Voodoo?

or

What did Vul* Do?

Bob Cox – 20 Feb 2009

FMRI Discussion

*et al.

Starting Points & Assumptions

- You know what FMRI is
- You have at least skimmed Vul's paper
- I'm ***not*** planning to go through the paper itself in great detail
  - ***Nor*** am I going through the rebuttals in depth
  - **N.B.:** Herein "**Rebuttal**" as a heading means a point drawn from a rebuttal Web page/manuscript, not something I necessarily agree with *in toto*
- I'm going to outline the salient points, toss in some random criticisms, and then open it up
- **N.B.:** Herein, "**Vul**" means all the authors!

Outline of Vul's Argument

1. Correlations are too high to be plausible
   - Reliability of personality/emotional scores is about $0.8 = \sigma^2(\text{inter-subject}) / [\ \sigma^2(\text{inter}) + \sigma^2(\text{intra})\ ]$
   - Reliability of FMRI regression results is about 0.7
   - Therefore maximum plausible true correlation coefficient *r* between scores and FMRI is about $0.75=\sqrt{(0.8\times0.7)}$
2. Selecting voxels from which to report correlations ***from the same data*** used to calculate the correlations biases the results high by an unknown amount
   - And therefore the results have no meaning at all ***!***

Very High Correlations!

- $r > 0.8$ is a pretty high correlation between a physiogical measurement (FMRI) and a behavioral measurement
- **Vul point #1**:
  - Reliability (test-retest) of FMRI and psycho-social measurements can't support such large *r* values
- **Rebuttal points**:
  - FMRI reliability depends partly on how much time series data goes into each individual subject's map, and can be significantly higher than Vul states
  - Reliability of some psycho-social scores is much higher than Vul states [I have no idea myself]

First 3 Papers Vul Singles Out

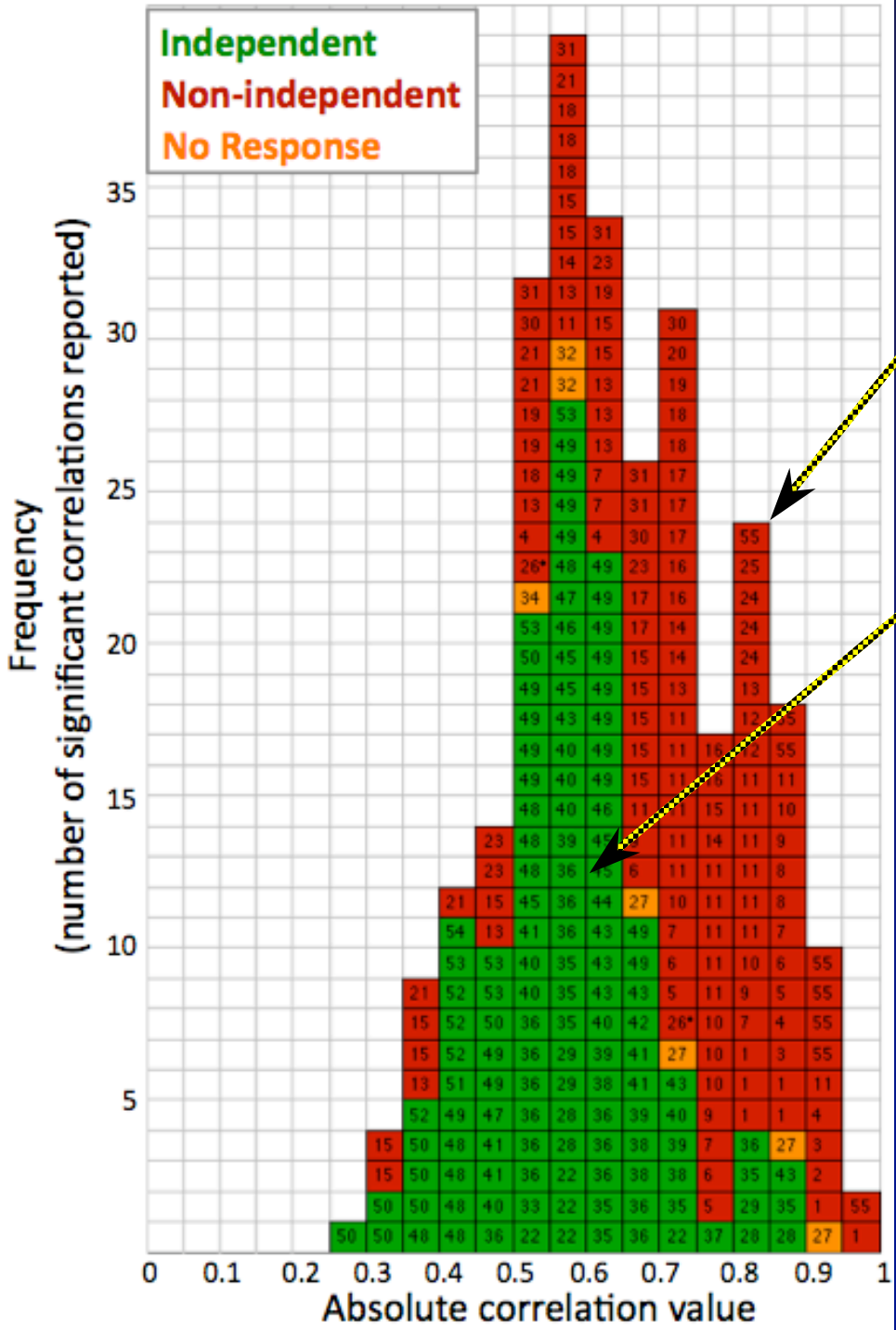
- Eisenberger 2003 (*Science*)
  - Best (peak value) $r = 0.88$ (!)
  - from 7.5 min of scanning, 60 *very* rapid event trials per condition; 8 mm blur; N=13 subjects
- Singer 2004 (*Science*)
  - Best (peak value) $r = 0.72$
  - 18 min EPI scanning, 20 trials per condition; 10 mm blur; N=16 subjects
- Sander 2005 (*NeuroImage*)
  - Best (peak value) $r = 0.96$ (!!)
  - unclear how much scanning (15-30 min?), 24 rapid event trials per condition; 8 mm blur; N=15 subjects
- Difficult to say ***exactly*** how they analyzed data

N.B.: must have $r > 0.5$ to have $p < 0.05$ in a *single voxels's test*, for these Ns

What is the Big Deal?

- **Vul point #2**: Non-independence ("circularity" or "selection bias"), abstractly expressed as:

1. Choose voxel set (ROIs) based on how well FMRI data (time series *or* regression maps) are correlated with something#1
   - Exclude below-threshold (e.g., low correlation) voxels from further analysis / testing / reporting
2. Report correlation of these selected voxels with something#2, *which is not entirely statistically independent of* something#1
3. Reported correlation will likely be biased high
   - and it's hard to know how much bias there is



Vul: Figure 5

- Red boxes = correlations taken from "non-independent" selection papers
- Green boxes = correlations taken from "independent" selection papers
- Rebuttals: Vul picked correlations from 'red' papers to make their point (their own "selection bias", which they *strongly* deny); *also*, mis-classified some papers as "green" or "red"

Reporting the ROI Correlation

- It's common to pick peak correlation or correlation coefficient at the peak "activation"
  - Will also bias results high, simply because even if all ***true*** correlations in ROI are equal, measured correlations will have random fluctuations
  - Picking the peak gives a nice *xyz* coordinate
- In any case (e.g., average across ROI **selected non-independently**, and/or use peak):
  - It's difficult to assess the ***accuracy*** (e.g., confidence interval) of the reported correlation
  - Even if ROI is selected stringently from FMRI activation, so that each voxel is "surely" active

Other & Lesser Points

- Methods sections are often imprecise, especially about the analysis details
  - Would be hard to replicate studies, even if they gave you all their data
- Some analyses contain stuff that is just plain wrong (but the import of which is unclear in results)
  - Eisenberger 2003: correction for multiple comparisons is incorrect (footnote 23 of that paper)
  - Vul caught this, but then they amusingly got the correction wrong *also* (fixed it when I emailed Vul)
- Amount of data gathered (for *Science* papers!) can be quite low **[***this is **my** point, not Vul's***]**
  - Eisenberger 2003: 7.5 min FMRI $\times$ 13 subjects

Different Degrees of Circularity/Bias

- ROI selection ***per subject*** by each subject's overall activation map (including the data to be correlated with behavioral measure)
- Peak voxel selection in an ROI ***per subject***
  - Does anyone do this?
- ROI or peak voxel selection ***across subjects***
  - That is, from the group activation map
  - **Especially bias-prone** if inter-subject ROI/voxel selection *only* uses using correlation with the behavioral measurement [Vul's Fig 3: next slide]
    - Vul claims this is very common, but that is not so clear
    - At least 1 person I know at NIMH says Vul mis-interpreted their paper and their response to the survey

Vul: Figure 3

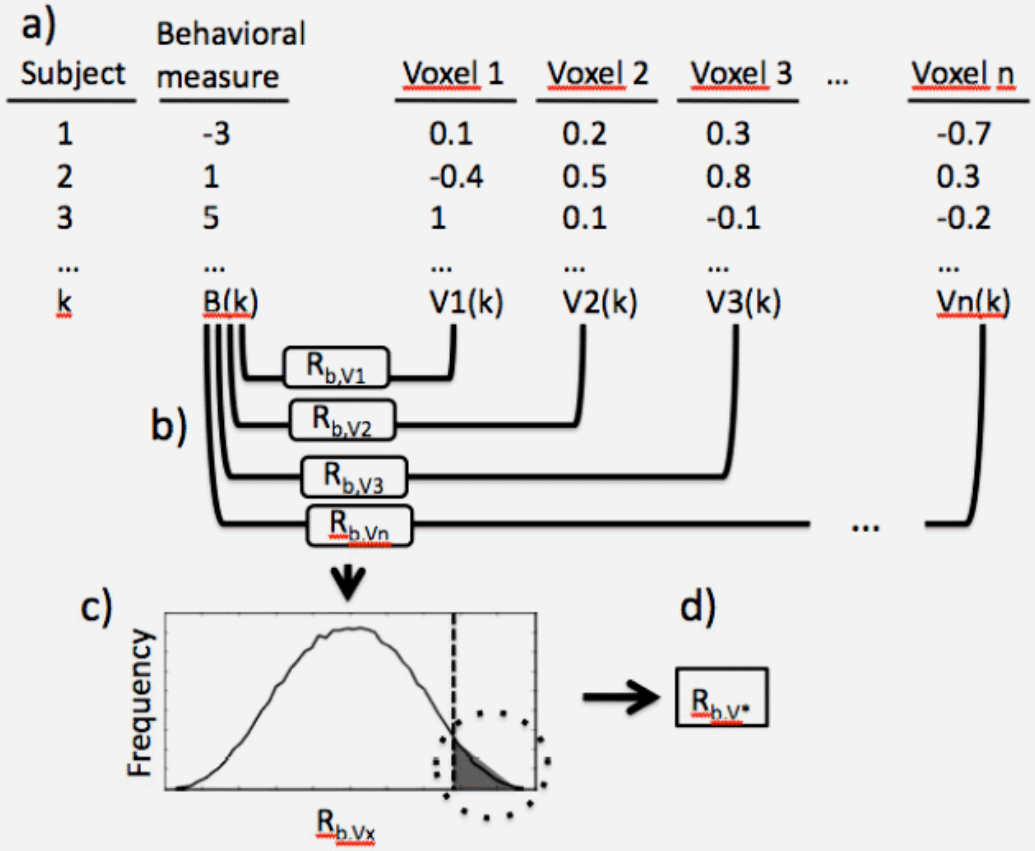


Figure 3: An illustration of the analysis employed by 54% of the papers surveyed. (a) From each subject, the researchers obtain a behavioral measure as well as BOLD measures from many voxels. (b) The activity in each voxel is correlated with the behavioral measure of interest across subjects. (c) From 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.

Aside: What is the Analysis Goal?

- Assessing the relation between brain activity level and some externally observed psycho-social-behavioral-emotional measurement(s)
- Providing statistical confidence of results (*p*)
  - Note that Vul argument has ***nothing*** to do with 3D ANOVA or LME (i.e., the most common forms of FMRI analysis and reports) group maps *per se*
- And some quantitative "strength" of the relationship (because we like numbers these days)
  - Vul paper conflates problems with strength measure with significance: **IMHO**, a serious error
- And some indication of where the relationship is strong: Coordinates; Anatomical region

My Take on Vul's Voodoo - 1

- Fig 4 [next slide] is *fundamentally misleading*
  - Has 10^8 independent uncorrelated data points
    - By selecting top 'voxels' with apparent high correlation, shows you can get a high "peak correlation" from correlation-free data
  - **But**: FMRI experiments (after blurring) typically have only about **10^4** independent data points (resolution elements × subjects) – a factor of 10,000 fewer
    - Rumpelstiltskin can't spin so little straw into gold
  - You can't draw quantitative conclusions about FMRI data analysis from this Figure***!***
    - Or from his related simulation in Appendix 2

Vul: Figure 4

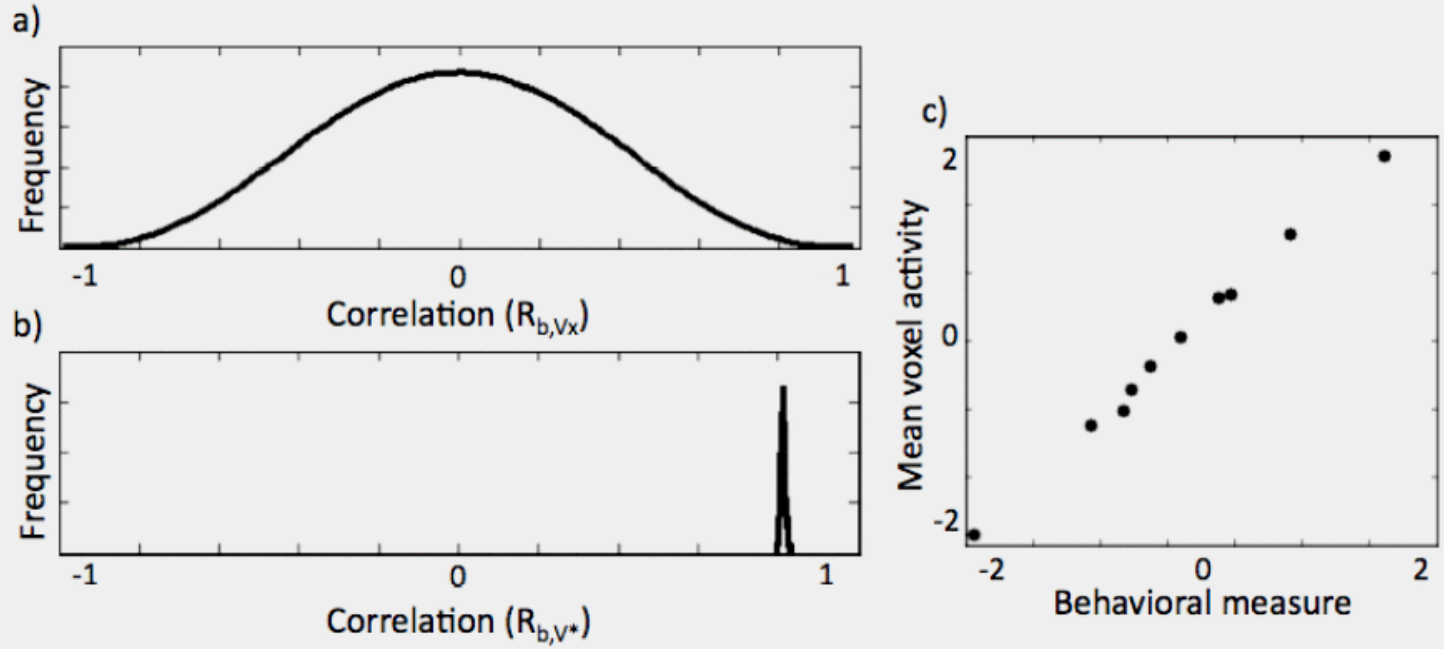


Figure 4: A simulation of a non-independent analysis on pure noise. We simulated 1000 experiments each with 10 subjects and 10000 voxels, and one individual difference measure. Each subjects' voxel activity and behavioral measure were independent 0-mean Gaussian noise. Thus, (a) the true distribution 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. (For similar exercises in other neuroimaging domains, see Appendix 2, and Baker, Hutchison, et al., 2007; Simmons et al., 2006; Kriegeskorte et al., 2008)

More on Vul's Fig 4 Argument

- The simulation shows the *possibility* of creating high apparent correlations when using a non-independent selection process
  - The weather station:stock price example is another example showing the *possibility* of completely specious results if you select from enough data
  - Examples are *not* intended to be a quantitative example of the problems appearing in FMRI
- But since the possibility exists, non-independent selection correlations are *completely meaningless – end of story*
  - **This is where I diverge from their argument**
  - **Unquantified bias doesn't imply no meaning**

My Take on Vul's Voodoo - 2

- They have one important point right:
  - Non-independent selection of points from which the correlation will be reported will bias the results high
  - And we can't estimate accurately how much bias there is (**but**: can't get r=0.8 results from true r=0!)
- They pound this point with provocative and annoying over-generalizations, misleading analogies, and unwarranted conclusions:
  - "**reported correlation coefficients mean almost nothing**" (but paper doesn't define "meaning")
  - "**many of the real relationships are probably far lower than the ones shown in green**" (for which no argument is presented, except innuendo)

My Take on Vul's Voodoo - 3

- **Plus** side:
  - Raising the issue of non-independent analysis and reporting, and for suggesting some alternatives
  - And the issue of reliability of the measurements being correlated
    - This subject deserves more attention (both on the FMRI and psychometric ends), especially if correlations might be used for predictions about individual patients
- **Minus** side:
  - Grotesquely misleading title
  - Exaggerated (non-realistic for FMRI) simulations
  - Grossly over-strong pronouncements
  - Their own methodological simplifications, errors, and imprecision
- My scoring: **+23 – 87 = –64** points* *[N.B.: sarcasm]

The Badness of the Bias?

- I've said the simulation examples can't be used as any kind of quantitative guide to decide how big the problem is in real FMRI datasets
- One feasible way to figure out what's up, Doc:
  - Carry out **realistic** simulations of FMRI datasets with various levels of correlations with an external variable (itself also corrupted by noise, of course)
  - Analyze these datasets with various actual processing strategies
  - Determine what the distribution of reported correlations vs. the underlying "truth": the correlations used to generate the simulated data
- Any volunteers?

What to Do with *Your* Data

- **Understand** what you are doing
- **Understand** your data
  - Look at it in different ways and at different stages of the analysis
- Avoid ***gross*** circularity in ROI selection
  - Select ROIs anatomically [completely non-circular]
  - "Punch out" ROI of 10 mm radius (say) around activations — whether or not all voxels are "active"
  - Select ROIs purely functionally, but from different imaging runs [mostly non-circular]
- Peak voxel correlations can be misleading
- What is it you are trying to say, anyway?