9.71 Functional MRI of High-Level Vision
Fall 2007

Some Tips on How to Critically Evaluate fMRI Studies
Nancy Kanwisher
9.71 Sept, 2007

1. First, figure out what question the researcher is asking and what answer they are giving to that question.

Ask yourself: Is this an interesting question? Does it have clear theoretical implications and if so what are they? Do you care about the result? Should anyone? Why? Are you *surprised* by the result? Situate the question in a broader theoretical context. If there is no such broader context, be worried.

2. The most critical aspect of the design of the experiment is: what is getting compared to what?

Make a list of all the mental functions that you think go on during the critical test condition. Then make a list of all the mental functions that are going on in the control condition, then see how many go on only (or more) in the test condition than the control condition.

Are the test and control conditions "minimal pairs"?

Be wary of very low baselines such as fixation.

Beware of comparisons of a very difficult task versus a very easy task. Lots of brain areas get activated by virtually any difficult task; this activation may be very nonspecific and may be difficult to interpret (without appropriate controls).

Watch out for attention confounds: is one condition much more interesting/engaging/attention capturing than another?

Watch out for eye movement confounds between conditions.

Imagine doing the task (or code up a version of it) and introspect while you do this mental simulation. What does it feel like and what do you think you would really do in this task?

3. Classic problems in analyses/inferences/conclusions to be wary of:

A. "Brain area X was activated by task Y."

i. Ask: task Y *compared to what*? Everything is a comparison, and many comparisons are uninformative/trivial.

ii. What else activates brain area X? the specific activation seen in task Y (compared to whatever) may not be so specific if prior studies have already been implicated the same area in dozens of other tasks/processes. Two classic cases in point: the anterior cingulate cortex (ACC), and the intraparietal sulcus (IPS).

iii. How strongly activated was that region? Not all 'activations" are the same - *Effect sizes matter!* If one condition produces a massive response compared to a given baseline, and another condition produces a very small but significant activation, the two "activations" are not the same. Brain imaging is the only field I can think of where people seem to think it is OK to report p levels without means; this is completely bizarre and can be highly misleading. (Note: many fMRI researchers appear to disagree with me on this one.)

B. “Because Region X responded significantly more strongly in Task A than control, but didn't respond significantly more strongly in Task B than control, it is selectively activated by Task A.”

A difference in significances is not necessarily a significant difference. If you want to claim that the region responds more to A than B, then *compare A to B.* Statistics are not transitive.

C. Claims of this form: “We found activation in the medial prefrontal cortex for tasks involving reasoning about other minds, consistent with numerous prior studies.”

Brains are as different across individuals as faces are, so what counts as the “same place” in the brain is not well defined across different brains. (Is the freckle on Joe’s nose in the same place as the freckle on Bob’s nose? This means *something,* but it is highly imprecise.) The prior activations for reasoning about other minds in the medial prefrontal cortex range up to 5 centimeters apart from each other. This is not the “same place” in any meaningful sense. The ‘same place” in the brain is only clearly defined within an individual subject. When pooling data across subjects, better options are to define regions of interest (functionally or anatomically) within each individual subject before pooling across subjects. Note that this is my opinion, and though many in the field agree with me, some disagree; see Friston et al (2006) and Saxe et al (2006) for a lively debate on this topic.

D. "The results of the present study demonstrate that Task A is carried out in a distributed network of cortical areas."

What has been learned here?

4. Some of the many ways to cheat:

A. Showing data from the “best voxel”.

With tens of thousands of voxels to chose from in an overall nosiy data set, some of them will look pretty good.

B. Showing “fitted data”. If you have a real effect, show your real data.

C. Showing activation maps that “look similar” or “look different”. There are many ways to chose particular slices, thresholds, etc to make activations look similar or different. If the claim is that they are similar or different, this should be tested statistically *on the exact same voxels*. Just showing similar-looking activations (especially in group data or across subjects) without statistically testing whether the *same voxels* are activated, is very weak. Beware of sneaky choice of slices; look at the anatomical images to see if it really is the same slices.

5. Some signs of a well done study:

A. The researchers show some raw data, e.g. nonfitted time courses or at least percent signal increases from fixation (or “beta weights”) in independently-defined regions of interest.

B. The critical result is replicated at least once.

C. More than one control condition is used, or the control condition is a “minimal pair”.

6. Some important general caveats about fMRI research:

A. Typical imaging parametrs include about *several hundred thousand neurons per voxel*! Most studies smooth their data and average across subjects which increases this number dramatically. It is a great miracle that we see anything at all with this method.

B. Temporal resolution of fMRI is lousy – at best a few 100 ms. Most of cognition happens in tens of milliseconds, not hundreds. So component steps cant usually be resolved.

C. fMRI activations do not imply necessity!