BASIC AND APPLIED SOCIAL PSYCHOLOGY
2016, VOL. 38, NO. 1, 41–46
http://dx.doi.org/10.1080/01973533.2016.1138863

Routledge
Taylor & Francis Group

Response to Comments by Chatterjee, Rose, and Sinha

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

Downloaded by [University of South Florida] at 03:40 04 February 2016

In our reanalysis of Chatterjee, Rose, and Sinha (2012), which appears earlier in this issue (Pashler, Rohrer, Abramson, Wolfson, & Harris, 2016/this issue), we brought to light two strange and (in our opinion) disturbing features of the data set from Study 3 (involving a task in which subjects are shown a word stem and provide the first word that comes to mind beginning with this stem). These peculiar features are as follows:

Reduplication effect: The 20 subjects who principally drove the main reported findings of the study showed an extraordinary level of similarity to each other in their choice of specific word-stem completions. This was true even for the nine filler word stems that were selected by the authors to have no apparent connection to the concepts and manipulations involved in the study. The same extreme level of reduplication was also present in these same subjects' responses to another completely separate set of stems (the eight "nontarget stems"—see our original article for details).

Mismatch effect: The data also featured some strange recurring word choices that did not even come close to fitting with the word stems reportedly used as stimuli in the study. For example, in response to the word stem SUPP__, six of 94 subjects gave the same (erroneous) completion SURGERY.

In our opinion, these two findings show that the published data are not real, raising very troubling questions about how the datafiles were created.

Background

Our interest in the findings in Chatterjee et al. (2012) began in May 2013 when we were sent a copy of the article by another investigator (Kathleen Vohs) who has pursued the topic of "money priming." Our curiosity was piqued by what we thought to be impressively large effect sizes in some of the studies of Chatterjee et al., and we therefore wrote to Chatterjee requesting the raw data from the paper. Some data were provided to us that summer. Examining these data in detail, we came to feel that the data provided for all three studies seemed strange in a variety of ways. Hoping to understand the situation better, that fall we requested more detailed data for Study 3 (viz., the subjects' actual word completion responses), which the authors were kind enough to provide to us. In February 2014 we brought the Reduplication Effect to the attention of the second author of the original article to see if he could shed any light on the massive and (in our view) strange overlap in word stem completion choices for the two key groups of subjects. During the rest of 2014 and early 2015, we pointed out to Rose some of the additional oddities in the Chatterjee et al. (2012) data in e-mail discussions. At no point over this period did Rose or any of his authors concede that there was anything wrong with their data or with the data collection and analysis procedures they had used.

Feeling that the oddities still lacked any sensible explanation, we wrote up a full description of what we had found and submitted the results for publication to this journal. The three independent reviewers all agreed that the level of reduplication was suggestive of corrupted data. (A fourth much more negative review was provided by one or more of the authors of Chatterjee et al.) We were invited to submit a revised manuscript. We did so, and this was then reviewed by an additional reviewer, who also seemed to find the Reduplication Effect troubling.

Shortly after our paper was accepted for publication, we learned that all of the original authors had apparently decided amongst themselves that Study 3 should be "retracted." As far as we know, they have not explained precisely what that means or exactly why they wish this partial retraction to take place, beyond referring to alleged "coding errors." (No retraction is yet reported on the website of the journal, and the editor of the journal has told us that no final decision has been

CONTACT Harold Pashler hpashler@ucsd.edu Department of Psychology, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093.
© 2016 Taylor & Francis

made; J. Steckel, personal communication, December 22, 2015.)

From the authors' commentaries on our paper, it seemed to us that two of the three authors (Rose and Sinha) wish it to be known that they had no personal involvement in the data analysis. Sinha (2016/this issue) stated that the first author (Chatterjee) was exclusively responsible for "data merging," data coding, and data analysis. Rose (2016/this issue) goes further to say that he had no involvement in either data collection or data analysis. (A footnote in the original paper had merely stated that "all authors contributed equally.")

Nonetheless, all three authors continue to contest our analyses in different ways. Their commentaries provide voluminous argumentation along with some simulations, as well as quotes from a statistical consultant colleague whose name is not mentioned in their commentaries. In length and level of detail, their responses are impressive. However, as far as we can tell, their arguments are mostly focused upon disputing relatively peripheral and inessential points in our original article (such as exactly how unusual the effect sizes in the studies are with respect to the literature and to underlying reality). But as to the two most crucial questions—*What caused the Reduplication Effect in their data set?* and *What caused the Mismatch Effect?*—we still do not see any comprehensible explanations that make sense in our eyes to explain the worrisome features of their published data.

Reduplication effect

To remind the reader, the Reduplication Effect noted by Pashler et al. (2016/this issue) consisted of an extraordinarily high level of similarity between the word-stem completions provided by two particular groups of subjects in Chatterjee et al. (2012). These groups, which we refer to as the "(5,0) subjects" and the "(0,5) subjects" (see our original article for details of exactly what these phrases mean), are the very subjects whose extreme scores on the two dependent variables did the most to drive the reported effects. In their responses to our article, the authors offer several arguments to contend that nothing is amiss.

First, they argue that even if the subjects with the duplicated filler responses were eliminated from the data set, there would still have been a trend toward the observed interaction in the remaining data. Even assuming the complete integrity of the remaining data (an assumption we would certainly not wish to rely upon), the presence of a weak trend in that data (they describe one effect as $p = .109$) would still have fallen well short of common journal standards for statistical significance. Had the authors submitted a manuscript reporting a trend with $p = .109$, we doubt that *Marketing Letters* reviewers would have recommended publication. Moreover, whereas Chatterjee et al. (2012) reported significant priming effects on both benefit words and cost words, without the two extreme subject groups in question we do not even see as much as a weak trend for benefit words ($p = .61$ by our calculation). Thus, we do not see how the authors' arguments on this point could rebut the most troubling potential scenarios for how and why the reduplication might have come to be present in the data.

Second, Chatterjee (2016/this issue) suggests that the reduplication might be a natural result of the operation of hitherto unknown psychological forces. After all, he argues, the people in the two extreme groups—the (0,5) and (5,0) subjects—were selected for having shown an extremely high level of priming. According to Chatterjee, this makes it "plausible that they could respond by accessing a similar constellation of words, even words that are not directly related to money" (p. 23). He adds, "… no one really knows the extent to which strong primes such as 'money' can activate distantly related concepts" (p. 23).

One problem here is that these subjects' filler and nontarget word choices lack any discernible commonalities with each other or connections with the concepts the authors hypothesized to be activated in the study (at least, we cannot see any commonalities, and the authors have not proposed any). So one would need to suppose that people are being drawn to a set of seemingly ineffable "strange attractors" in semantic space.

Most notable about the Reduplication Effect—as we discussed at some length in our article—is its remarkable *strength*. The authors' commentaries say little about this. It seems fanciful enough to imagine that a strange ineffable attractor force is drawing out common responses from the key subjects in the Chatterjee et al. (2012) Study 3 data, but to account for the results, one would have to further suppose that this force is more powerful than the sort of intuitively obvious and direct sorts of priming that are verified throughout the implicit memory literature of cognitive psychology (e.g., the tendency for people who have read the word QUININE to later complete the stem QU_ as QUININE). Doing our best to take this suggestion seriously, we even recoded a very large and credible data set (Kemps, Tiggemann, & Hollitt, 2014) to see if highly primable people might show *any* commonality in their choice of filler word-stem completions; the results indicated no evidence that they do.

Another problematic assumption underlying Chatterjee's (2016/this issue) argument is the idea that the people in the (5,0) and (0,5) groups could really

Downloaded by [University of South Florida] at 03:40 04 February 2016

Downloaded by [University of South Florida] at 03:40 04 February 2016

be starkly different than people in neighboring cells in the lattice shown in Figure 1 of our article (these neighbors were used as points of comparison in our resampling tests). A good reason to doubt this is that the dependent variables reflect the outcome of a binomial sampling process. It is true that on expectation, people who score 5 on the number of cost words produced are presumably slightly different in their average propensity for producing those words than are people who scored 4. However, the distributions of the two groups on the hypothesized latent variable of primability would have to be highly overlapping due to the stochastic nature of the process that put them in that node to start with. (By analogy, basketball players who score on five of 11 free throws cannot be radically better at the task than those who score four of 11, because luck and not just skill will have played such a big role in determining number of successes when there are just 11 total throws.) Yet our resampling tests showed that the reduplication in the (0,5) and (5,0) groups was far in excess of expectations even when the resampling was based just on the words produced by the subjects from nearest neighbor points within the lattice. In short, by every measure we can see, the reduplication is far too extreme to take seriously as a newly discovered natural causal mechanism. (If a psychological "strange attractor" effect for highly primable people really exists, this fact would represent a fascinating psychological discovery more momentous than any other findings that have emerged from priming studies. If anyone is inclined to believe this idea, that person should easily be able to confirm it with very simple additional studies, as it would have to be a very powerful effect indeed to have produced the reduplication evident in their data.)

Mismatch effect

The second oddity we described in our article was a mismatch between word-stem stimuli and the responses reported in the data files (an oddity uncovered by a reviewer of our paper, Professor Jelte Wicherts). According to the data set provided to us, six of 94 subjects prompted with the word stem SUPP___ completed this stem with SURGERY. Chatterjee (2016/this issue) offers seven possible factors that he says might explain these deviant patterns, such as "interaction of mood with an experimental manipulation," possible inequalities in the "ease of generating words from the word stems," and subjects' possible difficulties with English (p. 25).

Studying Chatterjee's list of seven suggestions, we cannot understand how any of them could really offer an explanation for the particular oddities observed. One suggestion that seems remotely promising is that "students may have just focused on the first letter of the stem" (p. 25; Sinha, 2016/this issue, makes a very similar suggestion about subjects possibly focusing "on only the first letter of the word stem," p. 38). But the problem here is that the six subjects in question did not choose a variety of words beginning with S, for which they would have had a vast number of options (SUGAR, SEX, SADNESS, etc.). But instead they all chose SURGERY! Not one of the other 88 subjects in the study chose any of the very long list of other words beginning with S that do not begin with SUPP. We are unable to think of any reasonable explanation for this pattern consistent with what is stated in the Method section of Chatterjee et al. (2016/this issue).

Another one of Chatterjee's (2016/this issue) suggestions is that subjects "may have talked among themselves" (p. 25). If true, this admits to a lack of care in data collection that we think most behavioral scientists would view very critically. But putting that aside, we still cannot see how it could have resulted in six people completing SUPP__ with SURGERY. Are we to suppose that one of the subjects said to others, "I think I am going to write down SURGERY as my completion for SUPP__" and the other subjects chose to imitate this choice? In our opinion, the scenario is preposterous.

A reasonable alternative hypothesis, of course, would be that this is still another sign that the data are not genuine and that perhaps there never really were six subjects who actually produced SURGERY. In examining the data in more detail, we recently noticed another interesting detail about the subject records containing SURGERY: Five of the six come from the same (0,0) node in the lattice shown in Figure 1 of our article. That is, all were subject records with zero benefit word and zero cost word completions (the sixth subject had a zero in one and a three in another). As far as we can tell, none of Chatterjee's explanations for the SURGERY responses would shed any light on why these subjects would have bunched up on the two key dependent variables in the study. By contrast, if these data records were fabricated or otherwise corrupted, it is easier to see how this might have happened.

Mismatch and reduplication co-occur: A new troubling observation

Thinking about Chatterjee's proposed explanations for the Reduplication Effect, we began to reflect on how the two patterns of oddity—Reduplication and Mismatch—might be related to each other. Why would the file contain six instances of SURGERY ostensibly

Downloaded by [University of South Florida] at 03:40 04 February 2016

provided by subjects in response to the cue SUPP__? Is it possible, we wondered, that these six data records were copied and pasted from a single original record containing the SURGERY response (perhaps with bits of random change sprinkled here and there to obscure what was being done)? If so, would these six subjects also be peculiarly alike in their filler word-stem completions (again, the filler stems were chosen to be unrelated to anything else going on in the experiment)?

We took all 15 possible pairings of the six subjects from the SURGERY group and computed the intersubject response-word distance for the filler stems only (the same measure used in Figure 2 of our reanalysis article). Sure enough, the mean distance between SURGERY subjects here was just 2.75 (the range went from 1 to 4). Comparing this value against Panels 2 and 3 of Figure 2 in our article, one sees that this distance measure is very abnormally low relative to the population of subject pairs as a whole. Indeed, it is about as low as the values in Panel 1 (the pairings from the (5,0) and (0,5) groups that we have been discussing at length here and in our article).

For three of the nine words, all six SURGERY subjects made the same choice. All six completed FO as FORT, whereas of the other 88 subjects, nine chose FORT. All six completed NA as NAME, whereas just 15 of the other 88 subjects completed it that way. And all six completed SPO as SPOT, whereas 14 of the other 88 subjects did the same.

To sum up the point, the six SURGERY subjects turn out to show extreme reduplication rates in their choices for the nine filler words. So we now have three "Reduplication Hot Spots" in the data set, rather than just the two discussed in our article: the (5,0) subjects, the (0,5) subjects, and the SURGERY subjects (who drew attention precisely because of the absurd response they supposedly provided in response to the target stem SUPP_).

This new observation is extremely pertinent to the interpretation offered by Chatterjee (2016/this issue) for the Reduplication Effect, because unlike the (5,0) and (0,5) subjects, *these subjects cannot be said to be high in primability*. In fact, if anything, they ought to be extremely *low* in primability. Thus, the explanation offered by Chatterjee for the reduplication effect in the (5,0) and (0,5) groups—that highly primable people are for some reason drawn to the same particular word choices that have no obvious semantic commonalities—falls flat here.

One alternative possibility with which this analysis is potentially consistent, in our opinion, is simple but disturbing: that three boluses of corrupted data were injected into what might have been a genuine data set from the start—the (5,0) group, the (0,5) group, and the SURGERY group, who as noted fell almost entirely within the (0,0) point in the lattice. A plausible reason for the possible injection of the (5,0) and (0,5) groups is obvious and has already been discussed, but what would be the motive for the injection of (0,0) data points? These data points would not have amplified the basic priming effects reported in the article (we have no idea what they might have done to the mediation results also reported in the same article, as we have not attempted to duplicate those). One possibility, however, is that (0,0) points might have been added to make the (5,0) and (0,5) data stand out less (at least for anyone casually exploring the data at the level of univariate distributions). After all, there were very few other subjects with zero cost words, a fact that might have been thought to make the presence of the suspect (5,0) and (0,5) data otherwise a bit conspicuous. That interpretation assumes fabrication, of course. There may be other interpretations that would involve equipment or human error, although we cannot see exactly how those could explain the patterns noted here.

One very reasonable concern with scrutinizing raw data sets as we have been doing here is the possibility that multiple hypothesis testing (more colloquially, "fishing around") may unacceptably drive up the chances of false alarms. If one does enough tests, one will eventually find some red flags in any data set—even where there is nothing truly amiss. However, in the current case it is amply clear why the SURGERY group became a singular focus of discussion as soon as Wicherts pointed it out, and the new finding that it shows exactly the same red flag as the other two groups of data picked out for other reasons can hardly be dismissed as the results of a fishing expedition.

Relatedly, we can probably infer from the thinness of the left tail of Figure 2, Panel B, that there cannot be very many more (if any) additional reduplication "hot spots" lurking within the data set. A very limited hypothesis-focused analysis has probably uncovered most, if not all, of the hot spots.

Other issues

Hypothetical constraints on data fabrication

Over the course of his commentary, Chatterjee (2016/this issue) suggests a number of potential constraints that he contends any data fabricator would be bound to follow, and then uses the fact that the data do not conform to these constraints to argue that the possibility of fabrication can be ruled out.

For example, he contends it should be assumed that "any data fabricator who was attempting to alter data would do so by copy-pasting entire records (rows) not columns" (p. 24). He also suggests it should be

Downloaded by [University of South Florida] at 03:40 04 February 2016

assumed that any data fabricator would sprinkle variability more or less uniformly across different rows, and thus, he argues, the fact there is quite a bit of variability of word-stem completion choices within a few columns of the most suspect sets of data records somehow rules out data fabrication. Most oddly, he even suggests that in order to use copy-and-paste data fabrication, there must have been a single original bona fide "donor" subject whose data actually showed a desired effect—as if a fabricator would find it too burdensome to create from scratch or manually edit a subject record before copying and pasting multiple copies of it.

All these constraints seem extremely unconvincing to us. We think any data fabricator with even a modicum of shrewdness would instinctively avoid following any simple and stereotyped fabrication strategy—precisely to make detection more complex (colloquially, to "cover their tracks"). If possessed of decent facility with a tool like Microsoft Excel, he or she (or they) could copy and paste rows here and there, edit the results a little bit, copy and paste the results from that, and so forth—maybe from time to time copying and pasting portions of columns or individual items, always adding new bits of noise by hand at each step. That (in our view, more realistic) sort of fabrication strategy would be unlikely to result in data that would conform to any of the constraints that Chatterjee conjures up. However, it might well trigger the very alarm bells brought out in our reanalysis of the data. (Again, we do not seek here to foreclose scenarios besides fabrication but rather to explain why the positive arguments against possible fabrication marshalled by Chatterjee [2016/this issue] do not seem to us to have any force.)

Chatterjee also complains that it is odd that we did not analyze and discuss the mood variable contained in one column of the Study 3 data set. Indeed, we do not think we ever looked at this variable, as it was not even discussed in their article. Chatterjee states that the data were "never analyzed" by himself or his collaborators, either, prior to the publication of Chatterjee et al. (2012) (p. 25). But he says that now that he has analyzed the effect of prime condition on mood, he finds theoretically congenial priming effects. These new positive findings, he claims, provide evidence that the data set must be bona fide. We do not understand this argument. If data involving some variables were fabricated, the same could be true for other variables. If some process besides fabrication corrupted some of the data in a biased fashion, it could have done the same to other variables. Similarly, we would not put much weight on Sinha's (2016/this issue) reported reanalysis of data after correcting for some sort of "coding error," because as far as we can tell, she offers no sensible account of the core anomalies discussed here.

Triangular-shaped distribution

Chatterjee claims that we overstated the peculiarity of the triangular-shaped distribution with its clumps at the extreme points (Figure 1 of Pashler et al., 2016). He says that simulations carried out by his unnamed consultant prove that the distribution is reasonable in shape. As far as we can tell, although these simulations (taken at face value) could potentially explain the triangular shape of the distribution, they do not provide any principled reason for the clumping at the extreme points in the lattice (5,0) and (0,5). Second, the assumption of an extreme negative correlation (needed to explain the triangular-shaped distribution) seems to have been conjured up without actually consulting data from the memory research field. Studies of intentional working memory tasks (where people are trying to remember material from, e.g., a brief display) seem generally to show negligible correlations across subjects between recall success for any one item and success on any other (Busey & Townsend, 2001; Fougnie, Suchow, & Alvarez, 2012). But the alleged priming examined by Chatterjee et al. (2012) involves implicit measures collected with subjects who were not trying to store anything in memory. To us, that would be a situation where the rationale for expecting negative correlations would be even *weaker* than it would be in the case of intentional working memory tasks.

Studies 1 and 2

Chatterjee also tells readers that we have "withdrawn" our previously expressed concerns about their data from Studies 1 and 2 (p. 19). The implication is that we have decided that the other two studies in their article deserve a clean bill of health. This is not the case. Chatterjee and his colleagues may stand behind these other studies, but we remain concerned about those data as well. In Pashler et al. (2016/this issue) we do not describe the oddities uncovered in the first two studies, because reviewers of an earlier version of our paper felt—as we did—that these oddities were arguably less compelling and certainly harder to quantify. Nonetheless, the peculiarities uncovered in the first two studies continue to strike us as troubling, especially in the context of what has come to light in Study 3, and we would invite others to examine these raw data as well (downloadable from http://laplab.ucsd.edu/Chattdata/).

To mention just one of these oddities, in Study 2, subjects indicated how much time they wanted to donate to a good cause, and brief exposure to a prime manipulation (Credit vs. Cash vs. Neutral) was reported

to have affected their donation decisions very dramatically. As one commonly finds in human judgments, integer values (like 2 hr or 6 hr donated) were offered by subjects much more often than noninteger values. More specifically, 166 of the 184 subjects' responses were integer values. Of the 18 noninteger responses, 13 appeared in just one of the three conditions (a highly significant difference), and moreover, 11 of the 18 cases in this one condition were duplicate values (7.5 hr in the Credit condition). It is not clear to us how to compute model-free measures of the likelihood of this distribution of noninteger values happening with real uncorrupted data, but we suspect that many readers with experience in behavioral data analysis will share our sense that it is more than a little peculiar—even allowing for the fact that coincidences sometimes occur in bona fide data sets.

Replication

Chatterjee (2016/this issue) refers to a supposed recent "replication" of Chatterjee et al. (2012). A direct replication would not by itself shed much light at all on the cause of the oddities in Study 3. In any case, however, it appears from what we have seen that the publication they cite did not even investigate word fragment completion.

Effect sizes

Both Chatterjee (2016/this issue) and Sinha (2016/this issue) argue strenuously against our suggestion that the effect sizes in Studies 1 and 2 are improbably large. As far as we can tell, we did not draw any strong conclusions from this point. We hope that over the next few years multiple independent replications will shed light on the true effect sizes for money priming using a variety of designs.

In summary, in our judgment the comments offered by the original authors do not improve the credibility of the data collection and analysis presented by Chatterjee et al. (2012). The authors say they are seeking to retract one study, but they have not put to rest the many worrisome issues that their raw data have brought to light. It may very well be that others will be able to figure out how innocuous mistakes or malfunctions could have caused the oddities discussed here. In their various comments on our reanalysis, however, the original authors do not seem to have achieved that purpose.

References

Busey, T. A., & Townsend, J. T. (2001). Independent sampling vs interitem dependencies in whole report processing: Contributions of processing architecture and variable attention. *Journal of Mathematical Psychology*, *45*, 283–323. doi:10.1006/jmps.2000.1317

Chatterjee, P. (2016/this issue). Response to Pashler et al. (2016). *Basic and Applied Social Psychology*, *38*(1), 19–29. doi:10.1080/01973533.2015.1129335

Chatterjee, P., Rose, R. L., & Sinha, J. (2012). Why money meanings matter in decisions to donate time and money. *Marketing Letters*, *21*(2), 1–10.

Fougnie, D., Suchow, J. W., & Alvarez, G. A. (2012). Variability in the quality of visual working memory. *Nature Communications*, *3*, 1229. doi:10.1038/ncomms2237

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

Pashler, H., Rohrer, D., Abramson, I., Wolfson, T., & Harris, C. R. (2016/this issue). A social priming data set with troubling oddities. *Basic and Applied Social Psychology*, *38*(1), 3–18. doi:10.1080/01973533.2015.1124767

Rose, R. L. (2016/this issue). Cautious thoughts on "A social priming data set with troubling oddities." *Basic and Applied Social Psychology*, *38*(1), 30–32. doi:10.1080/01973533.2015.1127237

Sinha, J. (2016/this issue). Selective literature review and selective data analyses: Implications for the (re)analysis of public access research data. *Basic and Applied Social Psychology*, *38*(1), 33–40. doi:10.1080/01973533.2015.1129336

Downloaded by [University of South Florida] at 03:40 04 February 2016