Invited reply submitted to *Perspectives on Psychological Science*.

Correlations in social neuroscience aren't voodoo: A reply to Vul et al.

Matthew D. Lieberman[1], Elliot T. Berkman[1], Tor D. Wager[2]

[1]Department of Psychology, University of California, Los Angeles

[2] Department of Psychology, Columbia University

ACKNOWLEDGMENTS: We would like to thank the following individuals (in alphabetical order) for feedback on drafts of this paper and relevant discussions: Arthur Aron, Mahzarin Banaji, Peter Bentler, Sarah Blakemore, Colin Camerer, Turhan Canli, Jessica Cohen, William Cunningham, Mark D'Esposito, Naomi Eisenberger, Emily Falk, Susan Fiske, Karl Friston, Chris Frith, Rita Goldstein, Didier Grandjean, Amanda Guyer, Christine Hooker, Christian Keysers, William Killgore, Ethan Kross, Claus Lamm, Martin Lindquist, Jason Mitchell, Dean Mobbs, Keely Muscatell, Thomas Nichols, Kevin Ochsner, John O'Doherty, Stephanie Ortigue, Jennifer Pfeifer, Daniel Pine, Russ Poldrack, Joshua Poore, Lian Rameson, Steve Reise, James Rilling, David Sander, Ajay Satpute, Sophie Schwartz, Tania Singer, Thomas Straube, Hidehiko Takahashi, Shelley Taylor, Alex Todorov, Patrik Vuilleumier, Paul Whalen, Kip Williams.

Correspondence should be addressed to:

Matthew Lieberman
Department of Psychology
University of California, Los Angeles
Los Angeles, CA 90095-1563
lieber@ucla.edu

ABSTRACT

Vul et al. claim that brain-personality correlations in many social neuroscience studies are "implausibly high," "likely…spurious," and "should not be believed." Several of their conclusions are incorrect due to flawed reasoning, statistical errors, and sampling anomalies. First, the conceptual issues discussed by Vul et al. have little to do with social neuroscience *per se* and are equally relevant for nearly all fMRI analyses that report measures of effect size from searches over multiple voxels or regions (r, t, or Z statistics). Second, Vul et al. incorrectly claim that whole-brain regression analyses use an invalid and "non-independent" two-step inferential procedure. We explain how whole-brain regressions are a valid single-step method of identifying brain regions that have reliable correlations with individual difference measures. Third, Vul et al. claim that large correlations obtained using whole-brain regression analyses may be the result of noise alone. We provide a simulation to demonstrate that typical fMRI sample sizes (N = 15-20) will only rarely produce large correlations in the absence of any true effect. Fourth, Vul et al. claim that the reported correlations are inflated to the point of being "implausibly high". Though biased *post hoc* correlation estimates are a well-known consequence of conducting multiple tests, Vul et al. make inaccurate assumptions when estimating the theoretical ceiling of such correlations. Moreover, Vul et al.'s own meta-analysis suggests that the magnitude of the bias is an increase of approximately .12, a rather modest estimate to inspire the label 'voodoo'. In addition, after correcting for likely restricted range in several of the "independent" correlations that Vul et al. treat as the gold standard, the means of the "non-independent" and "independent" correlations are nearly identical. Finally, it is troubling that almost 25% of the "non-independent" correlations in the papers reviewed by Vul et al. were omitted from their own meta-analysis without explanation.

The word "voodoo," as applied to science, carries a strong and specific connotation of fraudulence, as popularized by Robert Park's (2000) book, "Voodoo Science: The Road from Foolishness to Fraud." Thus, it is hard to construe the recent paper by Vul et al., entitled "Voodoo correlations in social neuroscience," as anything but a pointed attack on social neuroscience. Vul et al. claim that brain-personality correlations in many social neuroscience studies are "implausibly high" and "likely…spurious," and call for the authors to "correct the scientific record." Are brain-personality correlations in social neuroscience fraudulent, and are they deserving of such a strong and pointed epithet? The answer is a resounding no. Below, we address several of the claims made by Vul et al. and point out errors in their statistical reasoning and their meta-analytic assumptions and procedures that largely invalidate their conclusions.

Is the issue about correlations in social neuroscience?

From the title and abstract, a reader would be forgiven for thinking that the Vul et al. paper had something to do with social neuroscience specifically and the correlations reported in social neuroscience studies. In fact, the issues that they discuss, whether correct or not, apply equally to cognitive, clinical, developmental, affective, personality, and social neuroscience, and to procedures used in other fields as well (e.g. genetics). Any whole-brain regression using an individual difference variable such as age, clinical severity, average reaction time, error rate, or brain activity from a seed region will face the same set of issues regarding effect size estimates under multiple comparisons. Similarly, any whole-brain analysis, whether a regression or a simple contrast comparing condition A to condition B, is suspect under the terms they describe (see also Vul & Kanwisher, in press). In other words, there is nothing intrinsically specific to social neuroscience or correlational analyses. Vul et al. acknowledge this, noting that they selected social neuroscience because it was "the area where these correlations came to our attention" (p. 3). Unfortunately, the title, abstract, and much of the paper leave an impression that the issues are unique to this area.

Do whole-brain correlations use a "non-independent" two-step inference procedure?

Vul et al. (p. 11) contend that correlations resulting from a search across multiple brain regions (or brain "voxels"), the dominant method in neuroimaging research, is a two-step procedure in which the method used to select voxels to test (correlation) and the test performed on the resulting regions (correlation) are not independent. The clearest account of this comes from another paper by Vul and Kanwisher (in press) in which they describe the analogous situation in whole-brain contrast analyses suggesting that, "If one selects only voxels in which condition A produces a greater signal change than condition B, and then evaluates whether the signal change for conditions A and B differ in those voxels using the same data, the second analysis is not independent of the selection criteria" (p. 2). This statement is clearly pointing to the existence of two steps, each involving an inferential procedure, with the second inference guaranteed to produce significant results because of its non-independence from the first inference.

We don't know of any researchers who conduct their analyses this way.[1] When a whole-brain regression analysis is conducted, the goal is typically to identify regions of the brain whose activity shows a reliable non-zero correlation with another individual difference variable. A likelihood estimate that this correlation was produced in the absence of any true effect (e.g. a p-value) is computed for every voxel in the brain without any selection of voxels to test. This is the

[1] We were able to contact authors from 23 of the 28 "non-independent" papers reviewed by Vul et al. Each of the contacted authors reported that they used a single-step inferential procedure similar to the single-step procedure we describe, rather than the two-step procedure described by Vul et al. Several authors expressed frustration that the multiple choice questions asked by Vul et al. did not allow the authors to indicate whether they used one or two inferential steps, contributing to Vul et al.'s misrepresentation of how these studies were conducted.

only inferential step in the procedure, and standard corrections for multiple tests are implemented to avoid false positive results. Subsequently, descriptive statistics (e.g. effect sizes) are reported on a subset of voxels or clusters. The descriptive statistics reported are not an additional inferential step, so there is no "second analysis." For any particular sample size, the t and r-values are merely re-descriptions of the p-values obtained in the one inferential step and provide no additional inferential information of their own. Demonstrating that such r-values do not violate a theoretical upper limit, as Vul et al. suggest, is a separate issue that we address in a later section.

In sum, despite Vul et al.'s characterizing whole-brain regressions as "seriously defective" (p. 22), they provide a valid test, in a single inferential step, of which regions show a reliable linear relation with an individual difference measure. What reported correlations from whole-brain regressions really show is evidence for a non-zero effect, which is what they were designed to test. It is also true that the reported effect sizes (r, t, Z) from whole-brain analyses will be inflated (i.e. over-estimated relative to the population effect size) on average. However, as we detail below, the magnitude of the inflation may be far less than Vul et al. would have readers believe.

How often do large correlations occur without any true effect?

Vul et al. imply that the correlations in at least a sizeable subset of social neuroscience studies are not based on any true underlying relationship between psychological and neural variables (hence the terms "voodoo" and "spurious"). For all statistical tests, there is some likelihood that the observed result is spurious and the true population effect size is zero. This likelihood is what p-values estimate. A p-value of .05 in any research domain suggests that the observed effect would have occurred by chance in 5% of experimental samples. Because a typical whole-brain analysis involves thousands of tests, the likelihood of false positives is much greater, and thus correction for multiple comparisons is essential.

Although spurious correlations will occur (see Figure 4 from Vul et al. on a simulation assuming N=10), the critical question in the context of correlational analyses in fMRI is how often large correlations such as those targeted by Vul et al. will occur in the absence of any true effect—and, when prior anatomical hypotheses are available, how often they will occur in the expected anatomical locations. To assess how frequently they might occur in a typical whole-brain regression analysis, we conducted a simulation (see Figure 1). We examined how often correlations $\geq .80$ are expected to be observed anywhere in the brain in the absence of any true signal (this depends on the sample size and number of effective independent comparisons; see Figure 1 legend for details). With N = 18 subjects (the average N was 18.25 in the social neuroscience studies reviewed by Vul et al.), 76% of the simulated "studies" reported no correlation of $r \geq .80$ by chance anywhere in the (simulated) brain. Only 2% reported 2 or more false positive correlations. This suggests that in actual studies with similar properties and multiple comparison procedures, the majority of reported effects of this magnitude reflect a true underlying relationship.

Additionally, false positive activations are likely to be randomly and uniformly distributed throughout the brain. If each of the social neuroscience studies in question had reported no more than one or two significant correlations, in regions were apparently uniformly distributed over the brain across studies, there would be reason to question whether they were meaningful as a set. However, many studies report multiple correlated regions (consistent with the notion of distributed networks underlying social and affective phenomena) in the same approximate brain areas.

For example, among the articles critiqued by Vul et al. are studies examining fear of pain (Ochsner et al., 2006), empathy for pain (Singer et al., 2004; Singer et al., 2006), and social pain (Eisenberger, Lieberman, & Williams, 2003). In each of these pain-related studies, significant correlations were reported between individual difference measures and activity in the dorsal anterior cingulate cortex, a region central to the experience of pain (Price, 2000). The results of these

studies are clearly not distributed uniformly over the brain, as would be expected if these correlations were spurious. The same point is made by meta-analyses of the neuroimaging literature on emotion, which clearly show "hot spots" of consistently replicated activity across laboratories and task variants (Kober et al., 2008; Wager et al., 2008). Importantly, our meta-analyses suggest that, to a first order of approximation, results from studies of social and emotional processes are no more randomly distributed across the brain than studies in other areas of cognitive neuroscience such as working memory (Wager & Smith, 2003), controlled response selection (Nee, Wager, & Jonides, 2007), and long-term memory (van Snellenberg & Wager, in press).

In sum, even without considering any prior anatomical hypotheses, most, but not all, of the large correlations that Vul et al. target are likely to represent real relationships between brain activity and psychological variables. Furthermore, the use of prior anatomical hypotheses that limit false positive findings are the rule, rather than the exception. It is difficult to reasonably claim that the correlations, as a set, are "voodoo."

How inflated are "non-independent" correlations?

It is a statistical property of any analysis in which multiple tests are conducted that observed effect sizes in significant tests will be inflated—i.e., larger than would be expected in a repeated sample (Tukey, 1977). Vul et al. suggest that so-called "non-independent" correlations (descriptive correlation results from significant regions in voxel-wise searches) resulting from whole-brain analyses are "inflated to the point of being completely untrustworthy" (p. 20) and "should not be believed" (p. 22). While acknowledging that there is inflation in such correlations, it would be useful to know just how inflated they are in the case of the social neuroscience findings criticized by Vul et al.

Although it is impossible to know for sure, the meta-analysis by Vul et al. provides some measure of this inflation within the social neuroscience literature. In their Figure 5, Vul et al. plot the strength of correlations using what they deem acceptable "independent" procedures in green and so-called "non-independent" (biased) correlations in red. The general inference to be drawn is that relative to the gold-standard "independent" correlations, the "non-independent" correlations have higher values and are therefore systematically inflated.

In order to assess the average magnitude of the "independent" and "dependent" correlations, we collected all the papers cited in Vul et al.'s meta-analysis and extracted all of the correlations that met the inclusion criteria they describe. In doing so, we were surprised to find several anomalies between the set of correlations included in the Vul et al. meta-analysis and the set of correlations actually in the papers. We identified 54 correlations in the papers used in their meta-analysis that met their inclusion criteria, but were omitted from the meta-analysis without explanation. We also found 3 "correlations" that were included but were not actually correlations (see Appendix 1 for a breakdown). Among the "non-independent" correlations, almost 25% of the correlations reported in the original papers were not included in Vul et al.'s meta-analysis. The vast majority of the omitted correlations (50 of 54) and mistakenly included effects (3 of 3), if properly included or excluded, work against their hypothesis of inflated correlations due to "non-independent" correlation reporting.[2]

[2] Of the 41 omitted "non-independent" correlations, 38 had values lower than the mean of included "non-independent" correlations. The mean of the omitted "non-independent" correlations (.61) was significantly lower than the included "non-independent" correlations (.69), $t=4.06$, $p<.001$. Of the 13 omitted "independent" correlations, 12 had values higher than the mean of the included "independent" correlations. The mean of the omitted "independent" correlations (.63) was significantly higher than the included "independent" correlations (.57), $t=2.74$, $p<.01$. Of the 3 included non-independent correlations that should have been omitted, all 3 had values higher than the mean of the included non-independent correlations.

Based solely on the correlations that Vul et al. selected to include in their meta-analysis, the mean of "non-independent" correlations (average r = .69) is higher than the mean of the "independent" correlations (average r = .57), t = 5.31, p < .001 (see Figure 2a). This would suggest an average inflation of .12 – which is not insignificant, but hardly worthy of the label "voodoo." However, there is reason to believe that within this sample of correlations, the estimate of the inflation may itself be inflated.

Are "independent" correlations unbiased estimates?

The accuracy of correlation estimates relative to population values depends on the details of the study procedures in complex ways, and there are several potential sources of bias in the "independent" correlations that Vul et al. consider the gold standard. To illustrate this complexity, at least one known statistical effect causes many of the correlations in the "independent" analyses to be systematically *under*-estimated. Why would this be the case? Half of the "independent" correlations were computed on voxels or clusters selected from analyses of group-average contrast effects (e.g. voxels that were more active in task A than task B without regard for the individual difference variable). Because low variability is one of two factors that increase t-values, selecting voxels with high t-values for subsequent correlation analyses will tend to select voxels with low variability across subjects. This selection procedure restricts the range of the brain data and works against finding correlations with other variables.[3]

We re-analyzed the correlations in Vul et al's meta-analysis by (a) applying a correction for restricted range to the 58 correlations obtained using the procedure likely to result in restricted range, (b) including the previously omitted correlations, and (c) removing the three non-correlations that were mistakenly included in the original meta-analysis. "Independent" correlations based on anatomically-defined regions of interest do not have restricted range and thus were not corrected. Because we do not have access to the raw fMRI data from each of the surveyed studies, we estimated the full and restricted sample variances needed for the correction formula from one of our data sets and applied these variances to all of the "independent" correlations in the meta-analysis.[4]

In our re-analysis, there was no longer any difference between the "independent" (average r = .70) and the "non-independent" (average r = .69) correlation distributions, t = -0.57, p > .10 (see

---

[3] When a subsample has systematically lower variance than the full sample (i.e. restriction of range), correlations between the subsample and individual difference measures will produce correlation values that are smaller than the true correlation in the population (Thorndike, 1949). To give a simple analogy, imagine a correlation of .65 exists between age and spelling ability in 5 to 18 year olds. If we only sample 9 and 9.5 year olds, the observed correlation between age and spelling will be lower because we will have sampled from a restricted range of the age variable. Fortunately, the restriction of range effect can be corrected using the following formula from Cohen, Cohen, West, and Aiken (2003, p. 58) if the variance of the restricted sample and full sample are known.

$$\tilde{r}_{YX} = \frac{r_{YX_c}(sd_X / sd_{X_c})}{\sqrt{1 + r^2_{YX_c}\left(\left(sd^2_X / sd^2_{X_c}\right) - 1\right)}}$$

[4] For the full sample variance, we extracted data from a set of voxels distributed throughout the brain selected without consideration of t-test values. For the restricted sample variance, we extracted data from voxels with a significant group effect, as was typical of the "independent" studies. As expected, the average standard deviation in the full (2.82) and restricted samples (1.33) were significantly different from one another, t=4.63, p < .001.

Figure 2b).[5] Thus, when adjusted for restriction of range, the "independent" and "non-independent" samples of correlations do not support Vul et al.'s assertion of massive inflation. This should be seen as an exercise rather than a complete analysis, because we could not compute the variance for the full and restricted samples in each study, and because we did not attempt to take all other possible sources of bias into account. Indeed, calculating the bias in effect size would be at least as complex as determining a valid multiple comparisons corrections threshold, which requires detailed information about the data covariance structure in each study. Nevertheless, it does suggest that whatever inflation does exist may be far more modest and less troubling than Vul et al.'s characterization suggests.

Are such large correlations theoretically possible?

The upper limit on the observed correlation between two measures is constrained by the square root of the product of the reliabilities of the two measures. Vul et al. suggest that many "non-independent" correlations violate this theoretical upper limit. Based on a handful of non-social neuroscience studies that examined the reliability of fMRI data, Vul et al. provide estimates of what they believe a likely average reliability is for fMRI data (~.70). Similarly, they suggest that personality measures are likely to have reliabilities in the .70 - .80 range. Applying the products of the reliabilities formula, they conclude that the maximum upper bound for observable correlations is .74.

It is troubling that Vul et al. would make the bold claim that observed correlations from social neuroscience above .74 are "impossibly high" and above the "theoretical upper bound". This claim is based on a rough estimate of reliability that is then generalized across a range of measures. If we estimated that grocery store items cost, on average, about $3, would it then be theoretically impossible to find a $12 item? They make this claim despite the facts that (a) fMRI reliability has never been assessed for social neuroscience tasks; (b) if one is generalizing from previously measured reliabilities to measures with unknown reliability, it is the highest known reliabilities, not the average, that might best describe the theoretical maximum correlation; and (c) Vul et al. acknowledge in footnote 18 that some "independent" correlations are above .74 due to sampling fluctuations of observed correlations, an acknowledgement that should extend to the "non-independent" correlations.[6]

If we assume that brain regions in fMRI studies can have reliabilities above .90, as multiple studies have demonstrated, then the reliability of the individual difference measures actually used becomes critical. Consider, for example, the correlation (r = .88) between a social distress measure and activation in the dorsal anterior cingulate cortex during a social pain manipulation (Eisenberger, Lieberman, & Williams, 2003) that is singled out by Vul et al from the first page of their article. If one generically assumes that individual difference measures will all have reliabilities of .70 - .80, one would falsely conclude that the observed correlation in that study is not theoretically possible. However, multiple studies have reported reliabilities for this social distress measure between .92 and .98 (Oaten, Williams, Jones, & Zadro, 2008; Van Beest & Williams, 2006).[7] Applying

---

[5] Of several formulas considered for restricted range correction, the Cohen et al. (2003) formula that we used was the most conservative. Using Thorndike's formula (1949), the "independent" correlations actually become significantly higher than the "non-independent" correlations. Also, if only the correlations that Vul et al. included are used in the correction for restricted range analysis, the results are the same – there is no longer a significant difference between the samples.

[6] After correcting for restricted range, 46% of the "independent" correlations are above .74 and thus also violate Vul et al.'s theoretical upper bound.

[7] Given that one of the authors of the Vul et al. article emailed one of the authors of the Eisenberger et al. article about reliabilities for the social distress measure and further inquired specifically about one of the .92 reliabilities (K. D.

reliabilities of .90 for fMRI and .95 for the social distress measure yields a theoretical upper limit on the correlation of .92. Thus, by Vul et al.'s own criteria, a .88 correlation is theoretically possible in this case. This is just one example, but it points to the more general mistake of making claims about the theoretical upper bound of correlations based on approximate guesses of the measures' reliability.

Conclusions

Our reply has focused on several misconceptions in the Vul et al. paper that unfortunately have been sensationalized by the authors, as well as in the media, even prior to publication. Because social neuroscience has garnered a lot of attention in a short period of time, singling it out for criticism may make for better headlines. As this article makes clear, however, Vul et al.'s criticisms rest on shaky ground at best.

The substantive issues under consideration have nothing to do with social neuroscience specifically. Vul et al. describe a two-step inferential procedure that would be bad science if anyone did it, but as far as we know, nobody does[8]. As long as standard procedures for addressing the issue of multiple comparisons are applied in a reasonable sample size, large correlations will occur by chance only rarely, and most observed effects will reflect true underlying relationships. Vul et al.'s own meta-analysis suggests that the "non-independent" correlations are only modestly inflated, calling into question the use of labels such as "voodoo" and "untrustworthy." Finally, Vul et al. make incorrect inferences when attempting to use average expected reliabilities to inform on the theoretically possible observed correlations.

Ultimately, we should all be mindful that the effect sizes from whole-brain analyses are likely to be inflated, but confident in the knowledge that such correlations reflect meaningful relationships between psychological and neural variables to the extent that valid multiple comparisons procedures are used. There are various ways to balance the concerns of false positive results and sensitivity to true effects, and social neuroscience correlations use widely accepted practices from cognitive neuroscience. These practices will no doubt continue to evolve. In the mean time, we'll keep doing the science of exploring how the brain interacts with the social and emotional worlds we live in.

Williams, *personal communication*), it is disappointing that Vul et al. did not indicate that this .88 correlation was not violating the theoretical upper limit for this study.

[8] An important general lesson from this discussion is that post-hoc correlations will tend to be inflated—a statistical phenomenon understood since the 1800's—and should not be taken at face value as estimates of the correlation magnitude. As with any behavioral study of correlations, to quantify the exact magnitude of the predictive relationship of one variable on a second variable, cross-validation techniques should be used, as Vul et al. suggest. However, this valid point should not be taken as support for Vul et al.'s argument that the hypothesis-testing framework used to analyze brain-behavior correlations is flawed. This is not the case.

References

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied Multiple Regression/Correlational Analysis for the Behavioral Sciences*. Mahwah, NJ: Lawrence Earlbuam.

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

Kober, H., Barrett, L. F., Joseph, J., Bliss-Moreau, E., Lindquist, K., & Wager, T. D. (2008). Functional grouping and cortical-subcortical interactions in emotion: A meta-analysis of neuroimaging studies. *Neuroimage, 42*, 998-1031.

Nee, D. E., Wager, T. D., & Jonides, J. (2007). Interference resolution: Insights from a meta-analysis of neuroimaging tasks. *Cognitive, Affective, and Behavioral Neuroscience, 7*, 1-17.

Nichols, T. & Hayasaka, S. (2003). Controlling the familywise error rate in functional neuroimaging: A comparative review. *Statistical Methods in Medical Research, 12*, 419-446.

Oaten, M., Williams, K. D., Jones, A., & Zadro, L. (2008). The effects of ostracism on self-regulation in the socially anxious. *Journal of Social and Clinical Psychology, 27*, 471-504

Ochsner, K.N., Ludlow, D.H., Knierim, K., Hanelin, J., Ramachandran, T., Glover, G.C., & Mackey, S.C. (2006). Neural correlates of individual differences in pain-related fear and anxiety. *Pain,* 120, 69-77.

Park, R.L. (2000) *Voodoo Science: The Road from Foolishness to Fraud,* New York: Oxford University Press.

Price, D. D. (2000). Psychological and neural mechanisms of the affective dimension of pain. *Science, 288*, 1769-1772.

Singer, T., Seymour, B., O'Doherty, J., Kaube, H., Dolan, R., & Frith, C. D. (2004). Empathy for pain involves the affective but not sensory components of pain. *Science, 303,* 1157-1162.

Singer, T., Seymour, B., O'Doherty, J. P., Stephan, K. E., Dolan, R. J., & Frith, C. D. (2006). Empathetic neural responses are modulated by the perceived fairness of others. *Nature, 439,* 466-469.

Thorndike, R. L. (1949). *Personnel selection*. New York: John Wiley.

Van Beest, I., & Williams, K. D. (2006). When inclusion costs and ostracism pays, ostracism still hurts. *Journal of Personality and Social Psychology, 91,* 918–928.

van Snellenberg, J. X., & Wager, T. D. (in press). Cognitive and Motivational Functions of the Human Prefrontal Cortex. In E. Goldberg & D. Bougakov (Eds.), *Luria's Legacy in the 21st Century*. Oxford: Oxford University Press.

Vul, E., Harris, C., Winkielman, P., Pashler, H. (in press) Voodoo correlations in social neuroscience. *Perspectives on Psychological Science*.

Vul, E & Kanwisher, N (in press) "Begging the question: The non-independence error in fMRI data analysis." To appear in Hanson, S. & Bunzl, M (Eds.), Foundations and Philosophy for Neuroimaging.

Wager, T. D., Barrett, L. F., Bliss-Moreau, E., Lindquist, K., Duncan, S., Kober, H., et al. (2008). The Neuroimaging of Emotion. In M. Lewis, J. M. Haviland-Jones & L. F. Barrett (Eds.), *Handbook of Emotions* (3rd ed., pp. 249-271). New York: Guilford Press.

Wager, T. D., & Smith, E. E. (2003). Neuroimaging studies of working memory: a meta-analysis. *Cognitive, Affective, and Behavioral Neuroscience, 3*, 255-274.

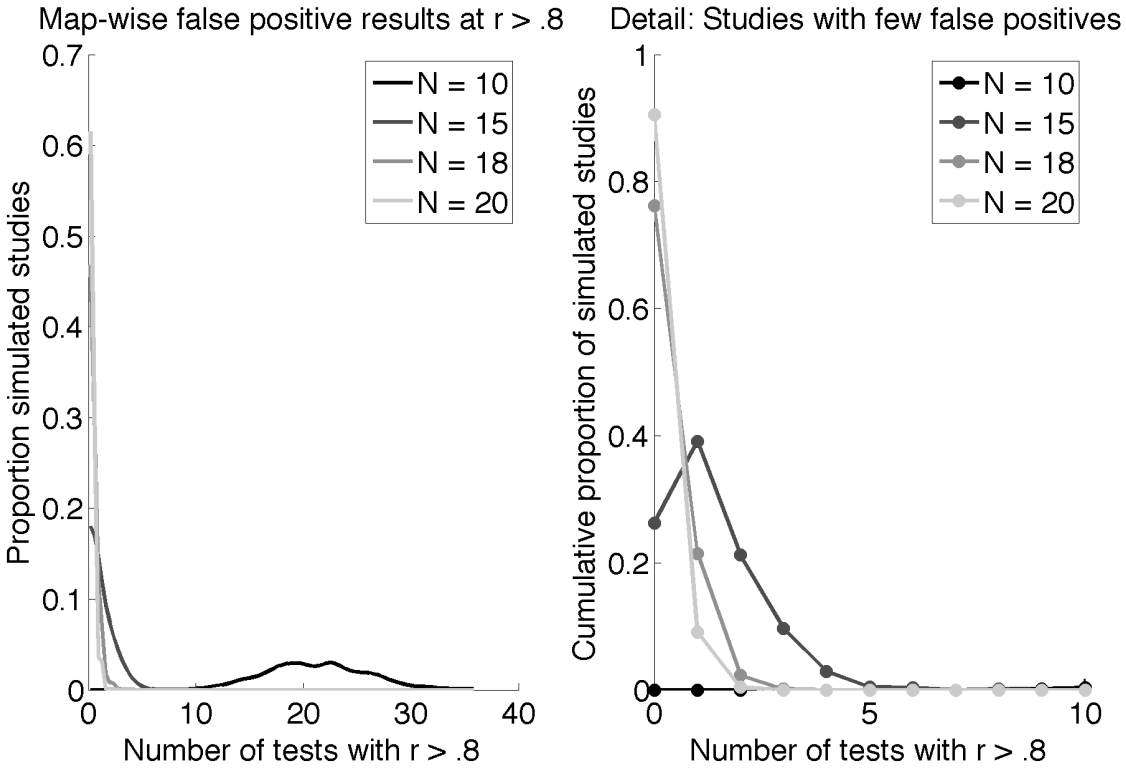
Figure Captions

Figure 1. A simulation of the number of high false positive correlations (correlations above 0.8) that might reasonably occur in a typical whole-brain regression analysis. We conducted 1,000 simulated whole-brain regression analyses in which brain and covariate values were independent Gaussian random variables. The procedures and assumptions are described in more detail below. The left panel shows a histogram of the number of simulated studies (y-axis) that yielded a given number of tests with $r > 0.8$ anywhere in the brain map (x-axis). Studies with $N = 10$ subjects, as in Vul et al.'s simulation, yielded high numbers of false positive tests (typically 15 to 25). Studies with $N = 18$ subjects (the mean of the criticized studies) yielded very few false positive results. The right panel shows details of the histogram between 0 and 10 false positive results. With $N = 18$, 76% of studies showed no false positive results at $r > .8$, 21% showed a single false-positive test, and 2% showed exactly two false-positive tests.

These results are illustrative rather than exact; the actual false positive rate depends on details of the noise structure in the data, and can be estimated using nonparametric methods on the full data set. The results presented here depend principally on the sample size (N), the number of effective independent tests (NEIT) performed in the whole-brain analysis, and standard assumptions of independence and normally distributed data. To estimate the NEIT, we used the p-value thresholds for 11 independent whole-brain analyses reported in Nichols and Hayasaka (2003) that yield $p < 0.05$ with familywise error-rate correction for multiple comparisons as assessed by Statistical Nonparametric Mapping software. We then equated this p-value threshold to a Bonferroni correction based on an unknown number of independent comparisons, and solved for the unknown NEIT for each study. Averaging over the 11 contrast maps yielded an average of 7768 independent comparisons. Individual studies may vary substantially from this average. Dividing the number of voxels in each map by the NEIT for each study and averaging yielded a mean of 25.3 voxels per test; thus, each false positive result can be thought of as a significant region encompassing 25 voxels.

Figure 2. Distribution of "independent" and "non-independent" correlations uncorrected and corrected for restriction of range, based on papers included in the meta-analysis by Vul et al. (A) A reconstruction of the correlations plotted in Figure 5 of Vul et al. Here correlations are plotted as a percentage of total correlations of each type. In this display, "non-independent" correlations (average r=.69) are inflated relative to the "independent" correlations (average r=.57) by an average of .12. (B) A re-analysis of the data from the studies included in the meta-analysis by Vul et al. "Independent" correlations using a procedure likely to result in restricted range issues were corrected; 52 correlations in the relevant papers that were omitted by Vul et al. were included; and 3 "correlations" that were not actually correlations were removed. In the re-analysis, the "non-independent" correlations (average r=.69) are no longer observed to be inflated relative to "independent" correlations (average r=.70).

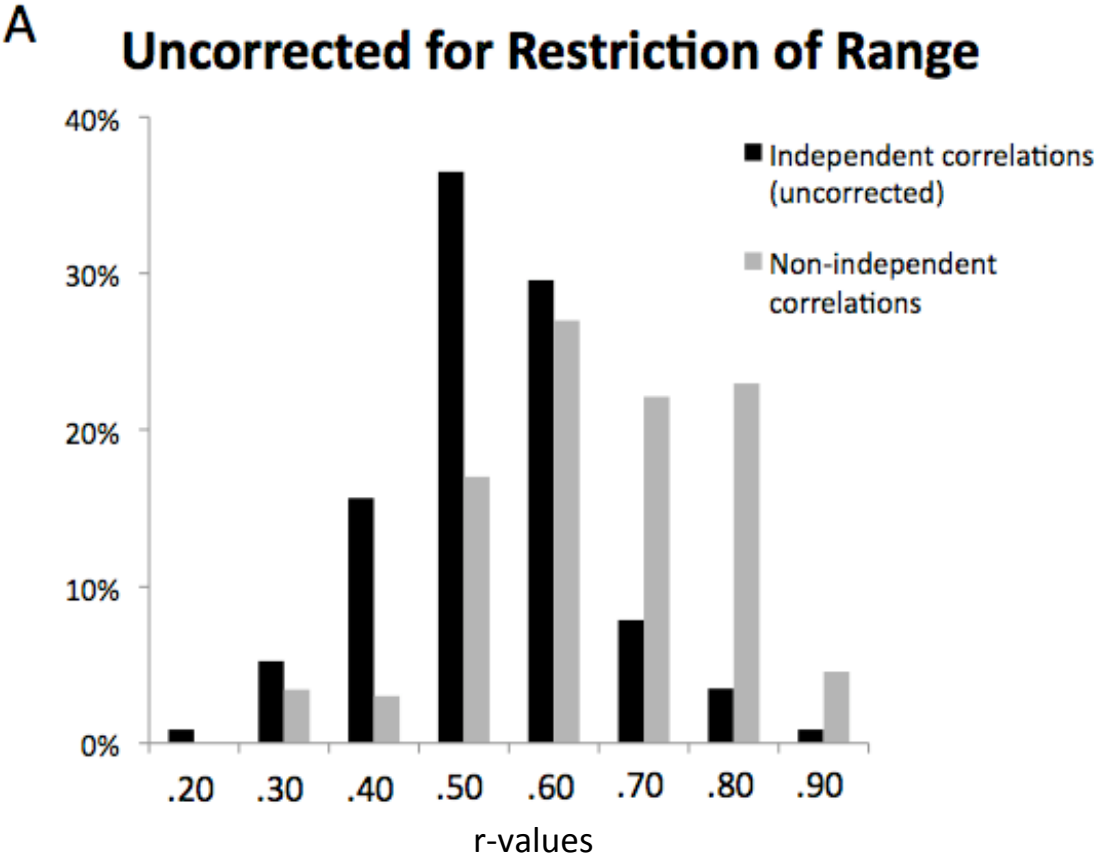
Figure 1.



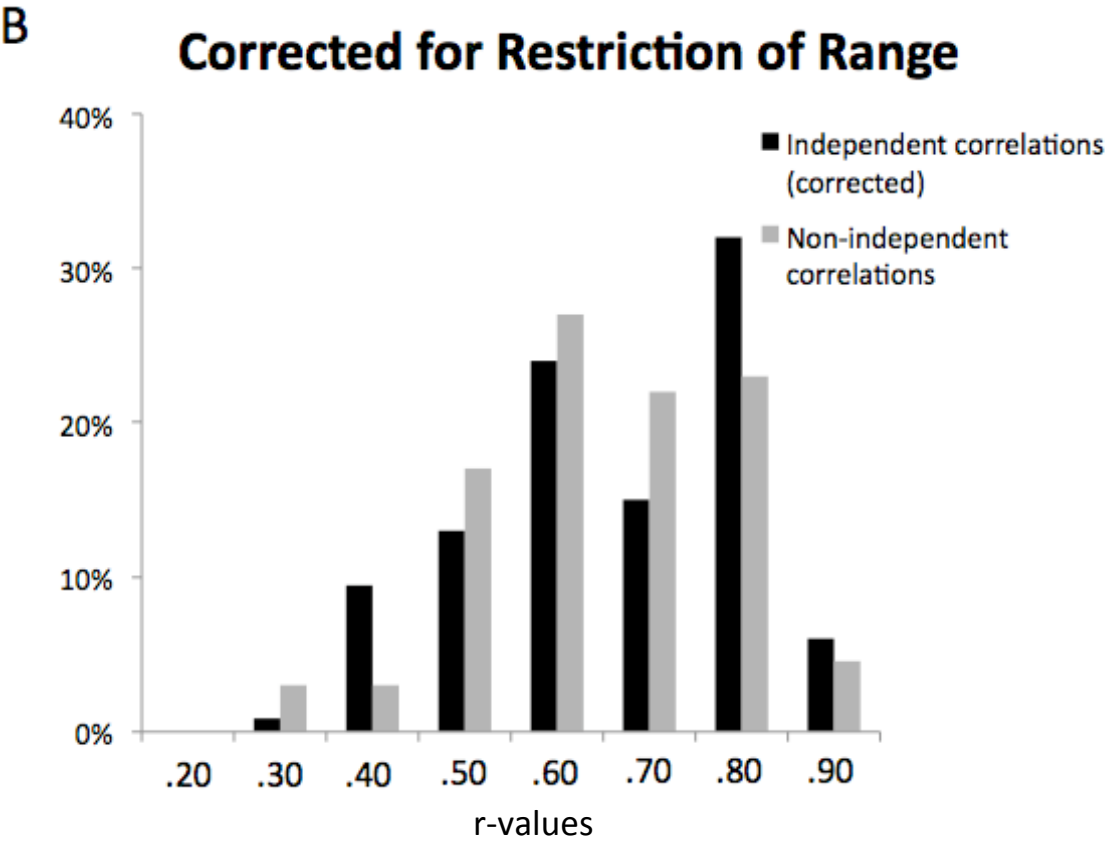
Likelihood that particular numbers of false positive tests will occur (at a threshold of r > 0.8)

Sample size	0	1	2	3	4	5	6 or more
15	26.3%	39.1%	21.2%	9.7%	2.9%	0.4%	0.4%
18	76.2%	21.4%	2.3%	0.1%	0.0%	0.0%	0.0%
20	90.5%	9.1%	0.4%	0.0%	0.0%	0.0%	0.0%

A


Uncorrected for Restriction of Range
40%
30%
20%
10%
0%
Independent correlations (uncorrected)
Non-independent correlations
.20
.30
.40
.50
.60
.70
.80
.90
r-values

B


Corrected for Restriction of Range
40%
30%
20%
10%
0%
Independent correlations (corrected)
Non-independent correlations
.20
.30
.40
.50
.60
.70
.80
.90
r-values

APPENDIX: Sampling errors in Vul et al. article

1. In study #4 (Ochsner et al., 2006), one "non-independent" correlations was not included in the analysis.

2. In study #6 (Eisenberger et al., 2003), three "correlations" were included in the analysis, that were not in fact correlations. For three of the main effect analyses comparing exclusion to inclusion, the authors reported an effect size r-statistic, along with t and p. No individual difference variable was involved in these analyses.

3. In study #7 (Hooker et al., 2008), three "independent" correlations were not included in the analysis.

4. In study #21 (Rilling et al., 2007), 35 "non-independent" correlations from Table 8 were not included and one other correlation from the manuscript was also not included. Although these correlations are listed as a table of r-values, it is conceivable that they were left out of the analysis because p-values were not presented. A simple calculation would have confirmed that with 22 subjects, nearly all of these correlations are significant at $p<.005$ (and most at $p<.001$) and thus met the sampling criteria.

5. In study #22 (Mobbs et al, 2005), 5 "non-independent" correlations were included in Figure 5. However, these correlations were calculated from ROI's obtained in a contrast analysis comparing two conditions and therefore should have been classified as "independent" correlations.

5. In study #31 (Singer et al., 2006), 4 "non-independent" correlations that are described in the text but given numerically in the supplementary materials (as indicated in the main text) were not included.

6. In study #39 (Posse et al., 2003), one "independent" correlation was not included in the analysis.

7. In study #45 (Leland et al., 2006), one "independent" correlation was not included in the analysis.

8. In study #53 (Kross et al., 2007), three "independent" correlations were not included in the analysis.