. Spo___
25. Reco____

Appendix C: Apparent errors in the execution of Study 4

1. In the data we received from the first author, Stems W18 and W22 are both FO. Yet Chatterjee et al. scored the data so that W18 is a stem that can be completed by a cost word, whereas W22 is a filler stem. Thus, when subjects chose "Food" for W22 it was not counted as a cost word.
2. Word 1 has prefix TRA (expecting "Travel", a cost word) and Word 15 has prefix TR (expecting "Trees," a benefit word). When subjects chose "Travel" for W15 it was not counted as a cost word.
3. The design of the study is such that all benefit words appear on the pros side of the Nature Conservancy statement (e.g., trees, certificates) and all cost words appear on the cons side of that statement (e.g., pocket, food). However, other words on pro or cons side of the Nature Conservancy match prefixes, including filler prefixes, and by logic of this study should also be counted as cost or benefit words. These are
   a. Nature (Nature Conservancy), matches filler prefix NA. Should be a benefit word.
   b. Chance (pro side, "chance to pay back community"), matches filler prefix CHA.

   Should be a benefit word.

   c. Trash (pro side, "planting trees, picking up trash"), matches both cost prefix TRA and benefit prefix TR. Should be a benefit word.
   d. Duty (cons side, "[…] expenses related to […] volunteer duties"), matches filler prefix DU. Should be a cost word. In fact the manuscript mentions that duty is a cost word, but it is not marked or counted as such in the the study data.

   All of these were choices made by some subjects.
4. Word 13 has prefix CE (expecting "Certificates", a benefit word). However, 77 out of 94 subjects entered a word that started with CA (Cat, Car, etc.).
5. Furthermore, some subjects entered "Call," a cost word, for W13. It was not counted as a cost word.
6. In the Nature Conservancy statement the word "Food" is offered on the cost side but the word "Meals" is offered on the benefit side, with no clear conceptual distinction between the two.
7. There are also separate priming materials (scrambling sentences), designed to make the two groups think about their respective conditions (cash or credit). Here we have:
   a. Both groups are additionally primed to the cost word "Travel." Scrambling sentence: "travel debt/necessary I to want."
   b. Cash group is additionally primed to the cost word "Food." Scrambling sentence: "he observes often people food."
   c. Both groups are additionally primed to the cost word "Food" in another scrambling sentence: "good she likes blue food."
   d. Both groups are additionally primed to the cost word "Time." Scrambling sentence: "ball the throw time high."
   e. "Money" was given as a priming word both to the cash and credit groups. Cash scrambled sentence: "money you to luck good." Credit scrambled sentence: "is money this fun good." The same word was apparently supposed to prime one group for credit and the other group for cash.

Appendix D: Details of resampling analysis of Experiment 3 data (shown in Figures 3 and 4)

The permutation procedure worked as follows:

1. Only responses to filler prefixes LA, BR, TAB, and SPO were used in this test. These were the stems not matched by any words in priming or stimulus materials, to make the analysis maximally conservative.
2. The local smoothing or near-neighbor method was used for this test. The analysis included only subjects at the two extreme nodes (five benefit words and 0 cost words, or 0 benefit words and five cost words), as well as their "near neighbors": subjects at most two "steps" away, including diagonally (i.e., (3,0), (3,1) and (3,2) are all included). The rationale is that subjects in each of the extreme nodes should be similar to their near neighbors and should show similar results in terms of word selection and duplication. Thus, not only are we computing probabilities using the same subjects who might potentially be special or highly primable in some way, but we are limiting them to the choices of subjects who differed little in their relative number of cost and benefit target completions.
3. The test was then conducted as follows:
   a. For each of the two corners (extreme node and its near neighbors), the four filler words were randomly reassigned between subjects. Reassignment was done by subject rather than by word (i.e., all

of one subject's words were reassigned to another subject in that corner).

b. Proportion of maximum duplication for each of the filler was computed for the subjects who formed the original extreme nodes. Proportion of maximum duplication is defined as the number of occurrences of the word most often repeated divided by the total number of words (nine for the (5,0) node and 11 for the (0,5) node).

c. This was repeated 20,000 times.

d. The original proportion of maximum word duplication of each of the four prefixes was compared with the permutation distribution of maximum word duplication for each node.

Figures 3 and 4 show the results of the resampling test for the extreme node with five benefit words but no cost words (5,0) group ($n = 9$) and the (0,5) group ($n = 11$).