Voodoo Correlations in Social Neuroscience

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

[1]Massachussetts Institute of Technology

[2]University of California, San Diego

[*]to whom correspondence should be addressed: hpashler@ucsd.edu

In Press, Perspectives on Psychological Science

Dec. 23, 2008

ACKNOWLEDGMENTS. Phil Nguyen provided invaluable assistance with literature review and management of the survey of researchers reported here, and Shirley Leong provided capable assistance with data management and analysis. We thank all the researchers who responded to our questionnaire. This work was supported by the National Institute of Mental Health (grant P50 MH0662286-01A1), Institute of Education Sciences (Grants R305H020061 and R305H040108 to H. Pashler), the National Science Foundation (Grant BCS-0720375 to H. Pashler; Grant SBE-0542013 to G. Cottrell), and a collaborative activity grant from the James S. McDonnell Foundation.

The authors gratefully acknowledge comments and suggestions from Chris Baker, Jon Baron, Hart Blanton, John Cacioppo, Max Coltheart, Danny Dilks, Victor Ferreira, Timothy Gentner, Michael Gorman, Alex Holcombe, David Huber, Richard Ivry, James C. Johnston, Nancy Kanwisher, Brian Knutson, Niko Kriegeskorte, James Kulik, Hans Op de Beeck, Russ Poldrack, Anina Rich, Seth Roberts, Rebecca Saxe, Jay Schulkin, John Serences, Mark Williams, John Wixted, Steven Yantis, and Galit Yovel.

Abstract

The newly emerging field of Social Neuroscience has drawn much attention in recent years, with high-profile studies frequently reporting extremely high (e.g., >.8) correlations between behavioral and self-report measures of personality or emotion and measures of brain activation obtained using fMRI. We show that these correlations often exceed what is statistically possible assuming the (evidently rather limited) reliability of both fMRI and personality/emotion measures. The implausibly high correlations are all the more puzzling because social-neuroscience method sections rarely contain sufficient detail to ascertain how these correlations were obtained. We surveyed authors of 54 articles that reported findings of this kind to determine the details of their analyses. More than half acknowledged using a strategy that computes separate correlations for individual voxels, and reports means of just the subset of voxels exceeding chosen thresholds. We show how this non-independent analysis grossly inflates correlations, while yielding reassuring-looking scattergrams. This analysis technique was used to obtain the vast majority of the implausibly high correlations in our survey sample. In addition, we argue that other analysis problems likely created entirely spurious correlations in some cases. We outline how the data from these studies could be reanalyzed with unbiased methods to provide the field with accurate estimates of the correlations in question. We urge authors to perform such reanalyses and to correct the scientific record.

A Puzzle: Remarkably High Correlations in Social Neuroscience

The field of social neuroscience (or social cognitive neuroscience, as it is also sometimes referred to) scarcely existed 10 years ago, and yet the field has already achieved a remarkable level of attention and prominence. Within the space of a few years, it has spawned several new journals (*Social Neuroscience, Social Cognitive and Affective Neuroscience)*, and is the focus of substantial new funding initiatives (National Institute of Mental Health, 2007), lavish attention from the popular press (Hurley, 2008) and the trade press of the psychological research community (e.g., *APS Observer,* Fiske, 2003). Perhaps even more impressive, however, is the number of papers from social neuroscience that have appeared in such prominent journals as *Science, Nature,* and *Nature Neuroscience.*

While the questions and methods used in social neuroscience research are quite diverse, a substantial number of widely cited papers in this field have reported a specific type of empirical finding that appears to bridge the divide between mind and brain; extremely high correlations between measures of individual differences relating to personality, emotionality and social behavior, and measures of brain activity obtained with functional magnetic resonance imaging (fMRI). We focus on

social neuroscience[1] here because this was the area where these correlations came to our attention; we have no basis for concluding that the problems discussed here are necessarily any worse in this area than in some other areas.

To take but a few examples of many studies that will be discussed below:

Eisenberger, Lieberman, and Williams (2003), writing in *Science*, described a game they created to expose individuals to social rejection in the laboratory. The authors measured the brain activity in 13 individuals at the same time as the actual rejection took place, and later obtained a self-report measure of how much distress the subject had experienced. Distress was correlated at r=.88 with activity in the anterior cingulate cortex (ACC).

In another *Science* paper, Singer et al. (2004) found that the magnitude of differential activation within the ACC and left insula induced by an empathy-related manipulation was correlated between .52 and .72 with two scales of emotional empathy (the Empathic Concern Scale of Davis, and the Balanced Emotional Empathy Scale of Mehrabian).

Writing in *NeuroImage*, Sander et al. (2005) reported that a subject's proneness to anxiety reactions (as measured by an index of the Behavioral Inhibition System; Carver and White, 1994) correlated at r=.96 with the difference in activation of the right cuneus to attended versus ignored angry speech.

In the review below, we will encounter many studies reporting similar sorts of correlations.

The work that led to the present article began when the present authors became puzzled about how such impressively high correlations could arise. We describe our efforts to resolve this puzzlement, and the conclusions that our inquiries have led us to.

Why should it be puzzling to find high correlations between brain activity and social and emotional measures? After all, if new techniques of social neuroscience are providing a deeper window on the link between brain and behavior, does it not make sense that researchers should be able to find the neural substrates of individual traits—and thus potentially bring to light stronger relationships than have often been found in purely behavioral studies?

The problem is this: It is a statistical fact (first noted by researchers in the field of classical psychometric test theory) that the strength of the correlation observed between measures A and B ($r_{ObservedA,ObservedB}$) reflects not only the strength of the relationship between the traits underlying A and B ($r_{A,B}$), but also the reliability of the measures of A and B ($Reliability_A$ and $Reliability_B$, respectively). In general,

$$r_{ObservedA,ObservedB} = r_{A,B} * \text{sqrt}(reliability_A * reliability_B)$$

Thus, the reliabilities of two measures provide an upper bound on the possible

[1] Social neuroscience relies on a variety of methodologies,including neuroimaging (e.g., fMRI, PET), patient studies (e.g., lesions), electrophysiology (e.g., EEG and EMG), animal research (e.g., cross-species comparisons), neuroendocrine, and neuroimmunological investigations (Harmon-Jones & Winkielman, 2007).

correlation that can be observed between the two measures (Nunnally, 1970)[2].

Reliability Estimates

So what are the reliabilities of fMRI and personality/emotional measures likely to be[3]? The reliability of personality and emotional scales varies between measures, and according to the number of items used in a particular assessment. However, test-retest reliabilities as high as .8 seem to be relatively uncommon, and usually found only with large and highly refined scales. Viswesvaran and Ones (2000) surveyed many studies on the reliability of the Big Five factors of personality, and concluded that the different scales have reliabilities ranging from .73 to .78. Hobbs and Fowler (1974) carefully assessed the reliability of the sub-scales of the MMPI, and found numbers ranging between .66 and .94, with an average of .84. In general, therefore, a range of .7 - .8 would seem to be a somewhat optimistic estimate for the smaller and more ad hoc scales used in much of the research described below, which could well have substantially lower reliabilities.

Less is known about the reliability of blood oxygenation level dependent (BOLD) signal measures in fMRI, but some relevant studies have recently been performed[4]. Kong et al. (2006) had subjects engage in six sessions of a finger tapping task while recording brain activation. They found test-retest correlations of the change in BOLD signal ranging between 0 and .76 for the set of areas that showed significant activity in all sessions[5]. Manoach et al (2001, their figure 1, p. 956) scanned subjects on two sessions of performance on the Sternberg memory scanning task, and found reliabilities ranging between .23 to .93, averaging .60. Aron, Gluck, and Poldrack (2006) had people perform a classification learning task on two separate occasions widely separated in time, and found voxel-level reliabilities with modal values (see their figure 5, p. 1005) a little bit below .8[6]. Johnstone et al. (2005, p. 1118) examined the stability of amygdala BOLD response to presentations of fearful faces in multiple sessions. Intraclass correlations for left and right amygdale regions of interest were in the range of .4 to .7 for the 2 sessions separated by 2 weeks. Thus, from the

---

[2] This is the case because the correlation coefficient is defined as the ratio between the covariance of two measures and the product of their standard deviations: $r_{x,y} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$. Real-world measurements will be corrupted by (independent) noise, thus the standard deviations of the measured distributions will be increased by the additional noise (whose magnitude is assessed by the measure's reliability). This will make the measured correlation lower than the true underlying correlation, by a factor equal to the geometric mean of reliabilities.

[3] We consider test-retest reliabilities here (rather than inter-item, or split-half reliability) because, for the most part, the studies we discuss gathered behavioral measure at different points in time than the fMRI data. In any case, internal reliability measures, like coefficient alpha, do not generally appear to be much higher in this domain.

[4] We focus here on studies that look at the reliability of BOLD activation measures, rather than the reliability of patterns of voxels exceeding specific thresholds, which tend to be substantially lower (e.g., Stark et al., 2004).

[5] It seems likely that restricting the reliability analysis to regions consistently active in all sessions would tend to overestimate the reliability of BOLD signal in general.

[6] They found somewhat higher reliabilities for voxels within a frontostriatal system that they believed was most specifically involved in carrying out the probabilistic classification learning.

literature that does exist, it would seem reasonable to suppose that fMRI measures computed at the voxel level will not often have reliabilities greater than about .7.

The Puzzle

This, then, is the puzzle. Measures of personality and emotion evidently do not often have reliabilities greater than .8. Neuroimaging measures seem typically to be reliable at .7 or less. If we assume that a neuroimaging study is performed in a case where the underlying correlation between activation in the brain area and the individual difference measure (i.e., the correlation that would be observed if there were no measurement error) is *perfect*[7] then the highest possible meaningful correlation that could be obtained would be sqrt(.8 * .7), or .74. Surprisingly, correlations exceeding this upper bound are often reported in recent social neuroscience literature.

Meta-Analysis Methods

We turned to the original papers to find out how common these remarkable correlations are, and what analyses might be yielding them. Unfortunately, after a brief review of several articles, it became apparent that the analyses employed varied greatly from one investigator to the next, and the exact methods were simply not made clear in the typically brief and sometimes opaque method sections.

To probe the issue further, we conducted a survey of the investigators. Our focus was confined to social neuroscience because this is the place where the remarkably-high correlations first drew our attention, and because they seem most prevalent here; however, we would not want the reader to think that any of the issues examined here are unique to this area. We proceeded as follows: First, we attempted to pull together as complete a sample as we could readily achieve of the social neuroscience literature reporting correlations between evoked BOLD activity and behavioral measures of individual differences in personality, emotionality, social behavior, and related domains (generally excluding psychopathological symptoms, however). Then we emailed the authors of the articles we identified, sending a brief survey to determine how the reported correlation values were computed.

Literature Review

Our literature review was conducted using the keyword "fMRI" (and variants), in conjunction with a list of social terms (e.g., "jealousy", "altruism", "personality", "grief", etc.). Within the articles retrieved by these searches, we selected all the articles we could find that reported across-subject correlations between a trait measure and evoked BOLD activity. This resulted in 54 articles, with 256 significant correlations between BOLD signal and a trait measure. It should be emphasized that we do not suppose this literature review

[7] There are several reasons why a true correlation of 1.0 seems highly unrealistic. First, for any behavioral trait, it is far-fetched to suppose that only one brain area influences this trait. Second, even if the neural underpinnings of a trait *were* confined to one particular region, it would seem to require an extraordinarily favorable set of coincidences for the BOLD signal (basically a blood flow measure) assessed in one particular stimulus or task contrast to capture *all function* relevant to the behavioral trait, which after all reflects the organization of complex neural circuitry residing in that brain area.

to be exhaustive. Undoubtedly we missed some papers reporting these kinds of numbers, but our sample seems likely to be quite representative, perhaps slanted toward papers that appeared in higher impact journals.

A histogram of these significant correlations is displayed in Figure 1. It can be seen that correlations in excess of .75 are plentiful indeed.

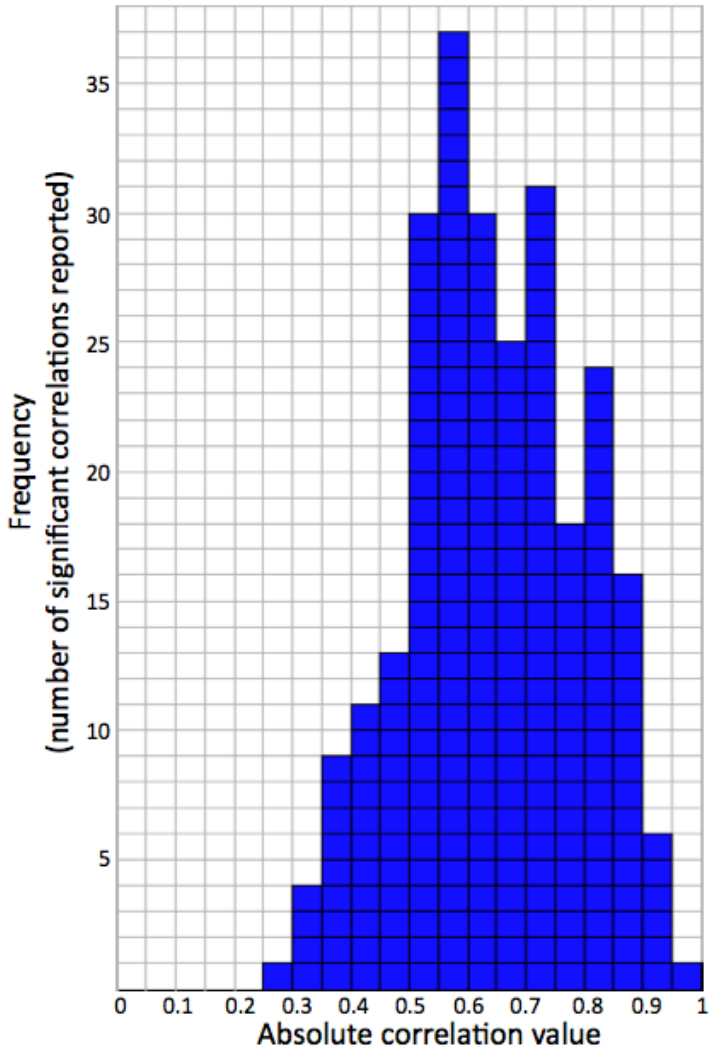

Figure 1: A histogram of the correlations between evoked BOLD response and behavioral measures of individual differences seen in the studies identified for analysis in the current article.

We turn next to the question: where do these numbers come from? Before doing so, we have to provide a bit of background for readers unfamiliar with methods in this area.

Elements of fMRI Analysis

For those not familiar with fMRI analysis, the essential steps in just about any neuroimaging study can be described rather simply (those familiar with the techniques may wish to skip this section). The output of an fMRI experiment typically consists of two types

of "3D pictures" (*image volumes*): "anatomical" (a high resolution scan that shows anatomical structure, not function) and "functional". Functional image volumes are lower resolution scans showing measurements reflecting, among other things, the amount of deoxygenated hemoglobin in the blood – blood oxygenation level dependent (BOLD) signal. A functional image volume is composed of many measurements of the BOLD signal in small, roughly cube-shaped, regions called "voxels" ('volumetric pixels'). The number of voxels in the whole image volume depends on the scanner settings, but it typically ranges between 10x64x64 and 30x128x128 voxels. Thus, each functional image contains somewhere between 40,000 and 500,000 voxels, with each of these voxels covering between 1 mm^3 (1x1x1 mm) and 125 mm^3 (5x5x5mm) of brain tissue (except for voxels outside of the brain). A new functional image volume is usually acquired every 2 or 3 seconds (TR, or repetition time) during a scan, so one ends up with a timeseries of these functional images.

These data are typically preprocessed to reduce noise and to allow comparisons between different brains. The preprocessing usually includes smoothing (averaging each voxel with its neighbors, weighted by some function that falls with distance, such as a Gaussian). The studies we focus on here ultimately compute correlations across subjects: in this kind of study, the voxels are usually mapped onto an *average brain* (although not always, e.g., Yovel & Kanwisher, 2005). A number of average-brain models exist, the most famous being Talairach (Talairach & Tournoux, 1988) and MNI (Evans et al. 1993), but some investigators compute an average brain model for their particular subjects, and normalize their functional image scans onto that model.

Following pre-processing, some measure of the activation in a given voxel needs to be derived to assess if it is related to what the person is doing, seeing, or feeling. The simplest procedure is just to extract the average activation in the voxel while the person does a task. However, because any task will engage most of the brain (from visual cortex to see the stimulus, to motor cortex to produce a response, and everything in between), fMRI researchers typically focus not on the activation in particular voxels during one task, but rather on a *contrast* between the activation arising when the person performs one task versus the activation arising when they do another. This is usually measured as follows: while functional images are being acquired, the subject does a mixed sequence of two different tasks (A,B,B,A,A,B,A, and so forth—where A might be *reading words* and B might be *looking at nonlinguistic patterns*). Thus, the experimenter ends up with two different time series to compare: the sequence of tasks the person performed and, separately for each voxel, the sequence of activation levels measured at that voxel. A regression analysis can now be performed to ask: "is this voxel's activity different when the subject was performing Task A compared to Task B"?

These basic steps common to most fMRI data analyses yield matrices consisting of tens or hundreds of thousands of numbers indicating activation levels. These can be (and indeed generally are) displayed as images. However, to obtain quantitative summaries of these results and do further statistics on them (such as correlating them with behavioral measures—the topic of the present article), an investigator must somehow select a subset of voxels and

aggregate measurements across them. This can be done in various ways. A subset of voxels in the whole brain image may be selected based on purely anatomical constraints (e.g., all voxels in a region generally agreed to represent the amygdala, or all voxels within a certain radius of some *a priori* specified brain coordinates). Alternatively, regions can be selected based on "functional constraints": meaning voxels are selected based on their activity pattern in functional scans. For example, one could select all the voxels for a particular subject that responded more to reading than to non-linguistic stimuli. Finally, voxels could be chosen based on some combination of anatomy and functional response.

In the papers we are focusing on here, the final result, as we have seen, was always a correlation value—a correlation between each person's score on some behavioral measure, and some summary statistic of their brain activation. The latter summary statistic reflects the activation or activation contrast within a certain set of voxels. In either case, the critical question is: how was this set of voxels selected? As we have seen, voxels may be selected based on anatomical criteria, functional criteria, or both. Within these broad options, there are a number of additional more fine-grained choices. It is hardly surprising, then, that brief method sections rarely suffice to describe how the analyses were done in adequate detail to really understand what choices were being made.

Survey methods

To learn more than the Method sections of these papers disclosed about the analyses that yielded these correlations, we emailed the corresponding authors of these articles. The exact wording of our questions is included in Appendix 1, but we often needed to send customized follow-up questions to figure out the exact details when the survey questions were misunderstood, or did not match our reading of the methods section.

In our survey we first inquired whether the fMRI signal measure that was correlated across subjects with a behavioral measure represented the average of some number of voxels, or instead, the activity from just one voxel that was deemed most informative (referred to as the peak voxel).

If it was the average of some number of voxels, we inquired about how those voxels were selected – asking whether they were selected based only on anatomy, only on the activation seen in those voxels, or both?

If activation was used to select voxels, or one voxel was determined to be most informative based on its activation, we asked what was the measure of activation used. Was it the difference in activation between two task conditions computed on individual subjects, or was it a measure of how this task contrast correlated with the individual difference measure?

Finally, if functional data were used to select the voxels, were they the same functional data as were used to define the reported correlation?

Survey participants

Of the 55 articles we found in our review, we received methodological details from 52, and 3 did not respond to repeated requests.

Survey Results

We display the raw results from our survey as the proportion of studies that investigators described with a particular answer to each question (Figure 2). Since

some questions only applied to a subset of participants, we display only the proportion of the relevant subset of studies.

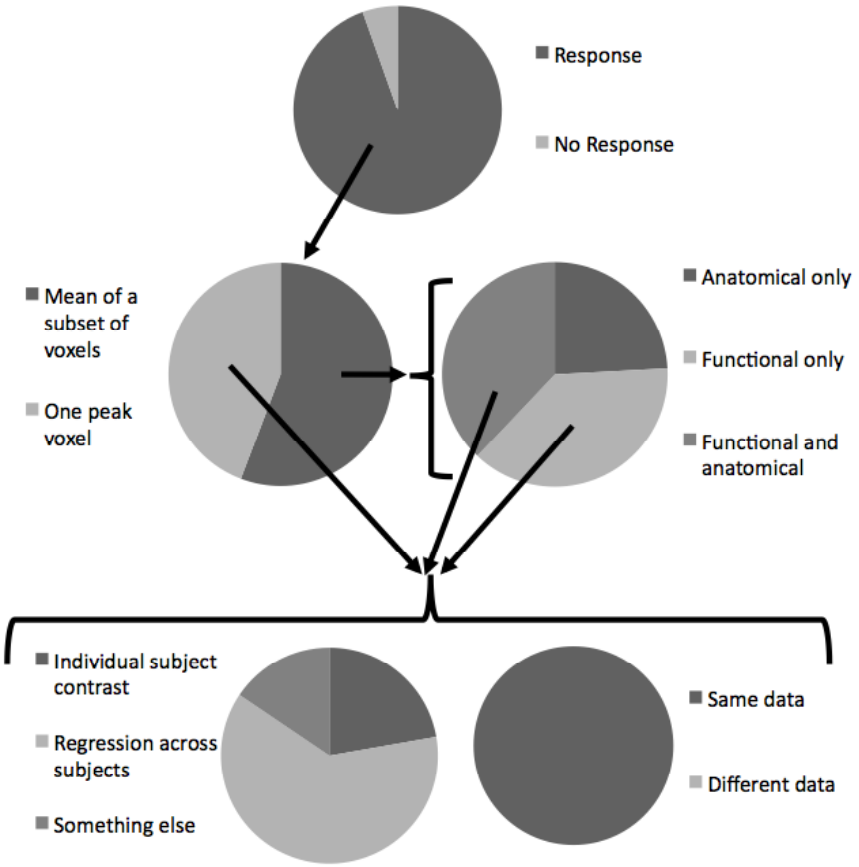

Figure 2. The results of our survey of social neuroscience individual-difference correlation methods. Of the 55 articles surveyed, the authors of 52 provided responses. Of those, 23 reported a correlation between behavior and one peak voxel; 29 reported the mean of a number of voxels. For those that reported the mean of a subset of voxels, 7 defined this subset purely anatomically, 11 used only functional constraints, and 11 used anatomical and functional constraints. Of the 45 studies that used functional constraints to choose voxels (either for averaging, or for finding the 'peak' voxel), 10 said they used functional measures defined within a given subject, 28 used the across-subject correlation to find voxels, and 7 did something else. All of the studies using functional constraints used the same data to select voxels, and then to measure the correlation. Notably, 54% of the surveyed studies selected voxels based on a correlation with the behavioral individual-differences measure, and then used those same data to compute a correlation within that subset of voxels.

The raw answers to our survey do not by themselves explain how the (implausibly high, or so we have argued) correlations were arrived at. The key, we believe, lies in the 54% of respondents who said that "regression across subjects" was the functional constraint

used to select voxels: indicating that voxels were selected because they correlated highly with the behavioral measure of interest.[8].

Figure 3 shows very concretely the sequence of steps that these respondents reported following in analyzing their data. A separate correlation across subjects was performed for each voxel within a specified brain region. Each correlation relates some measure of brain activity in that voxel (which might be a difference between responses in two tasks or in two conditions) with the behavioral measure for that individual. Thus, the number of correlations computed was equal to the number of voxels (meaning that in many cases, thousands of correlations were computed). At the next stage, the set of voxels for which this correlation exceeds some threshold were selected, and some measure of the relationships *for the voxels that exceed this threshold* was reported.

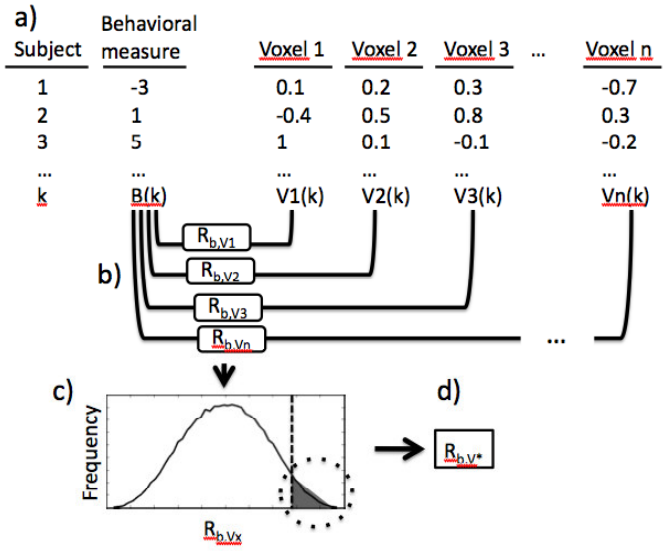

Figure 3: An illustration of the analysis employed by 54% of the papers surveyed. (a) From each subject, the researchers obtain a behavioral measure as well as BOLD measures from many voxels. (b) The activity in each voxel is correlated with the behavioral measure of interest across subjects. (c) From this set of correlations, researchers select those voxels that pass a statistical threshold, and (d) aggregate the fMRI signal across those voxels to derive a final measure of the correlation of BOLD signal and the behavioral measure.

[8] It is important to note that all of these studies also reported using the same data to compute the correlation as were initially used to select the subset of voxels.

What are the implications of selecting voxels in this fashion? Such an analysis will inflate observed across-subject correlations, and can even produce significant measures out of pure noise. The problem is illustrated in the simple simulation displayed in Figure 4: (a) investigator computes a separate correlation of the behavioral measure of interest with each of the voxels. Then, (b) those voxels that exhibited a sufficiently high correlation (passing a statistical threshold) are selected. Then an ostensible measure of the 'true' correlation is aggregated from the voxels that showed high correlations (e.g., by taking the mean of the voxels over the threshold). With enough voxels, such a biased analysis is guaranteed to produce high correlations even if none are truly present (Figure 4). Moreover, this analysis will produce visually pleasing scattergrams (e.g., Figure 4c) that will provide (quite meaningless) reassurance to the viewer that s/he is looking at a result that is solid, "not driven by outliers", etc.

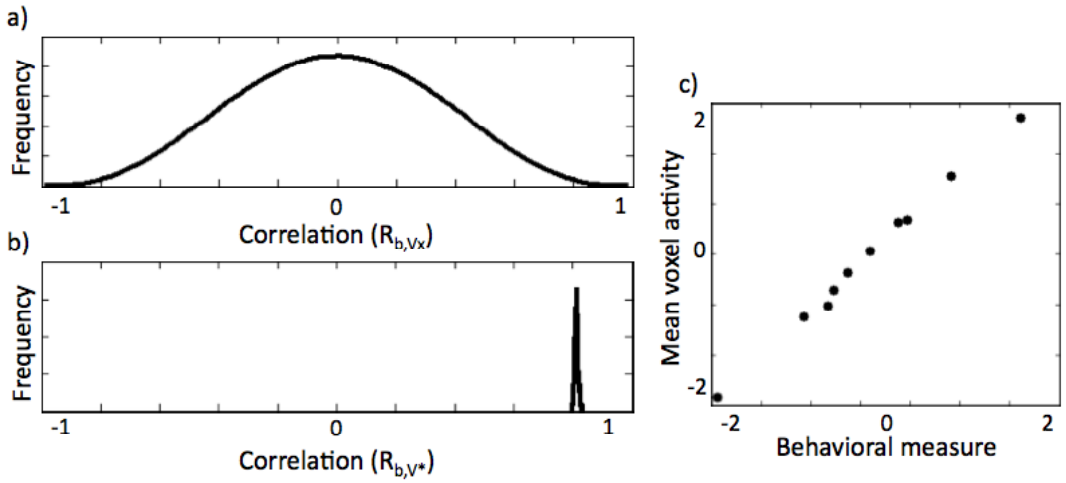

Figure 4: A simulation of a non-independent analysis on pure noise data (for similar exercises in other neuroimaging domains see Baker, Hutchison, et al, 2007; Simmons et al, 2006; Kriegeskorte et al, 2008). We simulated 1000 experiments each with 10 subjects and 10000 voxels, and one individual difference measure. Each subjects' voxel activity and behavioral measure were independent 0-mean Gaussian noise. Thus, (a) the true distribution of correlations between the behavioral measure and simulated voxel activity is distributed around 0, with random fluctuations resulting in a distribution that spans the range of possible correlations. (b) When a subset of voxels are selected for passing a statistical threshold (a positive correlation with $p<0.01$), the observed correlation of the mean 'activity' of those voxels is very high indeed. (c) If the BOLD activity from that subset of voxels is plotted as a function of the behavioral measure, a compelling scattergram may be produced.

The non-independence error

The fault seen in glaring form in Figure 4 will be referred to henceforth as the *non-independence error.* This approach amounts to selecting one or more voxels based on a functional analysis, and then reporting the results of the same analysis and functional data from just the selected voxels. This analysis distorts the results by selecting noise exhibiting the effect being searched for, and any measures obtained from such a non-independent analysis are biased and untrustworthy (for a formal discussion see Vul & Kanwisher, in press).

It may be easier to appreciate the gravity of the non-independence error by transposing it outside of neuroimaging. We (the authors of this paper) have identified a weather station whose temperature readings predict daily changes in the value of a specific set of stocks with a correlation of r=-0.87. For $50.00, we will provide the list of stocks to any interested reader. That way, you can buy the stocks every morning when the weather station posts a drop in temperature, and sell when the temperature goes up. Obviously, your potential profits here are enormous. But you may wonder: how did we find this correlation? The figure of -.87 was arrived at by separately computing the correlation between the readings of the weather station in Adak Island, Alaska, with each of the 3315 financial instruments available for the New York Stock Exchange (through the Mathematica function FinancialData) over the 10 days that the market was open between November 18th and December 3rd, 2008. We then averaged the correlation values of the stocks whose correlation exceeded a high threshold of our choosing, thus yielding the figure of -.87. Should you pay us for this investment strategy? Probably not: Of the 3,315 stocks assessed, some were sure to be correlated with the Adak Island temperature measurements simply by chance – and if we select just those (as our selection process would do), there was no doubt we would find a high average correlation. Thus, the final measure (the average correlation of a subset of stocks) was not independent of the selection criteria (how stocks were chosen): this, in essence, is the non-independence error. The fact that random noise in previous stock fluctuations aligned with the temperature readings is no reason to suspect that future fluctuations can be predicted by the same measure, and one would be wise to keep one's money far away from us, or any other such investment advisor[9].

Variants of the non-independence error occur in many different types of neuroimaging studies and in many different domains. The non-independence error is by no means confined to social neuroscience, nor to studies correlating individual behavioral differences with evoked fMRI activity. (For broader discussions of how non-independent analyses produce misleading results in other domains, see: Vul & Kanwisher, in press, Kriegeskorte et al, 2008; Baker, Hutchinson, et al, 2007; Baker, Simmons, et al 2007; Simmons et al, 2006).

Our survey allows us to determine which of the studies were committing variants of the non-independence error by finding analyses in which researchers selected voxels (answered A or B to question 1) based on correlation with the across-

[9] See Taleb (2004) for a sustained and engaging argument that this error, in subtler and more disguised form, is actually a common one within the world of market trading and investment advising.

subject behavioral measure of interest (answered B or C to question 2, and B to question 3), then plotted or reported the observed correlations from just those voxels (answered A to question 4).

Results and Discussion

For maximum clarity, we will present the results of our survey, and our overall analysis of what it means for the social neuroscience literature, in the form of a number of questions and answers.

A. Are the correlation values reported in this literature meaningful?

Of the 52 articles we successfully surveyed, 28 provided responses indicating that a non-independent analysis, like the one portrayed in Figures 3 and 4, was used to obtain the across-subject correlations between evoked BOLD activity and a measure of individual differences. As we saw in Figure 4, a non-independent analysis systematically distorts any true correlations that might exist. Thus, in half of the studies we surveyed, the reported correlation coefficients mean almost nothing, because they are systematically inflated by the biased analysis. The magnitude of this distortion depends upon variables (such as the number of voxels within the brain, noise and signal variance, etc.) which a reader would have no way of knowing, so it is not possible to correct for it. The problem is exacerbated in the case of the 38% of our respondents who reported the correlation of the *peak voxel* (the voxel with the highest observed correlation) rather than the average of all voxels in a cluster passing some threshold.

Figure 5 shows the histogram of correlation values with which our investigation started, this time color-coded by whether or not such a non-independent analysis was employed in the article. It is reassuring to see that the mode of independently acquired (i.e., valid) correlation values (coded green) is indeed below the 'theoretical upper bound' we anticipated from classical test theory and the limited information we have on test reliability (described in the introduction). The overwhelming trend is for the larger correlations to be emerging from non-independent analyses that are statistically guaranteed to inflate the measured correlation values.

In looking at Figure 5, it is tempting to assume that the non-independent (red) correlations, had they been measured properly, would have values around the central tendency of the independent (green) correlations (around .6). Thus, one might say, "it is very unfortunate that the numbers were seriously exaggerated, but the real relationships here are still pretty impressive." In our view, any such inference is unwarranted; many of the real relationships are probably far lower than the ones shown in green. After all, the published studies reporting independent measures of correlations are still predominantly those that found significant effects (resulting in the well known publication bias for significant results; cf. Ioannidis, 2005), and correlations much lower than .5 would often not have been significant with these sample sizes. We would speculate that, properly measured, many of the "red correlations" would have been far lower still, and may not exist at all. (For a discussion of the relationship between the non-independence error and the use of spatial clustering thresholds, see Appendix 2.)

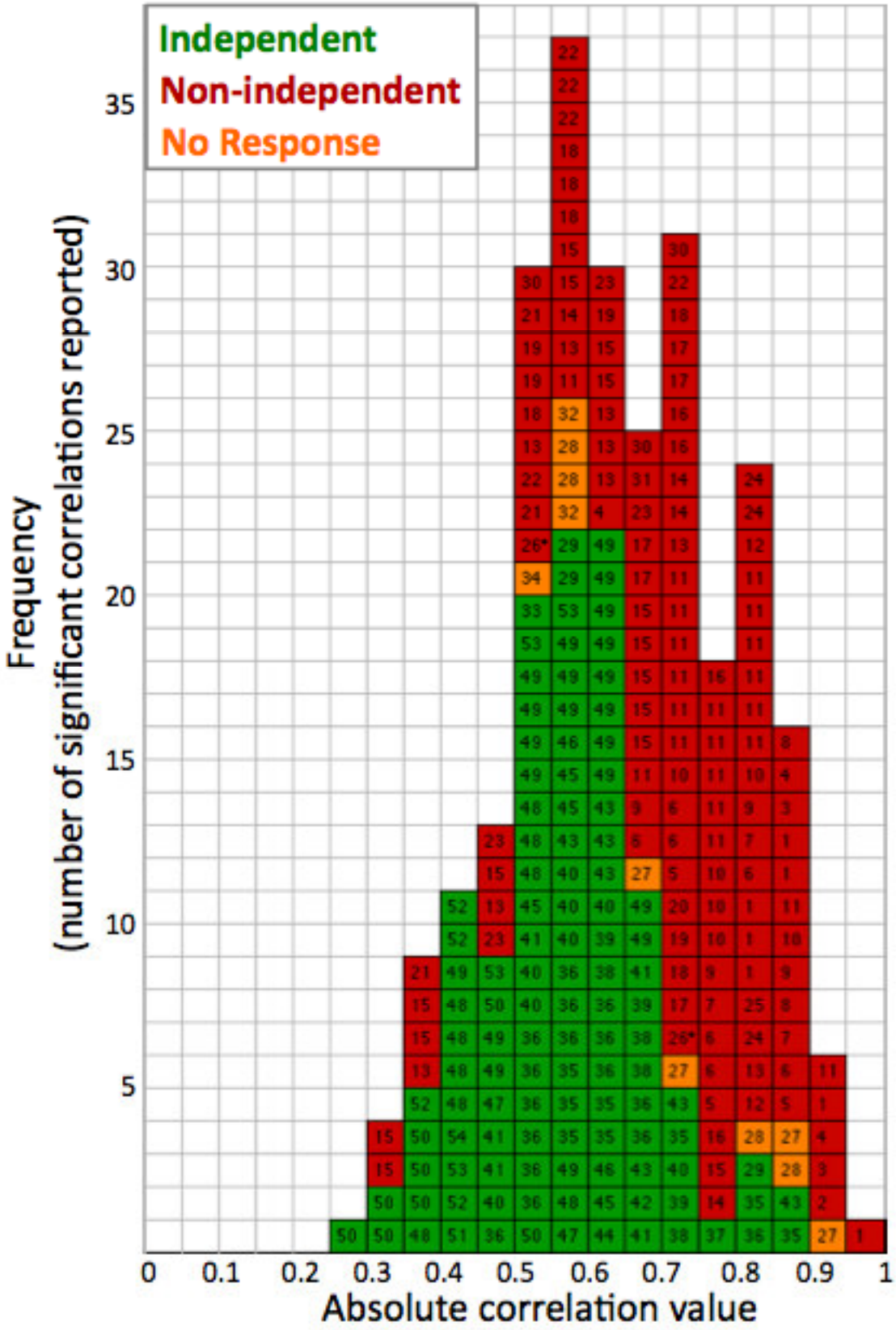
Independent
Non-independent
No Response
Frequency
(number of significant correlations reported)
35
30
25
20
15
10
5
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Absolute correlation value

*Figure 5. The histogram of the correlations values from the studies we surveyed (same data as Figure 1), this time, color-coded by whether or not the article from which this analysis originated used non-independent analyses. Correlations coded in green correspond to those that were achieved with independent analyses, avoiding the bias described in this paper. However, those in red correspond to the 54% of articles surveyed that reported conducting non-independent analyses – these correlation values are certain to be inflated. Entries in orange arise from papers whose authors chose not to respond to our survey. (See Table 1 below for key to article numbers; * study 26 carried out a slightly different, non-independent analysis: instead of explicitly selecting for a correlation between IAT and activation, they split the data into two groups, those with high IAT scores and those with low IAT scores, they then found voxels that showed a main effect between these two groups, and then computed a correlation within those voxels. This procedure is also non-independent, and will inflate correlations.)*

B. Is the problem being discussed here anything different than the well-known problem of multiple comparisons raising the probability of false alarms?

Every fMRI study involves vast numbers of voxels, and comparisons of one task to another involve computing a t-statistic and comparing it to some threshold. When numerous comparisons are made, adjustments of threshold are needed, and are commonly employed. The conventional approach involves finding voxels that exceed some arbitrarily high threshold of significance on a particular contrast (e.g., reading a word versus looking at random shapes). This multiple comparisons correction problem is well known and has received much attention.

The problem we describe arises when authors then report *secondary* statistics on the data in the voxels that were selected originally. In the case discussed in the present article, correlations are both the selection criterion and the secondary statistic.

When people compare reading a word versus reading a letter, and find brain areas with a t value of 13.2 (with 11 degrees of freedom, comparable to an r of .97, or an effect size of d=2.4), few people would interpret the t value as a measure of effect size. On the other hand, in the case of the r values under discussion here, we would contend that essentially everyone interprets them in that way.

C. What may be inferred from the scattergrams often exhibited in connection with non-independent analyses?

Many of the papers reporting biased correlation values display scattergram plots of evoked activity as a function of the behavioral measure. These plots are presumably included in order to show the reader that the correlation is not being driven by a few outliers, or by other aberrations in the data. However, when non-independent selection criteria are used to pick out a subset of voxels, the voxels passing this criterion will inevitably contain a large admixture of noise favoring the correlation (see the scattergram in Figure 4c for an example of a case where the relationship is pure noise). Thus, the shape of the resulting scattergrams provides no reliable indication about the nature of the

possible correlation signal underlying the noise, if any.

D. How can these same methods produce no correlations?

It may be as surprising to some readers, as it was to us, that a few papers reporting extraordinarily high correlations arrived at through non-independent analyses also reported some negative results (correlations that failed to reach significance). If the same analysis methods were applied to each correlation investigated, shouldn't the same correlation-amplifying bias apply to each one?

Indeed it should normally do so. However, with a bit of investigation, we were able to track down the source of (at least some of) the inconsistency: in certain papers, the bias inherent in non-independent analyses was sometimes wielded selectively, in such a way as to inflate certain correlations, but not others.

Take for instance Takahashi et al (2006), reporting an interaction in the presence of a correlation between evoked BOLD activity and rated jealousy in men and women: activity in the *insula* correlated with self-reported jealousy about emotional infidelity in men ($r=0.88$), but not women ($r=-0.03$). The opposite was true of activity in the *posterior STS* correlated with such self-reported jealousy in women ($r=0.88$), but not men ($r=-0.07$). At first blush, the scattergrams and correlations exhibit a very striking interaction (reported as significant at $p<0.001$). However, the insula activity corresponds to the peak voxel of a cluster that passed statistical threshold *for the correlation between rated jealousy and BOLD signal in males*; thus the observed correlation with rated jealousy in males was *non-independent* and biased, while the same correlation for rated jealousy in females was independent. The pSTS activity was *selected for correlating with rated jealousy in females*, and thus only the jealousy correlation in males was independent in that region.

It should come as no surprise, therefore, that such non-independently selected data produced a striking interaction in which the non-independent analyses showed high correlations while the independent analyses showed no correlation. Thus, the presence of the interaction, along with the magnitude of the correlations themselves, is quite meaningless and could have been obtained with completely random data like those utilized in the simulation shown in Figure 4.

E. But is there really any viable alternative to doing these non-independent analyses?

It is all very well to point out ways in which research methods fall short of the ideal. However, the ideal experiment and the ideal analysis are often out of reach, especially in fields like psychology and cognitive neuroscience. Perhaps we must settle for somewhat imperfect designs and methods to get any information whatsoever about across-subject brain-behavior correlations: Are any better methods available?

We contend that the answer is a clear-cut "Yes". These kinds of brain-behavior linkages can be readily investigated with designs that do not invite any of the rather disastrous complications that accompany the use of non-independent analyses.

One method is to select the voxels comprising different regions of interest

in a principled way that is "blind" to the correlations of those voxels with the behavioral measure and also mindful of the fact that individuals' brains are far from identical. For instance, to assess the relationship between ACC activity during exclusion and reactions to social rejection measured in a questionnaire, one would first put the social rejection data aside, and not "peek" at it while analyzing the fMRI data. The researcher can then define regions of interest in individual subjects in whatever way seems appropriate; e.g., by identifying voxels within the anatomical confines of the ACC that were significantly active for the excluded-included contrast (or, even better, using a different contrast, or different data, altogether). Once a subset of voxels is defined within an individual subject, one number should be aggregated from these voxels (e.g., the mean signal change). Only then are the behavioral data examined, and an unbiased correlation can be computed between the ACC region of interest and the behavioral measure. This method was used by a few of the authors of the current studies, e.g., Kross et al (2007). In addition to providing an unbiased measure of any relationships between evoked activity and individual differences, this 'functional Region of Interest' (fROI) method avoids implausible assumptions about voxel-wise correspondence across different individuals' functional anatomy[10] (Saxe, Brett, & Kanwisher, 2006).

If one feels that it makes sense to draw voxelwise correspondences between the functional anatomy of one subject and another, a second alternative exists: a 'split half' analysis. Here, half of the data are used to select a subset of voxels exhibiting the correlation of interest, and the other half of the data are used to measure the effect (examining the same voxels, but looking at different runs of the scanner). For example, if there are 4 runs in the social exclusion and 4 runs in the neutral condition, one can use 2 exclusion runs and 2 neutral runs to identify voxels that maximize the correlation, and *then test the correlation of the behavioral trait with these same voxels--but looking only at the other 2 runs.* Such a procedure uses *independent* data for voxel selection and the subsequent correlation test, and thus avoids the non-independence error[11].

[10] Although it is possible for voxels registered to the 'average brain' to be functionally matched across subjects, the variability in anatomical location of well-studied regions even in early visual cortex (V1, MT) and visual cognition (FFA) suggests to us that higher-level functions determining individual differences in personality and emotionality is not likely to be anatomically uniform across individuals (Saxe, Brett, & Kanwisher, 2006).

[11] At first blush, one might worry that using only half of the data to select the correlated regions will greatly decrease statistical power. However, there are two reasons why this should not be a concern. First, removing half of the data from each subject does not reduce the number of data-points that go into the across-subject correlation – it simply makes the estimate of BOLD activity for an individual subject more noisy (by a factor of sqrt(2)). This is not as detrimental to the ability to evaluate a correlation as reducing the number of data points. Second, stringent corrections for multiple comparisons are unnecessary for an independent split-half analysis, thus, a (reasonable) liberal threshold may be chosen to select the subset of voxels that correlate with the behavioral measure in the first half of the data. The statistical inference relies on the magnitude of the correlation observed in those voxels in the second half of the data – a single comparison, which will have ample power to detect any effect that may be close to significant in a properly corrected whole-brain analysis. For an even more data-efficient (but computationally intensive) independent validation technique, variants of the 'k-fold'

This straight-forward analysis may be computed on all of the suspect results noted in our paper thus far, and can be used to provide unbiased estimates of the correlations reported in these papers. Techniques of this kind (hold out validation and cross-validation) are used in a variety of fields (including fMRI) to evaluate the generality of conclusions when over-fitting is a possibility (Geisser, 1993) – as is the case when picking a small subset of many measured correlations as a measure of the true correlation.

It may often be advisable to use both of the methods just described, because they may find slightly different kinds of (real) patterns in the data. The first type of analysis focuses on the voxels that are most active in the task contrast at issue. This is a sensible place to look first to find relationships with individual differences. However, it is possible that the behavioral individual differences may be most closely associated with activity in some subset of voxels which may not show the greatest activity in this contrast. For example, it is possible that within the ACC there could be neural structures whose magnitude of response is related to rejection, even if the mean activation in those structures across subjects does not differ from zero.

F. Even if correlations were overestimated due to non-independent analyses, can't we at least be sure the correlations are statistically significant (and thus that there exists a real, nonzero, correlation)?

In most of the nonindependent analyses, the voxels included in the computation of the reported correlation were those that passed a threshold for significance that was based on some combination of the correlation value for each voxel and the spatial contiguity between the voxel and other elevated voxels--a threshold that typically included some ostensible adjustment for multiple comparisons. Given that, can we not be sure that there is a real, albeit weaker-than-reported, correlation? In principle, this ought to be the case – but only if the correction for multiple comparisons is appropriately implemented.

We did not explicitly survey the authors about their multiple comparisons correction procedures, but we do see evidence that the corrections used in this literature may often be less than trustworthy. The most common method of correcting for multiple comparisons used in this literature is family-wise error correction relying on "minimum cluster size thresholds"[12]. In this approach, the correlation in clusters of voxels is determined to be significant if the cluster contains a sufficiently large number of contiguous voxels each exceeding some statistical threshold. This procedure "relies on the assumption that areas of true neural activity will tend to stimulate signal changes over contiguous pixels" (Forman et al., 1995), i.e., "signal" will tend to show up as activity that extends beyond a single voxel, whereas statistical noise will generally be independent from one voxel to its neighboring voxel and thus will not usually appear in large clusters[13].

method can also be used (Brieman & Spector, 1992).

[12] See Appendix 2 for a discussion of whether the problem of inflated correlations is eliminated by the use of a cluster-based threshold.

[13] Technically, the rationale is somewhat more complicated and relies on estimates of the spatial correlations known to be present in the voxels

Given particular scan parameters[14], one can use various sophisticated techniques to compute the probability of falsely detecting a cluster of voxels (Type I error). This probability may be estimated using the *AlphaSim* tool from the program AFNI (Analysis for Functional NeuroImaging)[15] (Cox, 1996; Douglas Ward). We noticed that many papers in our sample chose p-thresholds of 0.005 and cluster size thresholds of 10, and stated that these choices were made relying upon Forman et al. (1995) as an authority. For instance, Eisenberger, et al. (2003) claimed that their analysis had a per-voxel false positive probability of "less than 0.000001." They used these thresholds on 19x64x64 imaging volumes at 3.125x3.125x4 mm, smoothed with 8 mm full-width at half-max Gaussian kernel. We were puzzled that these parameters would be able to reduce the rate of false alarms to the degree claimed, and so we investigated using AlphaSim. According to the *AlphaSim* simulations, pure noise data is likely to yield a cluster passing this threshold in nearly 100% of all runs (a per-voxel false alarm probability of 0.002)! To hold the false detection probability for a particular cluster below .000003 (thus keeping the overall probability of a false positive in the analysis below the commonly desired alpha level of 0.05), a far larger cluster size (namely, 56 voxels) would need to be used[16]. Thus, we suspect that the .000001 figure cited by Eisenberger et al. (2003) and other authors actually reflects a misinterpretation of Forman's simulations results[17]. It seems that ostensible corrections for multiple comparisons with the cluster size method are at least sometimes misapplied, and thus, even the statistical significance of some correlations in this literature may be questionable.

(e.g., due to smoothing). The smoothness assumption defines how likely it is for pure noise observations with these spatial statistics to contain clusters with a particular number of contiguous voxels exceeding statistical threshold.

[14] These parameters include: voxel dimensions, volume dimensions, smoothing parameter (sometimes data smoothness as estimated from the data), minimum cluster size, and minimum single-voxel p-threshold.

[15] The method used by AlphaSim differs subtly from that in SPM: AlphaSim allows users to estimate the smoothness of the data by entering the smoothing kernel – thus 'smoothness' amounts to the degree to which data were smoothed. In contrast, SPM computes a measure of 'smoothness' by measuring the spatial correlation in the data in addition to the smoothing parameter applied. Thus, simply entering the smoothing kernel into AlphaSim *underestimates* the smoothness of the data, and *underestimates* the probability of a falsely detected cluster. For our purposes, this means that the numbers obtained from AlphaSim will actually underestimate how large the clusters must be to reach a certain false alarm probability.

[16] Even if the brain occupied just one tenth of the imaging volume (7,700 voxels), the parameters described would falsely detect a cluster 60% of the time in pure noise – in this case, the appropriate minimum cluster size threshold would need to be 27, rather than 10, to reach a false detection rate of 0.05.

[17] The per-voxel false detection probabilities described by Eisenberger et al (and others) seem to come from Forman et al.'s Table 2C. Values in Forman et al's table report the probability of false alarms that cluster *within a single 2D slice* (a single 128x128 voxel slice, smoothed with a FWHM of 0.6*voxel size). However, the statistics of clusters in 2D (a slice) are very different from those of a 3D volume: there are many more opportunity for spatially clustering false alarm voxels in the 3D case, as compared to the 2D case. Moreover, the smoothing parameter used in the papers in question was much larger than 0.6*voxel size assumed by Forman in Table 2C (in Eisenberger et al., this was >2*voxel size). The smoothing, too, increases the chances of false alarms appearing in larger spatial clusters.

In general, it is important to keep in mind what statistics the *conclusions* of a particular paper rely on. In many papers, a liberal threshold is used to select an ROI (one that would be insufficiently conservative to address the multiple comparisons problem), and then *an independent* secondary statistic is computed on the ROI voxels. The conclusions of such papers usually rest on the secondary statistic computed within the ROI; what threshold was used to select the ROI voxels does not really matter. In the cases we discuss in this paper, *the secondary statistics are non-independent*, and are thus biased and meaningless. In these cases, the criteria used to *select* voxels becomes the *only* statistic which may legitimately be used to evaluate the results, and thus the selection criteria are of utmost importance for the conclusions of the paper.

It should be emphasized that we certainly do not contend that problems with corrections for multiple comparisons exist in all (or even a majority) of the papers surveyed. Many comparisons are corrected in a defensible fashion. Moreover, even papers using multiple comparisons corrections that, strictly speaking, rely on assumptions that were not really met, may report relationships that do indeed exist at least to some nonzero extent. In any case, we argue that (a) the actual correlation values reported by the non-independent analyses comprising over half of the studies we examined are sure to be inflated to the point of being completely untrustworthy, (b) assertions of statistical significance based on non-independent analyses require careful scrutiny—which does not always appear to have been done in the publication process. Perhaps most importantly, we argue (c) that if researchers would use the approaches recommended above (see Question D) they could avoid the whole treacherous terrain of non-independent analyses and its attendant uncertainties and complexities. In this way, the statistics would only need to be done once, the false alarm risk would be completely transparent, and there would be no need to use highly complex corrections for multiple comparisons that rest on hard-to-assess assumptions.

G. Well, in those cases where the correlation really is significant (i.e., nonzero), isn't that what matters, anyway? Does the actual correlation value really matter so much?

We contend that the magnitude, rather than the mere existence, of the correlation is what 'really matters'. A correlation of 0.96 (as in Sander et al., 2005), indicates that 92% of the variance in proneness to anxiety is predicted by the right cuneus response to angry speech. A relationship of such strength would be a milestone in understanding of brain-behavior linkages, full of promise for potential diagnostic and therapeutic spin-offs. In contrast, suppose—and here we speak purely hypothetically--the true correlation in this case were 0.1, accounting for 1% of the variance. The practical implications would be far less, and the scientific interest would be greatly reduced as well. A correlation of 0.1 could be mediated by a wide variety of highly indirect relationships devoid of any generality or interest. For instance, proneness to anxiety may lead people to breathe faster, drink more coffee, or make slightly different choices in which lipids they ingest. All of these are known to have

effects on BOLD responses (Weckesser et al, 1999; Mulderink et al., 2002; Noseworthy et al, 2003), and those effects could easily interact slightly with the specific hemodynamic responses of different brain areas. Or perhaps anxious people are more afraid than others of failing to follow task instructions and attend ever so slightly more to the required auditory stream. The weaker the correlation, the greater the number of indirect and uninteresting causal chains that might be accounting for it, and the greater the chance that the effect itself will appear and disappear in different samples in a completely inscrutable fashion (e.g., if the dietary propensities of anxious people in England differ from those of anxious people in Japan). We suspect that it is for this reason that the field of risk-factor epidemiology is said to have reached some consensus that findings involving modest but statistically significant risk ratios (e.g., ratios between 1.0 and 2.0) have not generally proven to be robust or important. It seems likely to us that most reviewers in behavioral and brain sciences also implicitly view correlation magnitude as important, and we suspect that the very fact that so many of the studies reviewed here appeared in high-impact journals partly reflects the high correlation values they reported.

Concluding Remarks

We began this article by arguing that many correlations reported in recent social neuroscience literature are "impossibly high". Correlations of this magnitude are unlikely to occur even if one makes the (implausible) assumption that the true underlying correlations -- the correlations that would be observed if there were no measurement error -- are perfect. We then went on to describe our efforts to figure out how these impossible results could possibly be arising. While the method sections of articles in this area did not provide much information about how analyses were being done, a survey of researchers provided a clear and worrisome picture. Over half of the investigators in this area used methods that are guaranteed to offer greatly inflated estimates of correlations. As seen in Figure 5, these procedures turn out to be associated with the great majority of the correlations in the literature that struck us as impossibly high[18].

Interestingly, we suspect that the problems brought to light here are ones that most editors and reviewers of studies using purely behavioral measures would usually be quite sensitive to. Suppose an author reported that a questionnaire measure was correlated with some target behavioral measure at r=.85, and that this number was arrived at by separately computing the correlation between the target measure and each of the items on the questionnaire, and reporting just the average of the highest-correlated questionnaire items. Moreover, to assess whether these highest-correlated questionnaire items were just the tail of a chance distribution across the many items, a filtering procedure had been used with properties too complex to derive analytically. We believe that few prestigious psychology journals would publish such findings. It may be that the problems are not being recognized in social neuroscience because of the

[18] The others (high green numbers in Figure 5) could simply reflect normal sampling variability of the sort found with any kind of imperfect measurement.

relative unfamiliarity of the measures, and the relatively greater complexity of the data analyses. Moreover, perhaps the fact that the papers report using procedures that include *some* precautions relating to the issue of multiple comparisons leads reviewers to assume that such matters are all well taken care of.

As discussed above, one thing our conclusions leave open is whether, behind any given inflated correlation, there is at least *some* real relationship—i.e. a true correlation higher than zero. Most investigators used thresholds that ostensibly correct for multiple comparisons but, we have argued, in some cases these corrections were seriously misapplied. Based on the analysis described above, we suspect that while in many cases the reported relationships probably reflect some underlying relationship (albeit a much weaker relationship than the numbers in the articles implied), it is quite possible that a considerable number of relationships reported in this literature are entirely illusory.

To sum up, then, we are led to conclude that a disturbingly large, and quite prominent, segment of social neuroscience research is using seriously defective research methods and producing a profusion of numbers that should not be believed.

A Suggestion to Investigators

Despite the dismal scenario painted in the last paragraph, we can end on a much more positive note. We pointed out earlier how investigators could have explored these behavioral trait- brain activity correlations using methods that do not have any of the logical and statistical deficiencies described here. The good news is that in almost all cases *the correct (and simpler) analyses can still be performed.* It is routine, and often required by journals and funders, for large neuroimaging data sets (which have usually been collected at great cost to public agencies) to be archived. Therefore, in most cases it is not too late to perform the analyses advocated here (or possibly others that also avoid the problem of non-independence). Thus, we urge investigators whose results have been questioned here to perform such analyses and to correct the record by publishing follow-up errata that provide valid numbers. At present, all studies performed using these methods have large question marks over them. Investigators can erase these question marks by re-analyzing their data with appropriate methods.

REFERENCES

Aron, A. R., Gluck, M. A., and Poldrack, R. A. (2006). Long-term test-retest reliability of functional MRI in a classification learning task. *Neuroimage,* 29, 1000-1006.

Baker, C. I., Hutchison, T. L., & Kanwisher, N. (2007). Does the fusiform face area contain subregions highly selective for nonfaces? *Nat Neurosci, 10*(1), 3-4.

Baker, C. I., Simmons, W. K., Bellgowan, P. S., & Kriegeskorte, N. (2007). *Circular inference in neuroscience: The dangers of double dipping.* Paper presented at the Society for Neuroscience, San Diego.

Brieman, L & Spector, P. (1992) Submodel selection and evaluation in regression. The X-random case. *International Statistical Review*, *60(3),* 291-319/.

Cox, R.W. (1996) AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. Computers and Biomedical Research, 29:162-173.

Evans, AC, Collins, DL, Mills, SR, Brown, ED, Kelly RL, & Peters TM. (1993) 3d statistical neuroanatomical models from 305 MRI volunteers. *Nuclear Science Symposium and Medical Imaging Conference, 3* 1813-1817.

Fiske, S. (2003). http://www.psychologicalscience.org/observer/getArticle.cfm?id=1242

K.J. Friston, A.P. Holmes, J.B. Poline, C.J. Price, and C. Frith. (1995) Detecting Activations in PET and fMRI: Levels of Inference and Power. *NeuroImage, 40*, 223-235.

Geisser, S. (1993). *Predictive Inference: An Introduction*, CRC Press.

Harmon-Jones, E. & Winkielman, P. (2007). Social Neuroscience. Integrating biological and psychological explanations of social behavior. Guilford Press. New York.

Hobbs, T. R., & Fowler, R. D. (1974). Reliability and scale equivalence of the Mini-Mult and MMPI. *Journal of Consulting and Clinical Psychology,* 1974, 42, 89-92.

Hurley, D. (2008) The Science of Sarcasm (Not That You Care). *New York Times*, June 3, 2008 http://www.nytimes.com/2008/06/03/health/research/03sarc.html?_r=1&oref=slogin

Johnstone, T., Somerville, L. H., Alexander, A. L., Oakes, T. R., Davidson, R., Kalin, N. H., and Whalen, P. J. (2005). Stability of amygdale BOLD response to fearful faces over multiple scan sessions. *NeuroImage,* 25, 1112-1123.

Kong, J., Gollub, R. L., Webb, J. M., Kong, J-T, Vangel, M. G., & Kwong, K. (2007). Test-retest study of fMRI signal change evoked by electroacupuncture stimulation. *NeuroImage,* 34, 1171-1181.

Kriegeskorte, N., Simmons, W.K., Bellgowan, P.S., & Baker, C.I. (2008) *Circular inference in neuroscience: The dangers of double dipping.* Paper presented at the Vision Science Society, Naples, FL.

Manoach, D.S., Halpern, E.F., Kramer, T.S., Chang, Y., Goff, D.C., Rauch, S.L., et al., 2001. Test–retest reliability of a functional MRI working memory paradigm in normal and schizophrenic subjects. Am. J. Psychiatry 158, 955– 958.

Mulderink, T. A., Gitelman, D. R., Mesulam, N. M., & Parrish, T. B. (2002). On the use of caffeine as a contrast booster for BOLD fMRI studies. *Neuroimage,* 15, 37-44.

National Institute of Mental Health (2007) New Social Neuroscience Grants to Help Unravel Autism, Anxiety Disorders (http://www.nimh.nih.gov/science-news/2007/new-social-neuroscience-grants-to-help-unravel-autism-anxiety-disorders.shtml)

Noseworthy M. D., Alfonsi, J., & Bells, S. (2003) Attenuation of brain BOLD response following lipid ingestion. *Human Brain Mapping*, 20, 116-21.

Nunnally, JC. Introduction to Psychological Measurement. New York: McGraw-Hill; 1970.

Saxe, R., Brett, M., Kanwisher, N. (2006). Divide and Conquer: A defense of functional localizers. Neuroimage May 1. 30(4):1088-96

Simmons, W. K., Matlis, S., Bellgowan, P. S., Bodurka, J., Barsalou, L. W., & Martin, A. (2006). Imaging the context-sensitivity of ventral temporal category representations using high-resolution fMRI. *Society for Neuroscience Abstracts*.

Stark, R., Schienle, A., Walter, B., Kirsch, P., Blecker, C., Ott, U., 2004. Hemodynamic effects of negative emotional pictures—a test– retest analysis. *Neuropsychobiology* 50, 108–118.

Taleb, N. (2004). *Fooled by Randomness: The Hidden Role of Chance in Life and in the Market.* New York: Thomson/Texere.

Talairach, J. & Tournoux, P. (1988) *Co-planar Stereotaxis Atlas of the Human Brain*. Thieme Medical Publishers, New York.

Viswesvaran, C., & Ones, D. S. (2000). Measurement error in "Big Five Factors" personality assessment: Reliability generalization across studies and measures. *Educational and Psychological Measurement*, 60, 224-235.

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

Weckesser, M., Posse, S., Olthoff, U., Kemna, L., Dager, S., Müller-Gärtner, H. W. (1999). Functional imaging of the visual cortex with bold-contrast MRI: hyperventilation decreases signal response. *Magnetic Resonance Medicine*, 41, 213-216.

Wei, X., Yoo, S.S., Dickey, C.C., Zou, K.H., Guttmann, C.R., Panych, L.P., (2004). Functional MRI of auditory verbal working memory: long-term reproducibility analysis. *NeuroImage* 21, 1000– 1008.

Yovel G., Kanwisher N. (2005) The neural basis of the behavioral face-inversion effect *Current Biology,* 15(24) 2256-62.

Table 1. The 54 surveyed articles listed in Figure 5.

Non Independent (red)

1	Sander, D., Grandjean, D., Pourtois, G., Schwartz, S., Seghier, M.L., Scherer, K.R., & Vuilleumier, P. (2005). Emotion and attention interactions in social cognition: Brain regions involved in processing anger prosody. *Neuroimage,* 28, 848–858.
2	Najib, A., Lorberbaum, J.P., Kose, S., Bohning, D.E., & George, M.S. (2004). Regional brain activity in women grieving a romantic relationship breakup. *American Journal of Psychiatry*, 161, 2245–2256.
3	Amin, Z., Constable, R.T., & Canli, T. (2004). Attentional bias for valenced stimuli as a function of personality in the dot-probe task. *Journal of Research in Personality*, 38(1), 15-23.
4	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.
5	Goldstein, R.Z., Tomasi, D., Alia-Klein, N., Cottone, L.A., Zhang, L., Telang, F., & Volkow, N.D. (2007a). Subjective sensitivity to monetary gradients is associated with frontolimbic activation to reward in cocaine abusers. *Drug and Alcohol Dependence*, 87(2–3), 233-240.
6	Eisenberger, N.I., Lieberman, M.D., & Williams, K.D. (2003). Does rejection hurt? An FMRI study of social exclusion. *Science*, 302, 290-292.
7	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(11), 2709-2724.
8	Takahashi, H., Matsuura, M., Yahata, N., Koeda, M., Suhara, T., & Okubo, Y. (2006). Men and women show distinct brain activations during imagery of sexual and emotional in.delity. *Neuroimage*, 32, 1299-1307.
9	Canli, T., Amin, Z., Haas, B., Omura, K., & Constable, R.T. (2004). A double dissociation between mood states and personality traits in the anterior cingulate. *Behavioral Neuroscience*, 118, 897-904.
10	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.
11	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.
12	Takahashi, H., Kato, M., Matsuura, M., Koeda, M., Yahata, N., Suhara, T., & Okubo Y. (2008). Neural correlates of human virtue judgment. *Cerebral Cortex*, 18(9), 1886-1891.
13	Britton, J.C., Ho, S.H., Taylor, S.F., & Liberzon, I. (2007). Neuroticism associated with neural activation patterns to positive stimuli. *Psychiatry Research: Neuroimaging*, 156(3), 263-267.
14	Straube, T., Mentzel, H.J., & Miltner, W.H. (2007). Waiting for spiders: brain activation during anticipatory anxiety in spider phobics. *Neuroimage*, 37:1427-

1436
15 Jabbi, M., Swart, M., & Keysers, C. (2007). Empathy for positive and negative emotions in the gustatory cortex. *NeuroImage,* 34, 1744-1753.
16 Killgore, W.D., Gruber, S.A., & Yurgelun-Todd, D.A. (2007): Depressed mood and lateralized prefrontal activity during a Stroop task in adolescent children. *Neuroscience Letters,* 416, 43-48.
17 Takahashi, H., Yahata, N., Koeda, M., Matsuda, T., Asai, K., & Okubo, Y. (2004). Brain activation associated with evaluative processes of guilt and embarrassment: an fMRI study. *Neuroimage*, 23, 967-974.
18 Aron, A., Fisher, H., Mashek, D.J., Strong, G., Li, H., & Brown, L.L. (2005). Reward, motivation, and emotion systems associated with early-stage intense romantic love. *Journal of Neurophysiology*, 94, 327-337.
19 Singer, T., Seymour, B., O'Doherty, J., Kaube, H., Dolan, R.J., & Frith, C.D. (2004). Empathy for pain involves the affective but not sensory components of pain. *Science*, 303, 1157-1162.
20 Canli, T., Sivers, H., Whitfield, S.L., Gotlib, I.H., & Gabrieli, J.D.E. (2002). Amygdala response to happy faces as a function of extraversion. *Science*, 296, 2191.
21 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.
22 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.
23 Somerville, L.H., Kim, H., Johnstone, T., Alexander, A.L., & Whalen, P.J. (2004). Human amygdala responses during presentation of happy and neutral faces: correlations with state anxiety. *Biological Psychiatry*, 55, 897–903.
24 Mantani, T., Okamoto, Y., Shirao, N., Okada, G., & Yamawaki, S. (2005). Reduced activation of posterior cingulate cortex during imagery in subjects with high degrees of alexithymia: a functional magnetic resonance imaging study. *Biological Psychiatry*, 57, 982-990.
25 Barrett, J., Pike, G.B., & Paus, T. (2004). The role of the anterior cingulate cortex in pitch variation during sad affect. *European Journal of Neuroscience*, 19(2), 458–464.
*26 Mitchell, J.P., Macrae, C.M., & Banaji, M.R. (2006). Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron*, 50, 655–663.
30 Reuter, M., Stark, R., Hennig, J., Walter, B., Kirsch, P., Schienle, A., & Vaitl, D. (2004). Personality and emotion: test of Gray's personality theory by means of an fMRI study. *Behavioral Neuroscience*, 118, 462– 469.
31 Singer, T., Seymour, B., O'Doherty, J.P., Stephan, K.E., Dolan, R.J., & Frith, C.D. (2006). Empathic neural responses are modulated by the perceived fairness of others. *Nature*, 439, 466–469.

No Response (orange)

27 Abler, B., Erk, S., Herwig, U., & Walter, H. (2007). Anticipation of aversive stimuli activates extended amygdala in unipolar depression. *Journal of Psychiatric Research*, 41, 511-522.

28 Phelps, E.A., O'Connor, K.J., Gatenby, J.C., Gore, J.C., Grillon, C., & Davis, M. (2001). Activation of the left amygdala to a cognitive representation of fear. *Nature Neuroscience*, 4, 437-441.

34 Glahn, D.C., Lovallo, W.R., & Fox, P.T. (2007). Reduced amygdala activation in young adults at high risk of alcoholism: studies from the Oklahoma Family Health Patterns Project. *Biological Psychiatry*, 61, 1306-1309.

Independent (green)

29 Lee, K.H., Brown ,W.H., Egleston, P.N., Green, R.D.J., Farrow, T.F.D., Hunter, M.D., Parks, R.W., Wilkinson, I.D., Spence, S.A., & Woodruff, P.W.R. (2006). A functional magnetic resonance imaging study of social cognition in schizophrenia during an acute episode and after recovery. *American Journal of Psychiatry,* 163, 1926-1933.

32 Phelps, E., O'Connor, K., Cunningham, W., Funayama, E., Gatenby, J., Gore, J., & Banaji, M.R. (2000). Performance on indirect measures of race evaluation predicts amygdala activation. *Journal of Cognitive Neuroscience,* 12, 729–738.

33 Coccaro, E.F., McCloskey, M.S., Fitzgerald, D.A., & Phan, K.L. (2007b). Amygdala and orbitofrontal reactivity to social threat in individuals with impulsive aggression. *Biological Psychiatry*, 62, 168-178.

35 McClernon, F.J., Hiott, F.B., Huettel, S.A., & Rose, J.E. (2005) Abstinence-induced changes in self-report craving correlate with event-related FMRI responses to smoking cues. *Neuropsychopharmacology*, 30, 1940-1947.

36 Herwig, U., Kaffenberger, T., Baumgartner, T., & Jancke, L. (2007). Neural correlates of a 'pessimistic' attitude when anticipating events of unknown emotional valence. *NeuroImage*, 34, 848-858.

37 Nitschke, J.B., Nelson, E.E., Rusch, B.D., Fox, A.S., Oakes, T.R., & Davidson, R.J. (2004). Orbitofrontal cortex tracks positive mood in mothers viewing pictures of their newborn infants. *Neuroimage*, 21, 583-592.

38 Lee, B.T., Cho, S.W., Khang, H.S., Lee, B.C., Choi, I.G., Lyoo, I.K., & Ham, B.J. (2007). The neural substrates of affective processing toward positive and negative affective pictures in patients with major depressive disorder. *Progress in Neuro-Psychopharmacology & Biological Psychiatry*, 31(7), 1487-1492.

39 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.

40 Paulus, M.P., Rogalsky, C., Simmons, A., Feinstein, J.S., & Stein, M.B. (2003). Increased activation in the right insula during risk-taking decision making is related to harm avoidance and neuroticism. *NeuroImage*, 19, 1439-1448.

41 Richeson, J.A., Baird, A.A., Gordon, H.L., Heatherton, T.F., Wyland, C.L., Trawalter, S., & Shelton, J.N. (2003). An fMRI investigation of the impact of

interracial contact on executive function. *Nature Neuroscience*, 6, 1323– 1328.
42 Rilling, J.K., Gutman, D.A., Zeh, T.R., Pagnoni, G., Berns, G.S., & Kilts, C.D. (2002). A neural basis for social cooperation. *Neuron*, 35, 395-405.
43 Heinz, A., Wrase, J., Kahnt, T., Beck, A., Bromand, Z., Grüsser, S.M., Kienast, T., Smolka, M.N., Flor, H., & Mann, K. (2007). Brain activation elicited by affectively positive stimuli is associated with a lower risk of relapse in detoxified alcoholic subjects. *Alcoholism Clinical and Experimental Research*, 31(7), 1138-1147.
44 Schneider, F., Habel, U., Kessler, C., Salloum, J. B., & Posse, S. (2000). Gender differences in regional cerebral activity during sadness. *Human Brain Mapping*, 9, 226-238.
45 Leland, D., Arce, E., Feinstein, J., & Paulus, M. (2006). Young adult stimulant users increased straital activation during uncertainty is related to impulsivity. *Neuroimage*, 33, 725–731.
46 Schneider, F., Weiss, U., Kessler, C., Salloum, J.B., Posse, S., Grodd, W., & Müller-Gärtner, H.W. (1998). Differential amygdala activation in schizophrenia during sadness. *Schizophrenia Research*, 34(3), 133–142.
47 Yucel, M., Lubman, D.I., Harrison, B.J., Fornito, A., Allen, N.B., Wellard, R.M., Roffel, K., Clarke, K., Wood, S.J., Forman, S.D., & Pantelis, C. (2007). A combined spectroscopic and functional MRI investigation of the dorsal anterior cingulate region in opiate addiction. *Molecular Psychiatry*, 12(611), 691–702.
48 Stein, M.B., Simmons, A.N., Feinstein, J.S., & Paulus, M.P. (2007). Increased amygdala and insula activation during emotion processing in anxiety-prone subjects. American *Journal of Psychiatry,* 164, 318–327.
49 Dannlowski, U., Ohrmann, P., Bauer, J., Kugel, H., Arolt, V., Heindel, W., & Suslow, T. (2007). Amygdala reactivity predicts automatic negative evaluations for facial emotions. Psychiatry Research: *Neuroimaging*, 154(1), 13–20.
50 Moriguchi, Y., Ohnishi, T., Lane, R.D., Maeda, M., Mori, T., Nemoto, K., Matsuda, H., & Komaki, G. (2006). Impaired self-awareness and theory of mind: an fMRI study of mentalizing in alexithymia. *Neuroimage*, 32(3), 1472-1482.
51 Habel, U.,Windischberger, C., Derntl, B., Robinson, S., Kryspin-Exner, I., Gur, R.C., & Moser, E. (2007). Amygdala activation and facial expressions: explicit emotion discrimination versus implicit emotion processing. *Neuropsychologia*, 45, 2369–2377.
52 Samanez-Larkin, G.R., Gibbs, S.E.B., Khanna, K., Nielsen, L., Carstensen, L.L., & Knutson, B. (2007). Anticipation of monetary gain but not loss in healthy older adults. *Nature Neuroscience,* 10, 787–791.
53 Kross, E., Egner, T., Ochsner, K., Hirsch, J., & Downey, G. (2007). Neural dynamics of rejection sensitivity. *Journal of Cognitive Neuroscience*, 19(6), 945-956.
54 Sanfey, A.G., Rilling, J.K., Aronson, J.A., Nystrom, L.E., & Cohen, J.D. (2003). The neural basis of economic decision-making in the Ultimatum Game. *Science*, 300, 1755-1758.

APPENDIX 1: fMRI Survey Question Text

Would you please be so kind as to answer a few very quick questions about the analysis that produced, i.e., the correlations on page XX. We expect this will just take you a minute or two at most.

To make this as quick as possible, we have framed these as multiple choice questions and listed the more common analysis procedures as options, but if you did something different, we'd be obliged if you would describe what you actually did.

The data plotted reflect the percent signal change or difference in parameter estimates (according to some contrast) of...

1. ...the average of a number of voxels.
2. ...one peak voxel that was most significant according to some functional measure.
3. ...something else?

If 1:

The voxels whose data were plotted (i.e., the "region of interest") were selected based on...

1a. ...*only* anatomical constraints (no functional data were used to define the region, e.g., all voxels representing the hippocampus).

1b. ...*only* functional constraints (voxels were selected if they passed some threshold according to a functional measure – no anatomical constraints were used; e.g., all voxels significant at p<.0001, or all voxels within a 5 mm radius of the peak voxel)

1c. ...anatomical and functional constraints (voxels were selected if they were within a particular region of the brain and passed some threshold according to a functional measure; e.g., all voxels significant at p<.0001 in the anterior cingulate)

1d. ...something else?

If you picked [1b, 1c, or 2] above could you please advise us about the following:

The functional measure used to select the voxel(s) plotted in the figure was...

[A]. ...a contrast within individual subjects (e.g., condition A greater than condition B at some p value for a given subject)

[B]. …the result of running a regression, across subjects, of the behavioral measure of interest against brain activation (for a contrast) at each voxel.

[C]. ...something else?

Finally: the fMRI data (runs/blocks/trials) displayed in the figure were...

[A]. ...the same data as those employed in the analysis used to select voxels (the functional localizer).

[B]. ...different data from those employed in the analysis used to select voxels (the functional localizer).

Thank you very much for giving us this information so that we can describe your study accurately in our review.

APPENDIX 2

G. Most papers use cluster size, not just a high threshold, to capture correlations. Does the inflation of correlation problem still exist in this case?

Yes. The problem arises from imposing any threshold which does not capture the full distribution of the 'true effect'. Since any true signal will also be corrupted by measurement noise, measurements of voxels that really do correlate with the behavioral measure of interest will also produce a distribution (although in this case the distribution will have a mean with a value that differs from zero). Imposing a threshold on this distribution will select only some samples – those with more favorable patterns of noise. If nearly the whole distribution is selected (statistical power is nearly 1) and there are no false alarm clusters, there would be no inflation. However, the lower the power, the more biased the selected subsample. Although cluster-size correction methods effectively increase power, they do not increase it sufficiently to mitigate bias. For simple whole-brain contrasts, cluster-size methods, appear to provide power that does not exceed 0.4 (and will more likely be substantially lower than that; Friston, Holmes, Poline, Price, and Frith, 1995). If statistical power is at 0.4, that means that only the top 40% of the true distribution will be selected – the mean of these selected samples will be very much higher than the true mean.

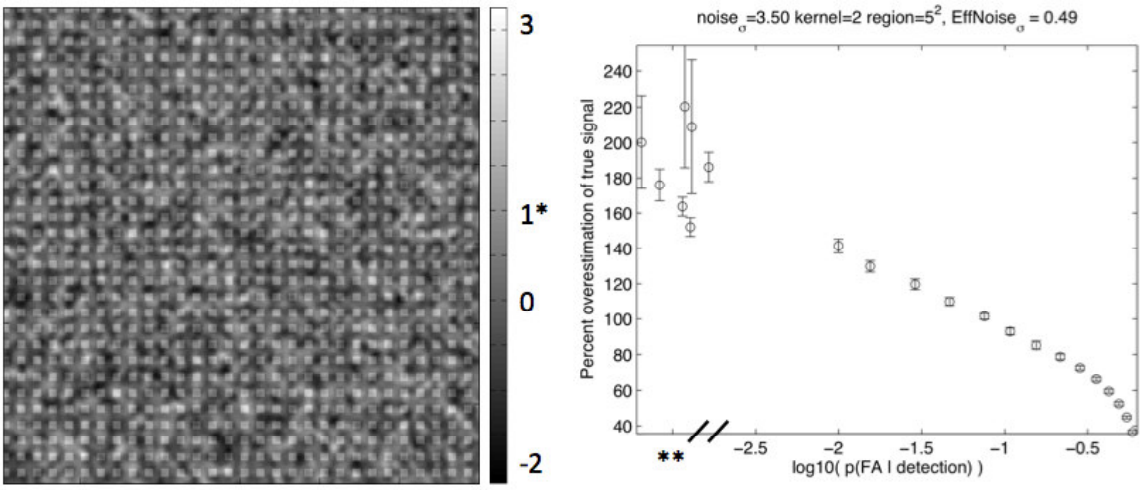

Figure A3: Simulation of cluster size correction and measure variable inflation.

For the moderately technical audience we provide a simplified cluster-size threshold simulation to show the magnitude with which the underlying signal can be inflated by an analysis procedure of roughly the sort we describe in this article. We generated a random 1000x1000 voxel slice (300x300 subset shown; the dimensions are irrelevant in our case, because we had a constant proportion of signal voxels) by generating random noise for each voxel (gaussian noise with mean 0 and standard deviation of 3.5). We blurred this slice with gaussian smoothing (kernel standard deviation = 2), thus inducing a spatial correlation between voxels, and resulting in an effective standard deviation of 0.5 per voxel. We then added "signals" to this noise: Signals were square "pulses" added to

randomly chosen 5X5 sub-regions of the matrix. Within one simulated matrix, 25% of the voxels were increased by 1. The color map shows measured intensity of a given voxel, with 0 being the noise average, 1 (marked with a *) the signal average.

We then did a simple cluster-search (finding 5x5 regions in which every voxel exceeded a particular threshold). We tried a number of different height thresholds, and for each threshold we measured the probability of a false alarm (the probability that a voxel that was within a 5x5 region in which all voxels passed threshold did not contain a true signal) -- the logarithm (base 10) of this probability is the x axis (-2 corresponds to p(FA) = 0.01, -0.3: p(FA) = 0.5). We also computed the inflation of the measured signal compared to the true signal in the detected voxels, as a percentage of true mean voxel amplitude; this is plotted on the y axis. "**" on the x-axis corresponds to simulated thresholds that did not produce <u>any</u> false alarm voxels in our simulations, thus, those reflect only regions that were entirely composed of signals. Error bars correspond to +/- 1.96 standard deviations across simulations for each threshold. (Naturally, low thresholds are on the right of the graph, producing many false alarms, high thresholds are on the left, producing few, if any, false alarms). A crude summary of the results of this simulation is that taking only signals that pass a threshold *always* inflates the underlying signal rather seriously (given thresholds that have a reasonable probability of false alarm), and as thresholds are raised to decrease false alarms, the signal inflation becomes even greater.