Reply to Comments on "Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition"

Edward Vul,[1] Christine Harris,[2] Piotr Winkielman,[2] and Harold Pashler[2]

[1]Massachussetts Institute of Technology and [2]University of California, San Diego

We are grateful to the commentators for providing many stimulating and valuable observations. The main point of our article was to call attention to the overestimation of individual differences correlations in a subset of neuroimaging papers. To structure our discussion of these comments, we list the main points from our paper, note where commentators have agreed or disagreed with each, and provide our own reactions to their comments.

MAIN POINTS

1. We claimed that approximately half of a sample of studies reporting individual difference correlations between brain activity and behavioral measures of personality, emotion, and social cognition computed the correlation by (a) computing a separate correlation across people for each voxel in a normalized brain, (b) identifying highly correlated voxels, and (c) calculating (and reporting) the mean (or peak) correlation from just these correlation "hot spots."

There does not appear to be any disagreement over this point. We have invited any author who believes that we wrongly described how they computed correlation values to write us and elaborate on their procedure; so far, this has revealed no misclassifications. Thus, our claims about how the correlations were calculated have not been disputed.

Nonetheless, a casual reader of Lieberman et al. (2009, this issue) might assume that there is some dispute here, as Lie-

berman et al. say we "incorrectly" described the "inferential procedure" of these studies. However, they are merely arguing that the correlation values were presented and interpreted differently than we say. Specifically, they contend that inferences in articles we criticized were based exclusively on the significance of the whole-brain analysis and that the reported correlation values were always understood to lack any meaningful interpretation and thus were never ascribed one. We will return to this issue below (see Point 3).

2. We argued that a correlation computed as described in Point 1 is statistically guaranteed to provide an inflated estimate of the true underlying (population) correlation—that is, it is guaranteed to provide a number whose expected value exceeds what would be obtained if the same measurements were made again in the same hotspots (with the same subjects or with different ones) and does so by an amount that cannot be calculated (and thus corrected for).

The reader will note that the statisticians who have commented (Lazar, 2009, this issue, and Lindquist & Gelman, 2009, this issue) take the validity of this point as a given, as does Feldman Barrett (2009, this issue), from a psychometric perspective; Yarkoni (2009, this issue) evidently agrees, and the commentators who weigh in to defend existing practices (Lieberman et al.; Nichols & Poline, 2009, this issue) also concede the point. Thus, there appears to be unanimous agreement about our central claim, namely, that the many nonindependent correlations that have been reported in the literature are inflated.

3. Our article implied that these nonindependent correlations (and accompanying scattergrams) were typically presented and interpreted as if they were "normal

correlations'' that could support population-level inferences about the strength of the relationship between brain activity and behavior (i.e., as unbiased measures of the effect size of a linear relationship, interpretable as any independent correlation).

We did not expect this point to be controversial, but it has inspired some strong protests. Lieberman et al. (as well as Jabbi et al., 2009) argue that these correlations were definitely not presented or used as "normal correlations" that support inferences about the magnitude of the correlation present in the population but were merely used as "descriptive statistics" embedded within significance tests from a whole-brain analysis, being ascribed no meaning beyond the outcome of the significance test.

As Lieberman et al. say, "For any particular sample size, . . . r values are merely redescriptions of the p values obtained in the one inferential step, and they provide no additional inferential information of their own" (p. 300). They also ask "If the reporting of correlation values and scatterplots is merely descriptive, then why do it? Vul et al. imply that its purpose is to sell correlations that appear to be very strong" (p. 300), but as Lieberman et al. explain, salesmanship was not the point: Scattergrams were plotted merely to show that the significant null-hypothesis tests (indicating the presence of a nonzero correlation) were not driven by outliers.

This complaint about our presentation led us to revisit the literature. Perhaps we had misunderstood the way correlations are interpreted in this literature and had failed to note that the correlation coefficients were actually presented and discussed merely as a description of the data, rather than as an inferential estimate of the strength of the correlation in the population. To find out, we reexamined not only the empirical articles, but also review articles and general-interest books written by researchers whose statistics we had labeled as nonindependent.

We believe that anyone examining this literature with this question in mind would have to conclude that the "selling" of the strength of these correlations has been at least as enthusiastic and unqualified as our article implied.

For example, Eisenberger, Lieberman, and Satpute (2005) compared a nonindependently computed fMRI–behavior correlation to an independently computed behavior–behavior correlation and concluded that "dACC reactivity was a substantially better predictor of interoceptive accuracy than self-reported neuroticism was, accounting for nearly five times the variance in interoceptive accuracy (74% vs. 16%). With the utilization of these types of methods, future personality research may have the potential to account for a substantially larger portion of the variation in human experience and behavior than has been possible with self-report measures alone" (p. 179). It does not seem plausible to us that comments like this reflect only inferences about the statistical significance of the dorsal anterior cingulate cortex correlation; instead, the (inflated) magnitude of the nonindependent correlation is interpreted in the

strongest possible way: as an estimate of the percentage of variance accounted for that is directly comparable to normal, independent correlations.

Similarly, Canli et al. (2001) say, "The strength of correlations between neural activation and personality dimensions was strong, especially compared with behavioral data one might encounter in other studies of personality and emotion" (p. 38). Again, the emphasis seems to be placed squarely on the alleged strength of a nonindependently computed correlation.

Review articles discussing this literature often describe the correlation coefficients with no qualification. For example, Eisenberger and Lieberman (2004) summarized their own findings as "magnitude of dACC activity correlated strongly with self-reports . . ." (p. 295) and showed scattergrams depicting one such correlation without mentioning the whole-brain analysis that was supposedly the only finding given any inferential weight. Similarly, when these correlations are communicated to the public, their strength is often focused upon—for instance, Ochsner was quoted in a Stanford University press bulletin as stating that the (nonindependent) correlations presented in Ochsner et al. (2006) were "insanely strong" (White, 2006, p. 1).

4. Our paper implied that the overestimation due to nonindependence is very likely to be large.

We did not attempt to quantify the degree of overestimation, but we did imply that it was not tiny, as based on our simulations (especially Appendix B). But what is the truth of the matter?

Lieberman et al. argue that the magnitude of overestimation can be determined by calculating the difference between the mean "red" and "green" correlations in our Figure 5, and they argue that, properly measured, this difference is not so large as our article suggested. As we stated in our original article (p. 281), we believe that comparison of red and green correlations provides no basis for estimating the magnitude of inflation. Even if it did, an estimate of the difference of mean correlations is of little value: The correlation coefficient is a nonlinear scale, so the idea that this complex form of mismeasurement would impose a constant additive increment seems to us mathematically impossible.

In any given study, the magnitude of inflation will vary not only as a function of factors that are easily determined (sample size, threshold, and number of voxels), but also as a function of some indeterminable factors (true effect size, noise of the measurements). As such, the magnitude of inflation will vary wildly and unpredictably from study to study, leaving no possibility to correct or estimate the inflation of any given result. We know of only one way to estimate how much inflation occurred in a study: to reanalyze the data using unbiased measures.

Unfortunately, so far, we have heard of just one such reanalysis. Poldrack and Mumford (in press) compared the results of a nonindependent correlation estimates with an independent reanalysis computed using cross-validation across runs (less conservative than a cross-validation across subjects, which may be more appropriate; see Feldman Barrett, 2009, and discussion

below). They found that a nonindependent correlation of ~.81 dropped to ~.56. This implies that, in this case, the nonindependent computation had overestimated effect size (as measured by r^2) by at least a factor of two. Of course, we cannot generalize these results to other nonindependent correlation estimates, and we cannot know the range of overestimation across studies until the data from many studies are reanalyzed. We reiterate our call for researchers to compute and report such reanalyses.

5. We argued that effect sizes of correlations are of great importance and implied that methods that merely test the statistical significance of a correlation without providing information on effect size are of limited value.

It is not clear whether any of the commentators disagree with this point. Perhaps Nichols and Poline do because they maintain that there is no problem with fMRI statistical practices, as methods are available to control false alarm rates in null-hypothesis tests computed over the whole brain. However, as we noted in our article, and as most statisticians and methodologists generally seem to agree (Thompson, 1996; Wilkinson and the Task Force on Statistical Inference, 1999), effect sizes are of great import.

One reason they are so important is that the weaker a correlation is, the more likely it is to reflect (a possible multiplicity of) highly indirect causal paths that are more likely to distract researchers than to enlighten them. Like Lindquist and Gelman, we suspect that brain–behavior correlations will rarely have a true value of exactly zero when precisely measured. As Nunnaly (1960) said, "If rejection of the null hypothesis were the real intention of psychological experiments, there would usually be no need to gather data" (p. 643). If small (but nonzero) correlations are the rule, what is learned from an analysis that enumerates brain areas in which a nonzero brain–behavior correlation exists, while providing no information about the strength of these correlations? In our view, not very much.

6. We pointed out that although most studies used appropriate multiple comparison corrections (and thus identified voxels that do indeed have nonzero correlations), these methods are not always correctly applied, even in articles that have passed peer review.

This point seems to have caused some confusion. Some, including Lieberman et al. and Jabbi et al., interpreted our article as condemning whole-brain analyses and implying that the entire literature is strewn with effects reflecting nothing but noise (our inclusion of all-noise simulations may have inadvertently encouraged this misinterpretation).

To clarify, we find nothing statistically inappropriate about using a properly corrected whole-brain analysis to identify regions with correlations deviating from zero (but see Point 5 above). Our primary objection pertains to the way that correlation magnitudes have been computed.

However, we also identified one multiple comparison correction error that can easily create apparent correlations out of

pure noise: a misreading of simulations by Forman et al. (1995). Specifically, this error entails adopting a cluster threshold with an extent of 10 voxels and a per-voxel p value of $<.005$ and assuming that these parameters provide a false alarm rate less than 0.000001. As we mentioned, this error can be found in Eisenberger, Lieberman, and Williams (2003) and a number of other articles (e.g., Taylor, Eisenberger, Saxbe, Lehman, & Lieberman, 2006), but it is definitely not present in most of the literature that we have been discussing; a review of the extent of this error is now in preparation.

The fact that such a mistake passed review in prestigious journals indicates that modern multiple-comparison correction procedures can be treacherous. We see this as another reason to prefer the independent analysis methods we recommend.

7. We suggested that independent (e.g., cross-validation) methods should be used to compute unbiased correlation coefficients, and pointed out that doing so not only provides a valid measure of effect size, but also allows for simpler and more transparent inferential tests (circumventing the pitfalls discussed in Point 6).

We are surprised that Lieberman et al. and Nichols and Poline agree that nonindependently computed correlations are biased and cannot support population-level inferences, and yet they do not embrace the cross-validation approach nor offer any other alternative. In essence then, they are implying that brain-imaging research on individual differences can proceed without any information about effect sizes of relationships. We doubt that this can be a promising strategy (see Point 5).

OTHER POINTS

Commentators made a number of additional important points, to which we now turn.

The Role of Sample Size

Lieberman et al. point out that our simulation using 10 subjects that produced a correlation of 0.8 from pure noise was not representative of the average number of subjects used in the overall set of studies that we reviewed (which had a mean sample size of 18). They note that samples with 18 subjects are much less likely to produce a correlation exceeding 0.8 out of pure noise. Indeed, as Yarkoni also points out, the magnitude of inflation is likely to be smaller with greater sample sizes. However, in the studies we surveyed, the sample sizes that actually produced correlations of 0.8 had a mean sample of 12 subjects, and none of them had more than 16 subjects. Thus, although Lieberman et al. were right to say that a study with 18 subjects is unlikely to produce a correlation of 0.8 from pure noise, they were wrong to assume that this sample size is representative of the studies that actually do produce such large correlations.

In his thoughtful commentary, Yarkoni goes farther to suggest that small sample sizes, rather than nonindependence, are

responsible for the inflated correlation estimates in this literature. He makes a very important observation, which was alluded to above but not discussed in our article: Big correlations tend to come from small studies. We can confirm his point within our sample: r^2 and $\log(n)$ are significantly negatively correlated both for nonindependent studies ($r = -0.62$) and for independent studies ($r = -0.58$, both $ps < .01$).

We believe that small sample size conspires with nonindependence and the number of voxels (measures) to produce misleading literature. If every researcher computed a few independent correlations and reported all of the findings, then the published numbers would be free of bias. On the other hand, if researchers test many hypotheses and report only the significant ones, the published literature will show a bias, even without biased analysis procedures (see also Ioannidis, 2008, for fascinating examples from epidemiology and medicine). This, of course, is the familiar "file-drawer problem" (Rosenthal, 1979).

In our view, the problem with nonindependent correlations is, in a sense, just another file-drawer problem, but exacerbated in two ways. First, nonindependent analyses build the file-drawer problem into the analysis procedure itself, rather than imposing it externally through biased publication choices. Second, a nonindependent analysis over the enormous number of measurements obtained from an fMRI experiment creates a file-drawer that is far larger than the most bloated file-drawer of an investigator doing independent tests one at a time; thus, the inflation of effect sizes will be larger.

That said, the underlying statistical issues raised by Yarkoni are important: The interactions between nonindependence, the number of comparisons, the number of subjects, and the statistical threshold used are complicated and need further analysis. Our simulations assumed only measurement error and neglected subject sampling variability; thus, we recommended cross-validation across runs. It may very well turn out that a more rigorous test (cross-validation across subjects) is needed to obtain valid generalizable numbers, for reasons that are described by Feldman Barrett.

Scope of Literature Review

Lieberman et al. criticize us for vaguely specifying the scope of our literature review. We plead guilty to this charge, but as far as we can tell, nothing important hinges on it. The nonindependent correlations that we described in social and personality neuroscience are common over the whole spectrum of fMRI research.

Missing Correlations in Earlier Version of Our Article

Lieberman et al. note that the version of our article that circulated on the Internet missed 54 correlations as well as a few other errors, and they imply that these omissions show signs of bias. The final version of our article (appearing in this journal) includes all of Lieberman et al.'s proposed corrections except for

35 correlations from an exploratory analysis in a paper by Rilling et al. (2007), whose relevance we dispute (see Footnote 10 in the original article).

Although it would have obviously been ironic (as well as improper) for us to have cherry-picked data to promote a campaign against cherry picking data, it would also have been self-defeating (after all, our chart was numerically coded with references to published articles and it was certain to be checked by authors). Moreover, it would also have been pointless, because the sole conclusion that our article drew from the distribution of independent and nonindependent correlation magnitudes was that nonindependent analyses are behind "the great majority of the correlations in the literature that struck us as impossibly high." This remains correct, and indeed overwhelmingly so, even with the 35 contested correlations from Rilling et al. (2007) included: 66 out of the 78 correlations that exceeded our initial "upper bound" estimate on plausible correlation magnitudes (.74) were computed nonindependently.

Replications

Lieberman et al. argue that some of the findings we criticize have stood up to replications. Unfortunately, the question of what should count as a replication for purposes of this discussion is not as simple as it might seem.

If the finding at issue is "a measure of Brain Area A accounts for roughly X percent of across-subject variation in Behavioral Measure Z," then what needs to be replicated is the correlation magnitude in a new sample using an independently localized, matching region (a nonindependent analysis can hardly validate another nonindependent analysis.) To our knowledge, this has never been done for any of the nonindependently computed correlations that we discussed.

If the conclusion is merely "Brain Area A correlates with Measure Z to a nonzero degree," then replication of a location is sufficient. But what constitutes a replication of a location? Answering this question requires quantifying uncertainty on the location of a cluster and deciding what should count as a sufficiently similar anatomical region (in different individuals with different neuroanatomy). In the absence of a review that grapples seriously with these issues, we are doubtful about loose claims of replications that are made without specifying the details of what is supposed to have been replicated and what would have been considered to be a nonreplication.

Restriction of Range

Lieberman et al. say that the difference between the independent and nonindependent correlations may be attributable to underestimation of independent correlations due to a restriction of range: that is, selecting regions based on a simple contrast of A > B will tend to select voxels with low variability across subjects, thus restricting the range of the data. Lieberman et al. then correct for this range restriction using some debatable

assumptions and suggest that independent correlations may actually be as high as the nonindependent correlations.

Fortunately, in our survey sample, we can test for the effect of restricted range by comparing the independent correlations in purely anatomical regions of interest (which are not affected by the restricted range issue) with the independent correlations obtained from orthogonal (noncorrelation) contrasts (which Lieberman et al. argued are underestimated due to a restriction of range). We find no difference between these groups—that is, no effect of restricted range ($p = .3$)—and Lieberman et al.'s calculation of a mean shift of ~ 0.13 is well outside the 95% confidence interval on the mean difference between these two sets of studies (-0.02 to 0.06). Thus, we are led to suspect that one or more of the assumptions that went into this correction by Lieberman et al. were false. In any case, given the fact that the mean difference between independent and nonindependent correlations provides no sound basis for estimating inflation due to nonindependence, we see little at stake here.

''Impossible'' Correlations

Lieberman et al. point out that we were incorrect in describing any particular correlation value as "impossibly large," and they imply that we underestimated reliabilities, whereas Nichols and Poline suggest that we are confusing bounds on samples with population correlations. We were too casual in describing how typically modest reliabilities constrain observable correlations—indeed, the only absolute bound one may put on a sample correlation is 1.0. Our estimate of 0.74 as an "upper bound" referred to the expected measured correlation under the implausible assumption that the true correlation underlying the noisy measurements is perfect (1.0). Correlations in excess of this "upper bound" are certainly possible, but unlikely—the larger and more frequent such correlations are, the less likely the set of correlations as a whole is to have arisen from unbiased measurements. It is therefore striking how many researchers have been reporting such unlikely correlations—a mystery that we believe is largely resolved by the findings of our survey.

The Past and Future of Nonindependence Problems

Authors commenting on our article have raised an important point: This problem is not new, and certainly not unique to social neuroscience, to fMRI, or to neuroscience. Rather this problem arises with all research methods that generate a great deal of data, and in which only some a priori unknown subset of the data is of special interest. The problem we call *nonindependence* (referring to the conditional dependence between voxel selection criteria and the effect size measure) has been called *selection bias* in survey sampling, *testing on training data* that results in *overfitting* in machine learning, *circularity* in logic, *double dipping* in fMRI (Kriegeskorte, Simmons, Bellgowan, & Baker, in press), and *the winner's curse* in auctioneering. Whatever name one prefers, the problem is the same: Estimates obtained from a subset of data selected for that particular measurement will be biased.

It is interesting that the first eruption of this issue that we have learned about took place in the field of psychometrics, when people constructed tests by selecting a subset of a large pool of potential items on the basis of their ability to predict some external criterion (like college graduation) and wished to say how accurate their test was.[1] Cureton (1950) experimented with completely random outcome data and found that when he assessed validity using the same data he had used to select the items, he obtained a high, but obviously spurious, measure of "validity." Cureton summed up his findings by saying, "When a validity coefficient is computed from the same data used in making an item analysis, this coefficient cannot be interpreted uncritically. And, contrary to many statements in the literature, it cannot be interpreted 'with caution' either. There is one clear interpretation for all such validity coefficients. This interpretation is 'Baloney!'" (p. 96). Though the details and the language may be different in every case, it would seem that the insight Cureton revealed is one that researchers in many fields are fated to rediscover.

REFERENCES

Canli, T., Zhao, Z., Desmond, J.E., Kang, E., Gross, J., & Gabrieli, J.D.E. (2001). An fMRI study of personality influences on brain reactivity to emotional stimuli. *Behavioral Neuroscience*, *115*, 33–42.

Cureton, E.E. (1950). Validity, reliability, and baloney. *Educational and Psychological Measurement*, *10*, 94–96.

Eisenberger, N.I., & Lieberman, M.D. (2004). Why rejection hurts: A common neural alarm system for physical and social pain. *Trends in Cognitive Neuroscience*, *8*, 294–300.

Eisenberger, N.I., Lieberman, M.D., & Satpute, A.B. (2005). Personality from a controlled processing perspective: An fMRI study of neuroticism, extraversion, and self-consciousness. *Cognitive, Affective & Behavioral Neuroscience*, *5*, 169–181.

Eisenberger, N.I., Lieberman, M.D., & Williams, K.D. (2003). Does rejection hurt? An FMRI study of social exclusion. *Science*, *302*, 290–292.

Feldman Barrett, L. (2009). Understanding the mind by measuring the brain: Lessons from measuring behavior (Commentary on Vul et al., 2009). *Perspectives on Psychological Science*, *4*, 314–318.

Forman, S.D., Cohen, J.D., Fitzgerald, M., Eddy, W.F., Mintun, M.A., & Noll, D.C. (1995). Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): Use of a cluster-size threshold. *Magnetic Resonance in Medicine*, *33*, 636–647.

Ioannidis, J. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*, 640–648.

Jabbi, M., Keysers, C., Singer, T., & Stephan, K.E. (2009). *Response to "Voodoo Correlations in Social Neuroscience" by Vul et al.* Retrieved from http://www.bcn-nic.nl/replyVul.pdf

Kriegeskorte, N., Simmons, W.K., Bellgowan, P.S.F., & Baker, C.I. (in press). Circular analysis in systems neuroscience: The dangers of double dipping. *Nature Neuroscience*.

Lazar, N.A. (2009). Discussion of "Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition" by Vul et al. (2009). *Perspectives on Psychological Science*, *4*, 308–309.

Lieberman, M.D., Berkman, E.T., & Wager, T.D. (2009). Correlations in social neuroscience aren't voodoo: Commentary on Vul et al. (2009). *Perspectives on Psychological Science*, *4*, 299–307.

Lindquist, M.A., & Gelman, A. (2009). Correlations and multiple comparisons in functional imaging: A statistical perspective (Commentary on Vul et al., 2009). *Perspectives on Psychological Science*, *4*, 310–313.

Nichols, T.E., & Poline, J.-B. (2009). Commentary on Vul et al.'s (2009) "Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition." *Perspectives on Psychological Science*, *4*, 291–293.

Nunnally, J. (1960). The place of statistics in psychology. *Educational and Psychological Measurement*, *20*, 641–650.

Poldrack, R., & Mumford, J. (in press). Independence in ROI analyses: Where's the voodoo? *Social, Cognitive, and Affective Neuroscience*.

Rilling, J.K., Glenn, A.L., Jairam, M.R., Pagnoni, G., Goldsmith, D.R., Elfenbein, H.A., & Lilienfeld, S.O. (2007). Neural correlates of social cooperation and non-cooperation as a function of psychopathy. *Biological Psychiatry*, *61*, 1260–1271.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*, 638–641.

Taylor, S.E., Eisenberger, N.I., Saxbe, D., Lehman, B.J., & Lieberman, M.D. (2006). Neural responses to emotional stimuli are associated with childhood family stress. *Biological Psychiatry*, *60*, 296–301.

Thompson, B. (1996). Statistical significance tests, effect size reporting, and the vain pursuit of pseudo-objectivity. *Theory & Psychology*, *9*, 191–196.

White, T. (2006, Feb. 1). *People who fear pain are more likely to suffer it*. Retrieved from http://news-service.stanford.edu/news/2006/february1/med-anxiety-020106.html

Wilkinson, L.The Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604.

Yarkoni, T. (2009). Big correlations in little studies: Inflated fMRI correlations reflect low statistical power (Commentary on Vul et al., 2009). *Perspectives on Psychological Science*, *4*, 294–298.