

Correlations Aren't Voodoo

Matthew D. Lieberman, Elliot T. Berkman, and Tor D. Wager

Correlations in Social Neuroscience Aren't Voodoo

Commentary on Vul et al. (2009)

Matthew D. Lieberman,¹ Elliot T. Berkman,¹ and Tor D. Wager²

¹*University of California, Los Angeles*, ²*Columbia University*

Address correspondence to Matthew Lieberman, Department of Psychology, University of California, Los Angeles, Los Angeles, CA 90095-1563; e-mail: lieber@ucla.edu.

ABSTRACT—Vul, Harris, Winkielman, and Pashler (2009, this issue) claim that many brain–personality correlations in fMRI studies are “likely...spurious” (p. xx), and “should not be believed” (p. xx). Several of their conclusions are incorrect. First, they incorrectly claim that whole-brain regressions use an invalid and “nonindependent” two-step inferential procedure, a determination based on a survey sent to researchers that only included nondiagnostic questions about descriptive process of plotting one’s data. We explain how whole-brain regressions are a valid single-step method of identifying brain regions that have reliable correlations with individual difference measures. Second, they claim that large correlations from whole-brain regression analyses may be the result of noise alone. We provide a simulation to demonstrate that typical fMRI sample sizes will only rarely produce large correlations in the absence of any true effect. Third, they 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, their own “meta-analysis” suggests that the magnitude of the bias is approximately .12— a rather modest bias.

In an article in this issue, Vul, Harris, Winkielman, & Pashler (2009, this issue) claim that brain–personality correlations in many social neuroscience studies and those in related fields are “implausibly high” (p. xx), “likely...spurious” (p. xx), and “should not be believed” (p. xx). The article was originally titled “Voodoo Correlations in Social Neuroscience” and was circulated widely in the scientific community, on the Internet, and in the popular press prior to publication. 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*. Though the title was subsequently changed to remove the word *voodoo*, the substance of the article and its connotations are unchanged: It is a pointed attack on social neuroscience. Much of the article’s prepublication impact was due to its aggressive tone, which is nearly unprecedented in the scientific literature and made it easy for the article to spread virally in the news. Thus, we felt it important to respond both to the tone and to the substantive arguments.

The trouble with the Vul et al. article is that it rests on a fundamental misconception about how statistical procedures are used in neuroimaging studies. They point out that post-hoc correlation estimates from whole-brain hypothesis testing procedures will tend to be greater than the true correlation value (this has been widely known but also widely underappreciated). However, they imply that post-hoc reporting of correlations constitutes an invalid inferential procedure, when in fact it is a descriptive procedure that is entirely valid. In addition, the quantitative claims that give their arguments the appearance of statistical rigor are based on problematic assumptions. Thus, it is ironic that Vul et al.’s article—which critiques social neuroscience as having achieved popularity in prominent journals and the press due to shaky statistical reasoning—itself achieved popularity based on problematic claims about the process of statistical inference. Our goal in this reply is to clarify the inferential procedures in question to set the record straight and to take a closer look at how conducting whole-brain correlation analyses might quantitatively impact correlation estimates.

DO WHOLE-BRAIN CORRELATIONS USE A “NONINDEPENDENT” TWO-STEP INFERENCE PROCEDURE?

Vul et al. (p. xx) 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 and suggest 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 nonindependence from the first inference.

The problem is that we know of no researchers who conduct their analyses this way. We were able to contact authors from 23 of the 28 nonindependent articles reviewed by Vul et al. Each of the contacted authors reported that they used a single-step inferential procedure, 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.

So what do these researchers actually do? When a whole-brain regression analysis is conducted, the goal is typically to identify regions of the brain whose activity shows a reliable nonzero 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 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 inferential step. For any particular sample size, the t and r values are merely redescrptions of the p values obtained in the one inferential step and provide no additional inferential information of their own.

The fact that Vul et al.'s questionnaire (see their Appendix A) only asks about plotting of correlations to determine whether a second inferential step has occurred is one of the primary sources of the misunderstanding that has emerged from their article. Vul et al. interpret plotting of data as a second inferential step, but this is incorrect. Plotting the correlation is a purely descriptive process, not an inferential process. Nevertheless, Vul clearly characterizes this as an example of the nonindependence error, "The most common, most simple, and most innocuous instance of nonindependence occurs when researchers simply plot (rather than test) the signal change in a set of voxels that were selected based on that same signal change" (Vul & Kanwisher, in press, p. 5). This statement implies that if a behavioral researcher correlated an outcome measure with extraversion, neuroticism, and psychopathy and found a significant relationship only with extraversion, then it would be an error to plot just the extraversion correlation. Although Vul et al. constructed the survey sent to authors with the intention of assessing which analyses used a second nonindependent inferential step, the questionnaire did not ask a single question about a second inferential step; it only asks about data plotting, which is nondiagnostic with respect to inferential methods.

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. Scatterplots provide an implicit check on underlying assumptions that must be met if any standard inferential

procedure is used. A correlation of $r = .7$ in a sample of 30 participants could, for example, be driven entirely by one or two outliers (constituting a violation of the normality assumption), and readers viewing the scatterplot would quickly see this and question the result. Thus, it is true that correlation scatterplots often look very compelling when r values are high, and they should not be taken as unbiased estimates of the population correlation coefficient, but they should be reported nonetheless.

In sum, despite Vul et al.'s characterizing whole-brain regressions as “seriously defective” (p. xx), 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 nonzero 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., overestimated 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 $N = 10$), one 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 spurious correlations might occur in a typical whole-brain regression analysis, we conducted a simulation (see Fig. 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 18 subjects (the average N was 18.25 in the 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 two or more false positive correlations. This suggests that in actual studies with similar properties and multiple comparison procedures, the great majority of reported effects of this magnitude reflect a true underlying relationship.

In addition, 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 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 in the same approximate brain areas, which is consistent with the notion of distributed networks underlying social and affective phenomena.

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, 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). It is important to note that 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 are results from 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 NONINDEPENDENT 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 nonindependent 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. xx) and “should not be believed” (p. xx). It is true that there is inflation in such correlations (though not because of any invalid inferential procedure), it would be useful to know just how inflated these correlations are in the social neuroscience findings they criticize.

Although it is impossible to know for sure, the “meta-analysis”¹ 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 to be acceptable independent procedures in green and so-called nonindependent (biased) correlations in red. The obvious conclusion to draw is that the nonindependent correlations have higher values than the gold-standard independent correlations, and thus they are systematically inflated.

To assess the average magnitude of the independent and nonindependent correlations, we collected all the articles 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 articles. We identified 54 correlations in the articles used in their meta-analysis that met their inclusion criteria, but were omitted from the meta-analysis without explanation. We also found three “correlations” in the meta-analysis that were really effect sizes associated with main effects rather than correlations (see the Appendix for a breakdown). Among the nonindependent correlations, almost 25% of the correlations reported in the original articles 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, would work against Vul et al.’s hypothesis of inflated correlations due to nonindependent correlation reporting (see Figure 2). In other words, the omitted correlations were not randomly distributed with respect to the group means, as would be expected from clerical errors. Of the 41 omitted nonindependent correlations, 38 had values lower than the mean of included nonindependent correlations. The mean of the omitted nonindependent correlations (.61) was significantly lower than the mean of the included nonindependent correlations (.69), $t(173) = 4.06, p < .001$. 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 mean of the included independent correlations (.57), $t(129) = 2.74$, $p < .01$. All three of the included nonindependent correlations that should have been omitted had values higher than the mean of the included nonindependent correlations.

Based solely on the correlations that Vul et al. included in their meta-analysis, the mean of the nonindependent correlations (average $r = .69$) is higher than the mean of the independent correlations (average $r = .57$), $t(254) = 5.31$, $p < .001$ (see Figure 3a). This would suggest an average inflation of .12, which is not insignificant, but hardly worthy of the attacks made by Vul et al. However, there are reasons to believe that the estimate of the inflation within this sample of correlations may itself be inflated.

One reason why independent correlations from region-of-interest (ROI) analyses will tend to be smaller on average than nonindependent correlations from whole-brain analyses has nothing to do with the validity of either method. The minimum reportable r value in a study depends on the p value threshold, which will typically differ between the ROI analyses (used to generate the independent correlations) and whole-brain analyses (used to generate the nonindependent correlations). If an ROI analysis is examining effects in two regions in a sample of 18 subjects, then the p value threshold is .025 for a corrected p value of .05, and thus the minimum reportable correlation would be an r of .51. In a whole-brain analysis of 18 subjects using a p value threshold of .005, the minimum reportable correlation is an r of .62, and at a p value threshold of .001, the minimum reportable correlation is an r of .69. Thus, a portion of the difference observed in their meta-analysis is due to these reporting constraints rather than the analytic method per se.

ARE INDEPENDENT CORRELATIONS UNBIASED ESTIMATES?

Although Vul et al. focus on potential bias in nonindependent correlations, another reason for mean differences in nonindependent and independent correlations is biases in the independent correlations. 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 to be the gold standard. To illustrate this complexity, we know of at least one statistical effect that causes many of the correlations in the independent analyses to be systematically underestimated. 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 in 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.²

We reanalyzed the correlations in Vul et al.'s meta-analysis by (a) applying a correction for restricted range to the 58 independent correlations obtained using the procedure likely to result in restricted range, (b) including the previously omitted correlations, and (c) removing the three noncorrelations 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.³

In our reanalysis, there was no longer any difference between the independent (average $r = .70$) and the nonindependent (average $r = .69$) correlation distributions, $t(304) = -0.57, p > .10$ (see

Figure 3b).⁴ Thus, when adjusted for restriction of range, the independent and nonindependent 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 as measured in a particular sample. Vul et al. suggest that many nonindependent correlations violate this upper limit on what should be observable. On the basis of a handful of 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” (p. xx) and above the “theoretical upper bound” (p. xx) of what can legitimately be observed. 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? Vul et

al. 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 observable; and (c) they acknowledge in Footnote 19 that some independent correlations are above .74 due to sampling fluctuations of observed correlations, an acknowledgement that should also extend to the nonindependent correlations.⁵

If we assume that brain regions in fMRI studies can have reliabilities above .90, as multiple studies have demonstrated (Aron, Gluck, & Poldrack, 2006; Fernandez et al., 2003), 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 et al., 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, then 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), a fact that Vul et al. were aware of.⁶ Applying reliabilities of .90 for fMRI and .95 for the social distress measure yields a theoretical upper limit on observable correlations at .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. article that unfortunately have been sensationalized by the authors and by the media 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.

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.⁷ They used a survey to assess which authors use this method, but they did not include any questions that would actually assess whether the nonindependence error had occurred. 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 nonindependent correlations are only modestly inflated, calling into question the use of labels such as “spurious” and “untrustworthy.” Finally, Vul et al. make incorrect assumptions 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 meantime, we'll keep doing the science of exploring how the brain interacts with the social and emotional worlds we live in.

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, Ray Dolan, 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, Antonio Rangel, Steve Reise, James Rilling, David Sander, Ajay Satpute, Sophie Schwartz, Tania Singer, Thomas Straube, Hidehiko Takahashi, Shelley Taylor, Alex Todorov, Patrik Vuilleumier, Paul Whalen, and Kip Williams.

REFERENCES

- Aron, A.R., Gluck, M.A., & Poldrack, R.A. (2006). Long-term test-retest reliability of functional MRI in a classification learning task. *NeuroImage*, *29*, 1000–1006.
- Cohen, J., Cohen, P., West, S.G., & Aiken, L.S. (2003). *Applied multiple regression/correlational analysis for the behavioral sciences*. Mahwah, NJ: Erlbaum.
- Eisenberger, N.I., Lieberman, M.D., & Williams, K.D. (2003). Does rejection hurt? An fMRI study of social exclusion. *Science*, *302*, 290–292.
- Fernández, G., Specht, K., Weis, S., Tendolkar, I., Reuber, M., Fell, J., et al. (2003). Intrasubject reproducibility of presurgical language lateralization and mapping using fMRI. *Neurology*, *60*, 969–975.
- Hooker, C.I., Verosky, S.C., Miyakawa, A., Knight, R.T., & D’Esposito, M. (2008). The influence of personality on neural mechanisms of observational fear and reward learning.

Neuropsychologia, 466, 2709–2724.

- 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.
- Kross, E., Egner, T., Ochsner, K., Hirsch, J., & Downey, G. (2007). Neural dynamics of rejection sensitivity. *Journal of Cognitive Neuroscience*, 19, 945–956.
- Leland, D., Arce, E., Feinstein, J., & Paulus, M. (2006). Young adult stimulant users increased striatal activation during uncertainty is related to impulsivity. *NeuroImage*, 33, 725–731.
- Mobbs, D., Hagan, C.C., Azim, E., Menon, V., & Reiss, A.L. (2005). Personality predicts activity in reward and emotional regions associated with humor. *Proceedings of the National Academy of Sciences, USA*, 102, 16502–16506.
- 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.
- Posse, S., Fitzgerald, D., Gao, K., Habel, U., Rosenberg, D., Moore, G.J., & Schneider, F. (2003).

- Real-time fMRI of temporolimbic regions detects amygdala activation during single-trial self-induced sadness. *NeuroImage*, *18*, 760–768.
- Price, D.D. (2000). Psychological and neural mechanisms of the affective dimension of pain. *Science*, *288*, 1769–1772.
- 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.
- 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.
- Tukey, J.W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.
- 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, United Kingdom: Oxford University Press.
- Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Voodoo correlations in social neuroscience. *Perspectives on Psychological Science*, *4*, xx–xx.
- Vul, E., & Kanwisher, N. (in press). Begging the question: The non-independence error in fMRI data analysis. In S. Hanson & M. Bunzl (Eds.), *Foundations and philosophy for neuroimaging*. Cambridge, MA: MIT Press.

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.

Fig. 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 left panel shows a histogram of the number of simulated studies (y axis) that yielded a given number of tests in which $r > 0.8$ anywhere in the brain map (x axis). Studies with 10 subjects, as in Vul et al.'s simulation, yielded high numbers of false positive tests (typically 15 to 25). Studies with 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 18 participants, 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 $p < .05$ with family-wise 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 7,768 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.

Fig. 2. Distribution of correlations in papers surveyed by Vul et al. but omitted from their meta-analysis. A: Independent correlations that were omitted from the Vul et al. meta-analysis. The dotted line indicates the mean of independent correlations (.57) that were included in their meta-analysis. Twelve of the 13 omitted independent correlations were higher than this mean. B: Nonindependent correlations that were omitted from the Vul et al. meta-analysis. The dotted line indicates the mean of nonindependent correlations (.69) that were included in their meta-analysis. Thirty-eight of the 41 omitted nonindependent correlations were lower than this mean.

Fig. 3. Distribution of independent and nonindependent 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. Correlations are plotted as a percentage of total correlations of each type. In this display, nonindependent correlations (average $r = .69$) are inflated relative to the independent correlations (average $r = .57$) by an average of .12. B: A reanalysis 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 reanalysis, the nonindependent correlations (average $r = .69$) are no longer observed to be inflated relative to independent correlations (average $r = .70$).

¹Although Vul et al. characterize their review as a meta-analysis, their selection of studies for inclusion appears biased and nonreproducible. The selection of studies includes articles with large correlations that Vul et al. were likely aware of prior to sampling the literature (i.e., those papers that brought the issue of large correlations to their attention). If Vul et al. knew the magnitude of the correlations in these articles and then chose search terms guaranteed to include these in the meta-analysis, this would seem to be the kind of sampling bias that Vul et al. accuse others of. In addition, the selection of studies in their review is not reproducible. Vul et al. indicate that they searched for “social terms (e.g., *jealousy*, *altruism*, *personality*, *grief*)” (p. x), which is obviously an incomplete description. However, just to take one example, we searched for *altruism* and found several other fMRI papers on empathy from the time period covered by the Vul et al. review that were omitted from the meta-analysis for no discernable reason. Given that a number of these studies replicate the Singer et al. (2004) findings, it again raises questions about the selective inclusion of studies in their review.

²When 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:

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

³For 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, $t(48) = 4.63, p < .001$.

⁴Of 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 nonindependent correlations. Also, if we only use the correlations that Vul et al. included in the correction for restricted range analysis, the results are the same—there is no longer a significant difference between the samples.

⁵After correcting for restricted range, 46% of the independent correlations are above .74 and thus also violate Vul et al.'s theoretical upper bound.

⁶One of the authors of the Vul et al. article emailed one of the authors of the Eisenberger et al. (2003) article about reliabilities for this social distress measure prior to the submission of their manuscript and further inquired specifically about one of the .92 reliabilities (K. D. Williams, personal communication, January 17, 2009). Consequently, it is disappointing that Vul et al. did not indicate that this .88 correlation was not violating the theoretical upper limit for this study.

⁷An important general lesson from this discussion is that post-hoc correlations will tend to be inflated—a statistical phenomenon understood since the 1800s—and should not be taken at face value as estimates of the correlation magnitude. As with any behavioral study of correlations, one should use cross-validation to quantify the exact magnitude of the predictive relationship of one

variable on a second variable, 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.

APPENDIX: SAMPLING ERRORS IN THE VUL ET AL. (2009) META-ANALYSIS

1. In Study 4 (Ochsner et al., 2006), one nonindependent correlation was not included in the analysis.
2. In Study 6 (Eisenberger et al., 2003), Vul et al. included three “correlations” 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, Verosky, Miyakawa, Knight, & D’Esposito, 2008), three independent correlations were not included in the analysis.
4. In Study 21 (Rilling et al., 2007), 35 nonindependent 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, Hagan, Azim, Menon, & Reiss, 2005), five nonindependent correlations were included in Figure 5. However, these correlations were calculated from ROIs obtained in a contrast analysis comparing two conditions, and they should have therefore been classified as independent correlations.

5. In Study 31 (Singer et al., 2006), four nonindependent correlations that are described in the text were not included, though they were listed numerically in the supplementary materials (as indicated in the main text).
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, Egner, Ochsner, Hirsch, & Downey, 2007), three independent correlations were not included in the analysis.