Assessing FMRI Reliability: Data from the FBIRN

Daniel H. Mathalon

Department of Psychiatry
UCSF

San Francisco VA Medical Center

FMRI Reliability

- Question: Is FMRI reliable?
- Refined Questions:
  - Is FMRI sufficiently reliable to be used to track changes in subjects over time?
  - Is FMRI sufficiently reliable to be used to track changes in patients with schizophrenia over time?
- How should we assess FMRI reliability?
  - What data can be used to support reliability?

NeuroImage **11**, 708–734 (2000)
doi:10.1006/nimg.2000.0562, available online at http://www.idealibrary.com on IDEAL®

Variability in fMRI: An Examination of Intersession Differences

D. J. McGonigle, A. M. Howseman, B. S. Athwal, K. J. Friston, R. S. J. Frackowiak, and A. P. Holmes[1]

Wellcome Department of Cognitive Neurology, Institute of Neurology, London WC1N 3BG, United Kingdom

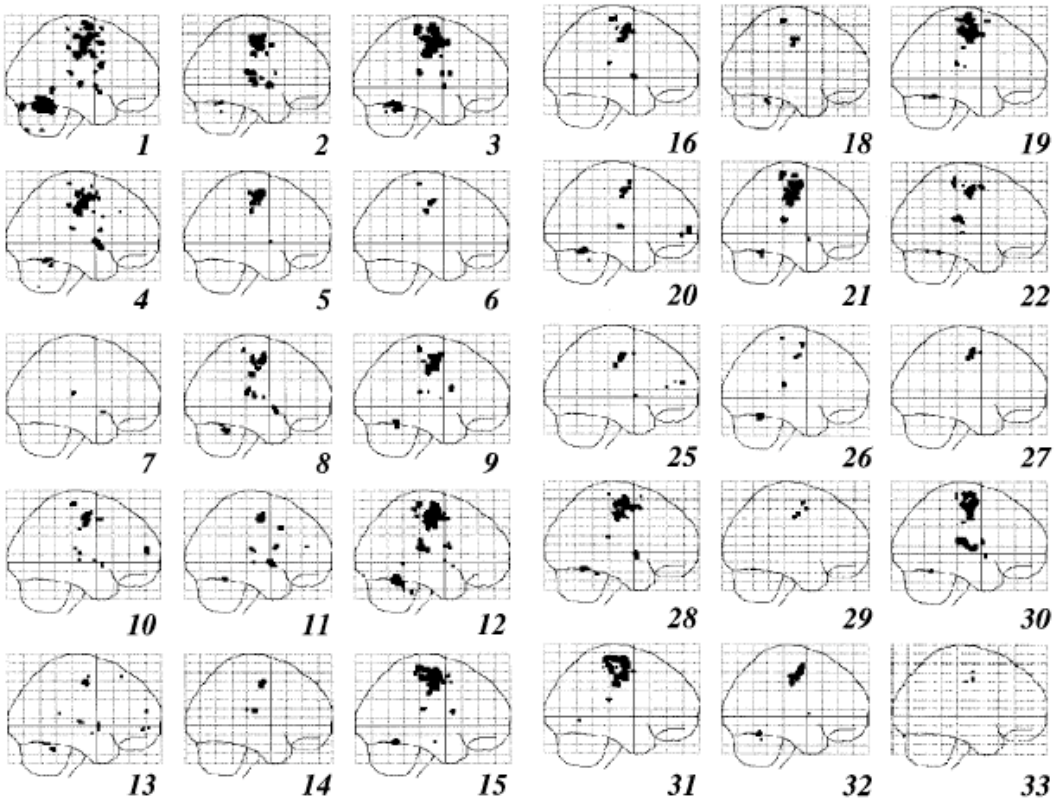
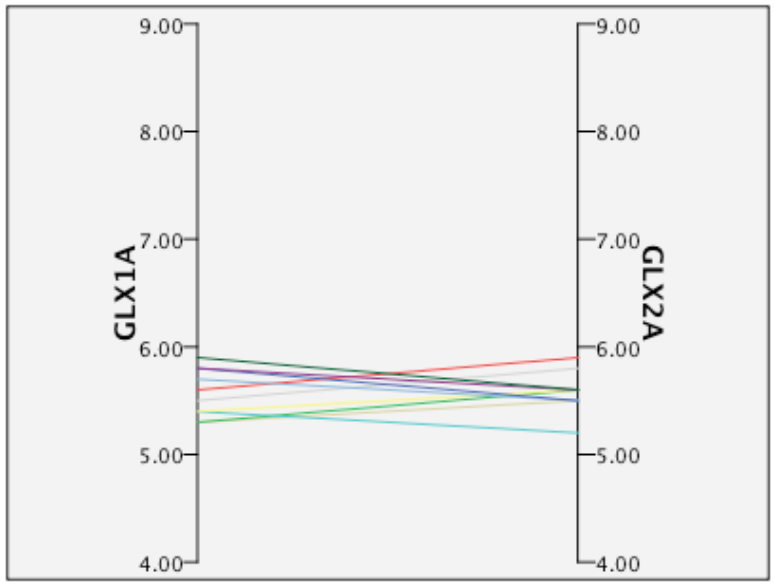


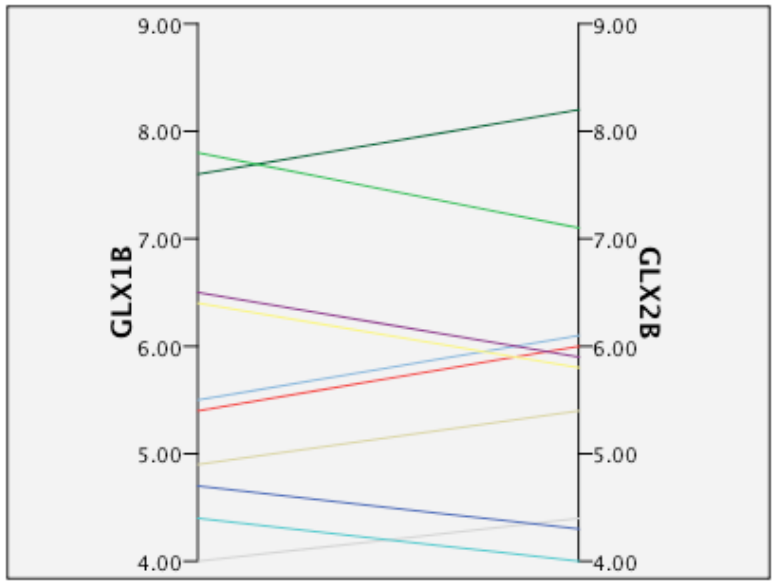
FIG. 2. Single-session sagittal MIPs for the motor paradigm. The number of each session is displayed below it. Although 33 sessions were collected, only 30 are shown here (sessions 17, 23, and 24 were rejected due to movement artifacts). All results are thresholded at $P < 0.05$ corrected for multiple comparisons unless otherwise stated.

Can single subject data be used to show that fMRI is unreliable?

- Not uncommon in grants to see pilot data from one or a few subjects tested on two or more occasions, showing a “small” percentage of variation in an FMRI signal over time.
- Measurements of variation in activation within a single subject over time do not usefully inform us about reliability.
  - The variation over time is “uncalibrated” by the degree of true variation in the measurement that exists across individuals.
  - Percentage of variation in a single subject may be useful to characterize the “precision” of the measurements (at least in the subjects tested), but not their “reliability”.

Hypothetical examples of 10 subjects tested twice in a (FMRI) test-retest reliability study

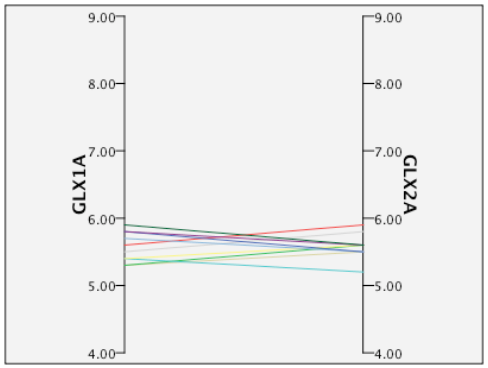




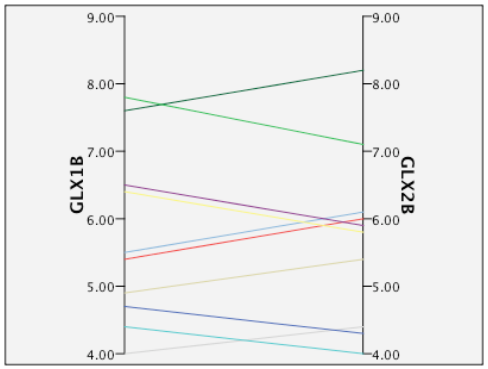
Note that in scenario A, subjects values range between values of 5-6, whereas in scenario B, values range between 4 – 8.

Percent Change in Signal Over Time: |t1-t2|/t1 * 100%

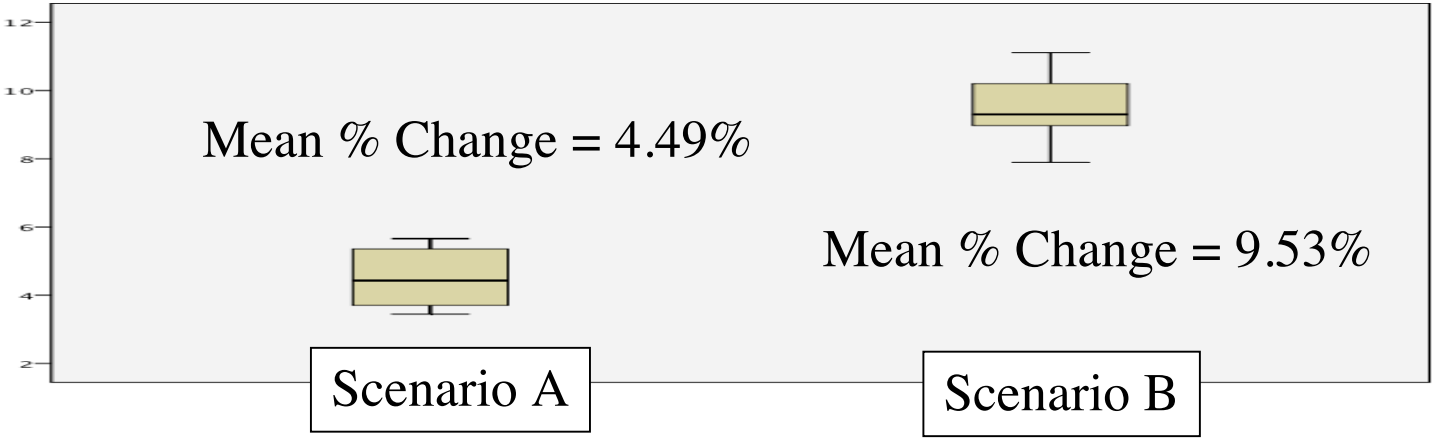
Scenario A



Scenario B



Box Plots of Percent Change Values

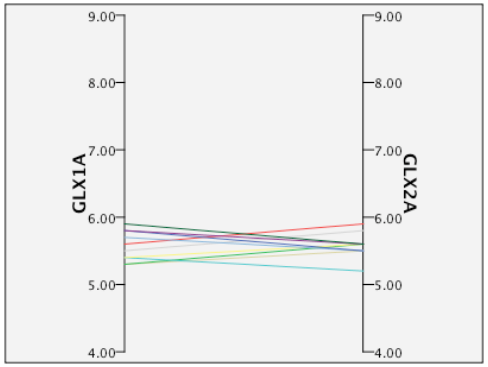


Note that the percent change values, averaged across subjects, is much worse in Scenario B than in Scenario A.

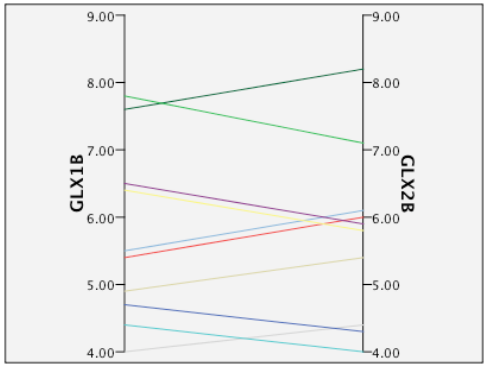
Q: Does this mean the reliability is lower in Scenario B?

A: NO! Reliability, calculated as Intraclass Correlation (ICC), is much worse in scenario A.

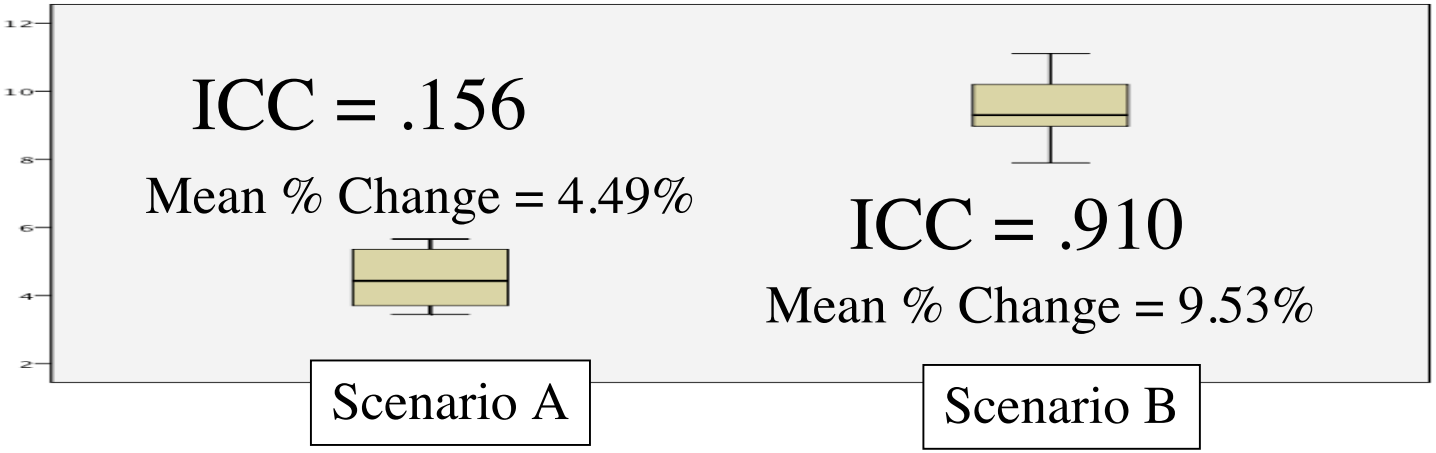
Scenario A



Scenario B



Box Plots of Percent Change Values



In Scenario A: The true variance among individuals in the population is relatively small, so the impact of an apparently small percent change from test to retest is much greater, substantially affecting our ability to consistently order individuals over time.

In Scenario B: The true variation among individuals in the population is relatively large, so the impact of the larger percent change from test to retest has less impact on our ability to differentiate among individuals consistently over time.

Why is Reliability Important?

- Reliability sets an upper limit on validity as reflected by the correlation between the fMRI measure and other measures of interest such as task performance, diagnosis, symptom severity, etc.
- It is impossible for an fMRI measure to correlate more highly with another measure than it correlates with itself.
  - Unless correlated measurement error contaminates the measurements.
- How do we assess reliability? ICC

Reliability as Defined in Classical Test Theory (CTT)

- Primary premise of CTT:
  - Observed test score (or fMRI measurement) = "true score" + "error"
  - In terms of observed variances:
    - $\sigma^2_{observed} = \sigma^2_{true} + \sigma^2_{error}$
      - "observed variance" = "true score variance" + "error variance"
- **Reliability = $\sigma^2_{true} / \sigma^2_{observed} = \sigma^2_{true} / (\sigma^2_{true} + \sigma^2_{error})$**
  - reliability = true variance / (true variance + error variance)
    - This is an intraclass correlation coefficient (ICC)
  - Conceptually, want to ascertain how much of observed test score variance is due to 'true score' variance versus 'error' variance.
  - There are a number of ways to quantify 'error variance'
    - 'Error' is a unitary construct in CTT (and error is 'bad').
    - Goal, then, is to reduce 'error' variance as much as possible
      - Standardization of measurement conditions (e.g., scanner performance, stimulus presentation characteristics, instructions and subject training, etc), making confounds constant across measurements
      - Aggregation --> more 'items' (for fMRI: trials, blocks, runs) are better (errors should cancel out).

Fundamental Equation

$$X = T + E$$

X = Observed score

T = True score

E = Error score

$$\text{Reliability} = \frac{\text{Variance of T}}{\text{Variance of X}}$$

The larger the variance of T in relation to X, the higher the reliability

Fundamental Equation

$$X = T + E$$

X = Observed score

T = True score

E = Error score

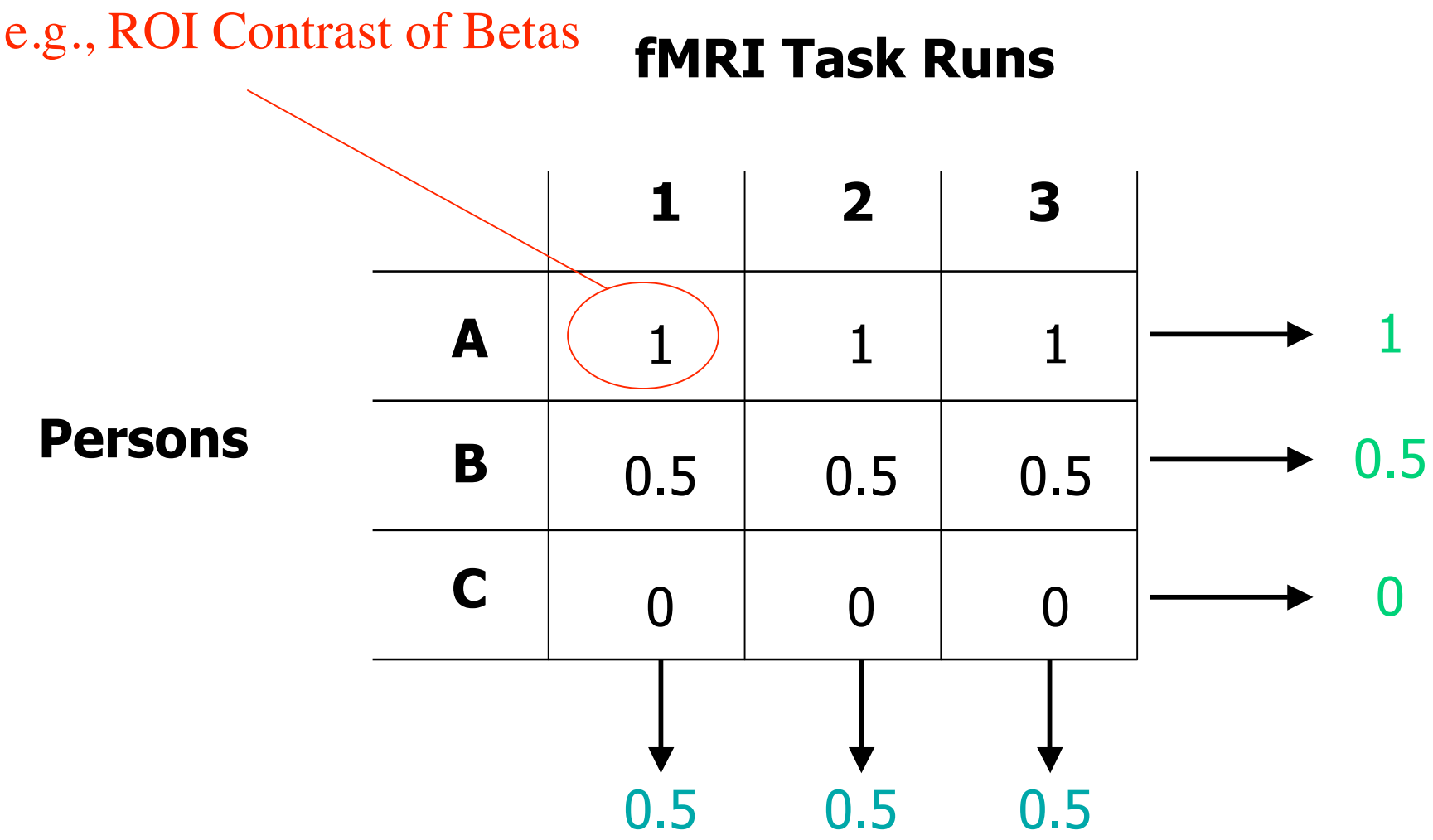
$$\text{Reliability} = \frac{\text{Variance of T}}{\text{Variance of X}}$$

$$\text{Reliability} = \frac{\text{Variance of T}}{\text{Var T + Var E}}$$

The larger the variance of T in relation to X, the higher the reliability.

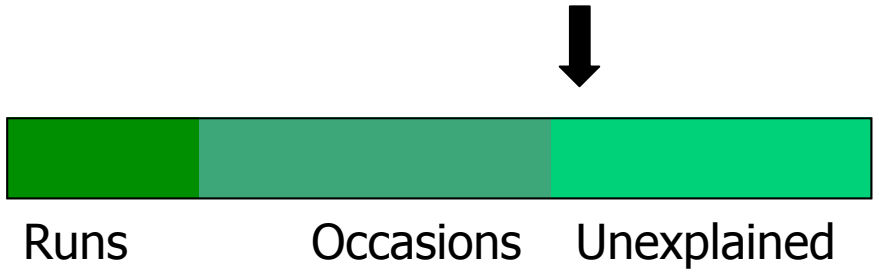
Variance of T is estimated by Person Variance in an ANOVA model in which Persons is a random effect.

Fundamental Equation



Multiple sources of error variance

$$\text{Reliability} = \frac{\text{Variance of T}}{\text{Var T + Var E}}$$



Generalizability Theory

- Cronbach's liberalization/expansion of CTT.
- Both CTT and G-Theory involve the concept of parallel measurements.
- However, in G-Theory, 'error' is not a unitary construct.
  - Goal is to decompose 'error variance' into as many measurable sources as possible.
    - Accomplished within an ANOVA framework
  - Consider multiple subjects measured over multiple fMRI task runs on multiple occasions (fully "crossed" design). What are the observable sources of variance?
    - Persons ("true score" or "universe score" variance)
    - Runs (inconsistencies due to stimulus choice, ordering, practice effects, fatigue, habituation).
    - Occasions (temporal specificity; effects of time; scanner drift; practice/exposure effects; habituation)
    - Persons x Runs (inconsistency over runs differs across subjects)
    - Persons x Occasion (inconsistency over occasions differs across subjects)
    - Persons x Runs x Occasions (highest order interactions + residual error)
- Distinctive characteristic:
  - G-theory allows inclusion of multiple sources of error in one reliability estimate

Generalizability Theory (cont'd)

- Dimensions or 'facets' of observation define boundaries in which observations are exchangeable.
  - Facet variances are usually considered 'random' effects in ANOVA framework.
- Reliability, then, depends on specific conditions and goals of measurement
  - there is not one single reliability coefficient that characterizes a test or measure like fMRI.

Generalizability Theory (Cont'd)

- Want to improve reliability of measurement?
  - Get people who differ as broadly as possible.
    - Increase "true score" variance (numerator in reliability)
    - Add to sample in whichever facets show the most variance--> because aggregation suppresses 'error'
  - General formulas are available to show how specific expected reliability coefficients would change with different sample sizes or different numbers of levels within a facet.

(Ideally) two steps in G analysis

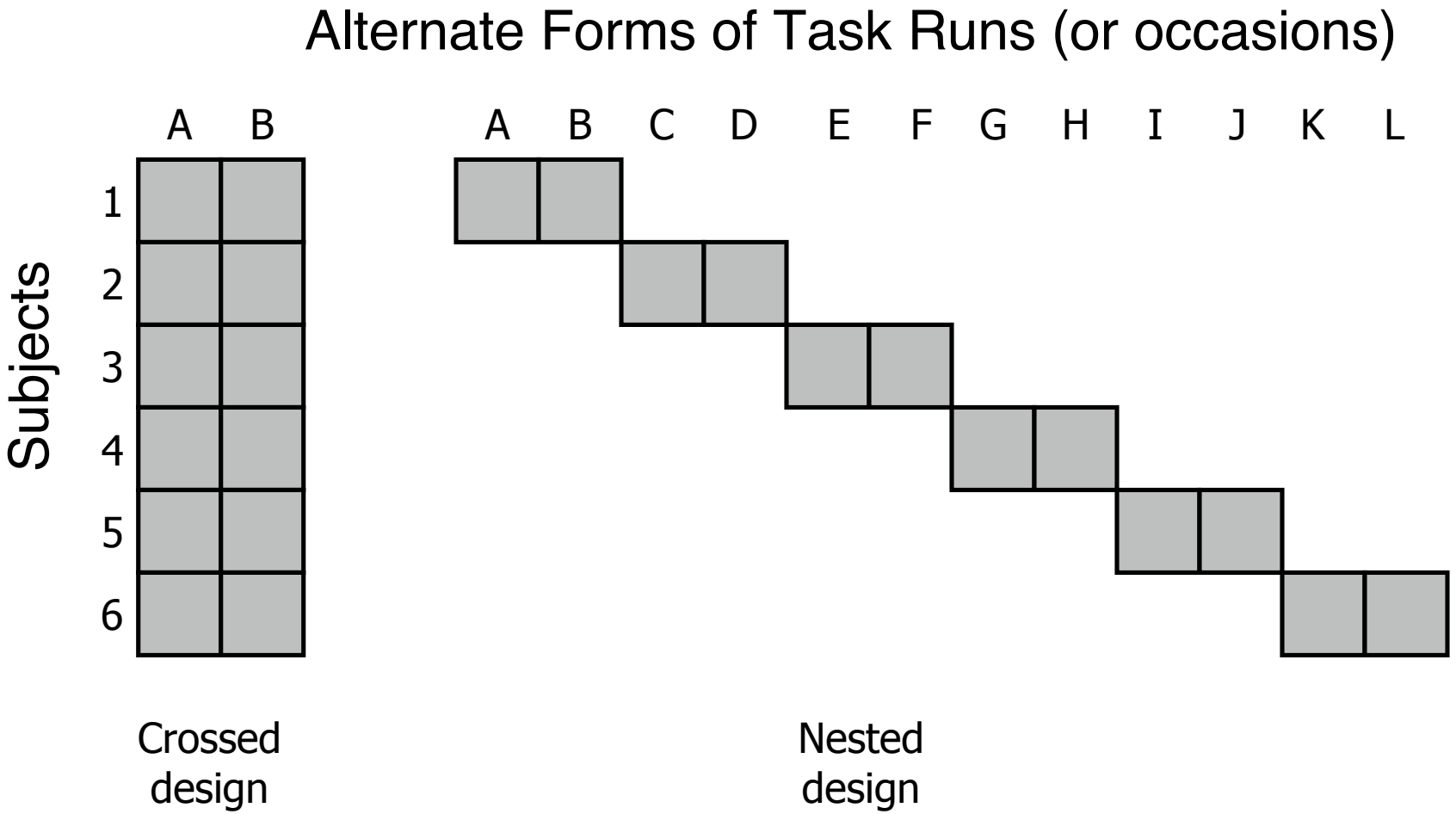
- G(eneralizability)-study:

  Estimation of sources of variance that influence the measurement (e.g., variance between subjects, runs and occasions)

- D(ecision)-study:

  Estimation of reliability indices as a function of concrete sample size(s) (e.g., number of runs, number of occasions)

Crossed vs nested Reliability Study Designs



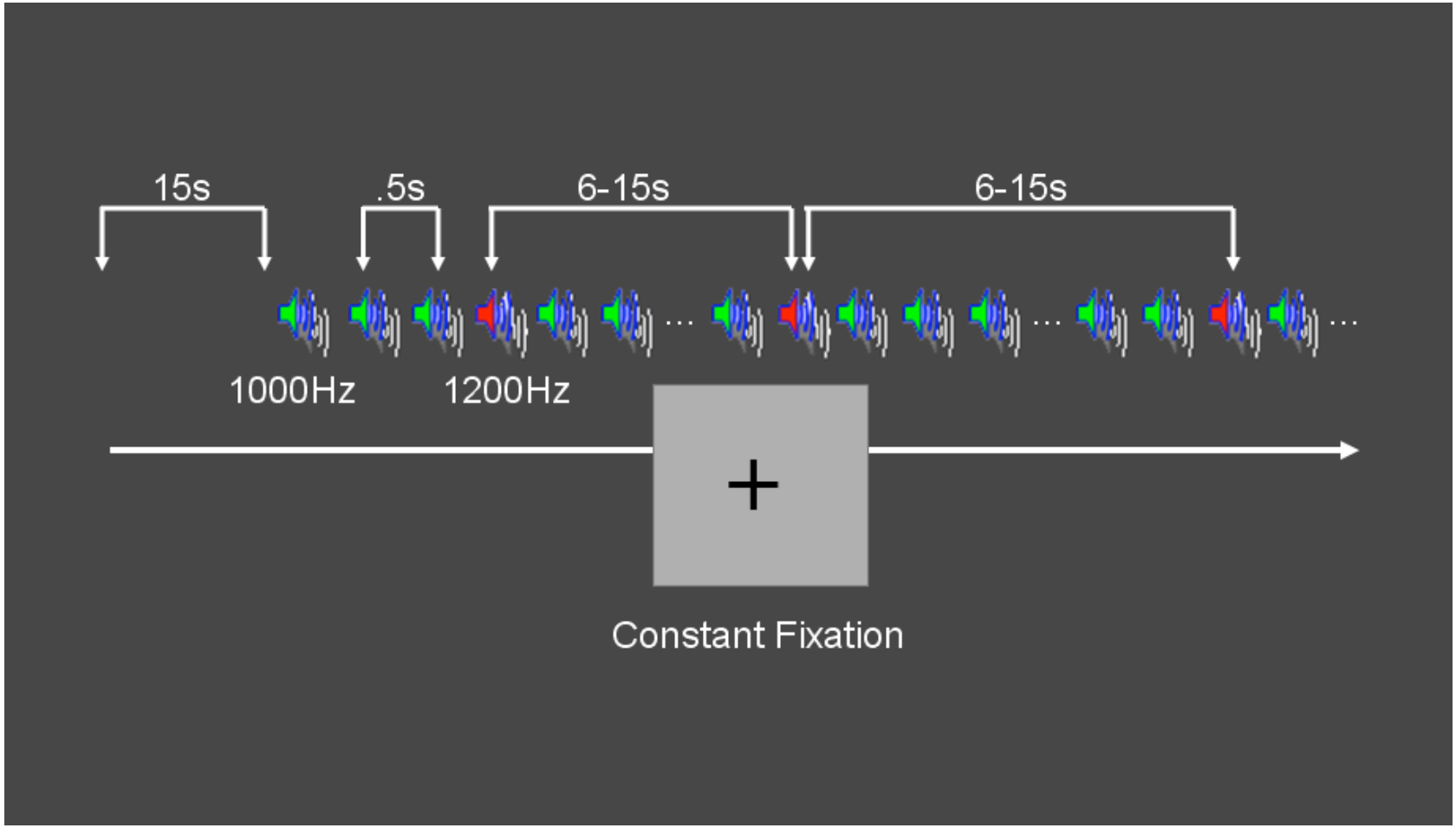
FBIRN: Phase II Study and the assessment of test-retest and cross-run reliability

- 9 Sites
  - Duke GE 4T
  - BWH GE 3T
  - MGH Siemens 3T
  - UCI Siemens 3T Allegra
  - UCLA Picker 1.5 T
  - University of Iowa Siemens 3T
  - University Minnesota Siemens 3T
  - New Mexico 3T Siemens
- Healthy controls (n=103) and patients with schizophrenia (n=94) recruited at each site.
- Two Scan Sessions within 2 weeks
  - **Auditory oddball task (68 controls and 66 patients) analyzed.**
  - SIRP task
  - Sensorimotor task

FBIRN: Auditory Oddball Task

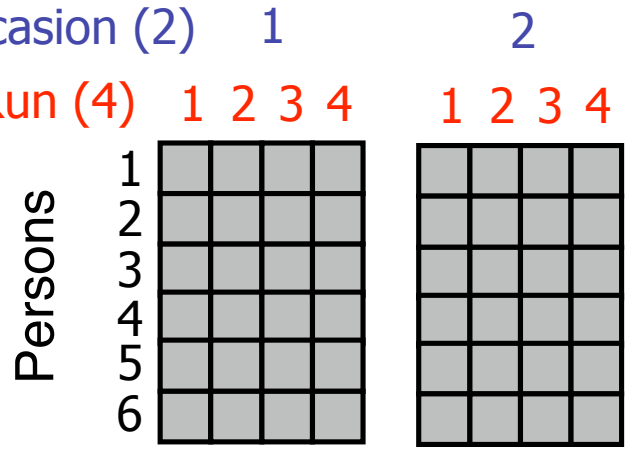
5% Target tones, to which subjects press response button.

95% Standard tones.



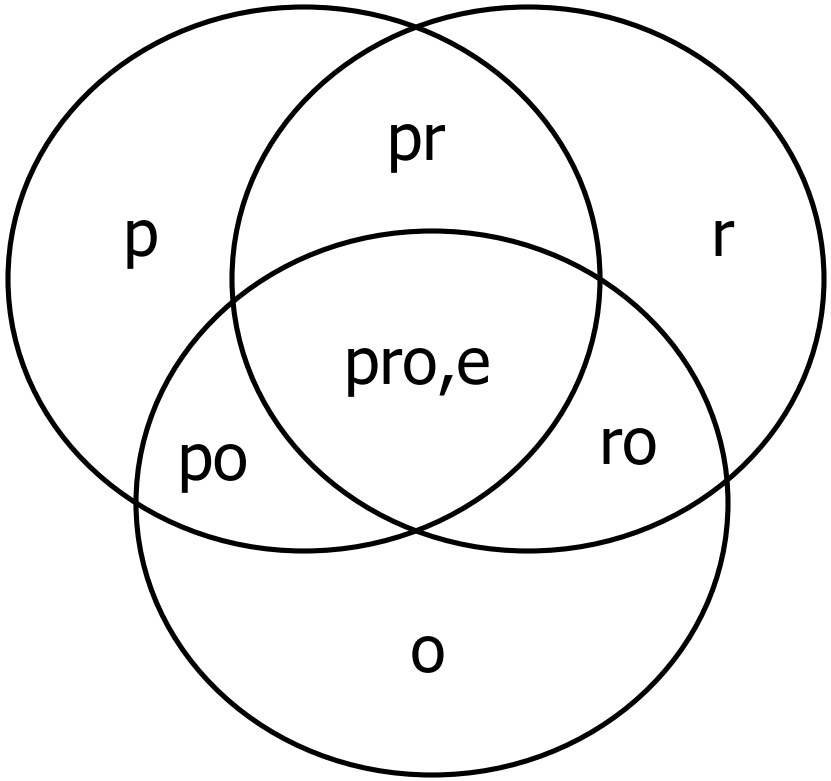
G-Study Design Within a Site

Persons x Occasion x Run



Sources of Variance for 2 facet crossed design

Person x Run x Occasion

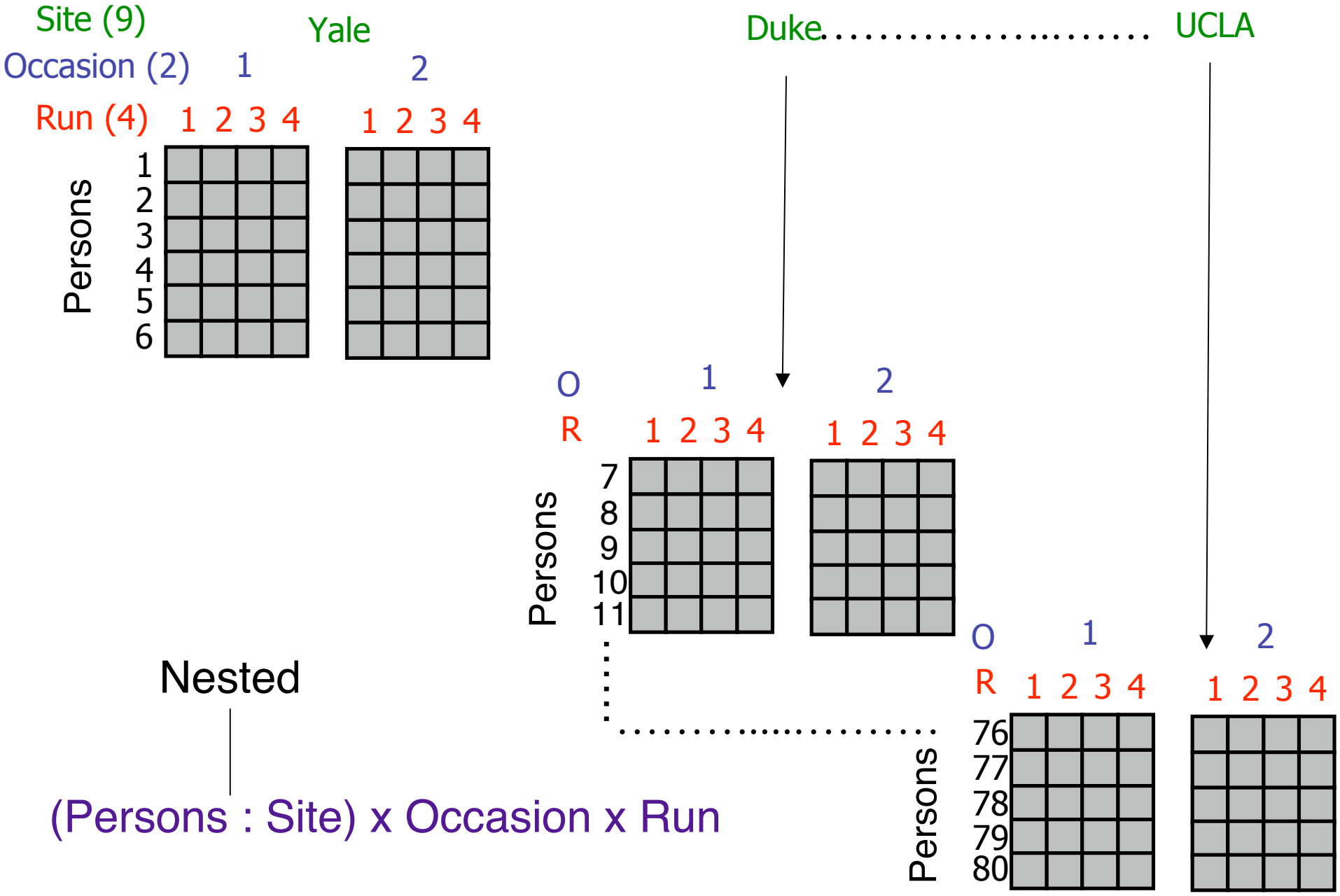


Estimable Variance Components from ANOVA Model:

p
o
r
pr
po
ro
pro, e

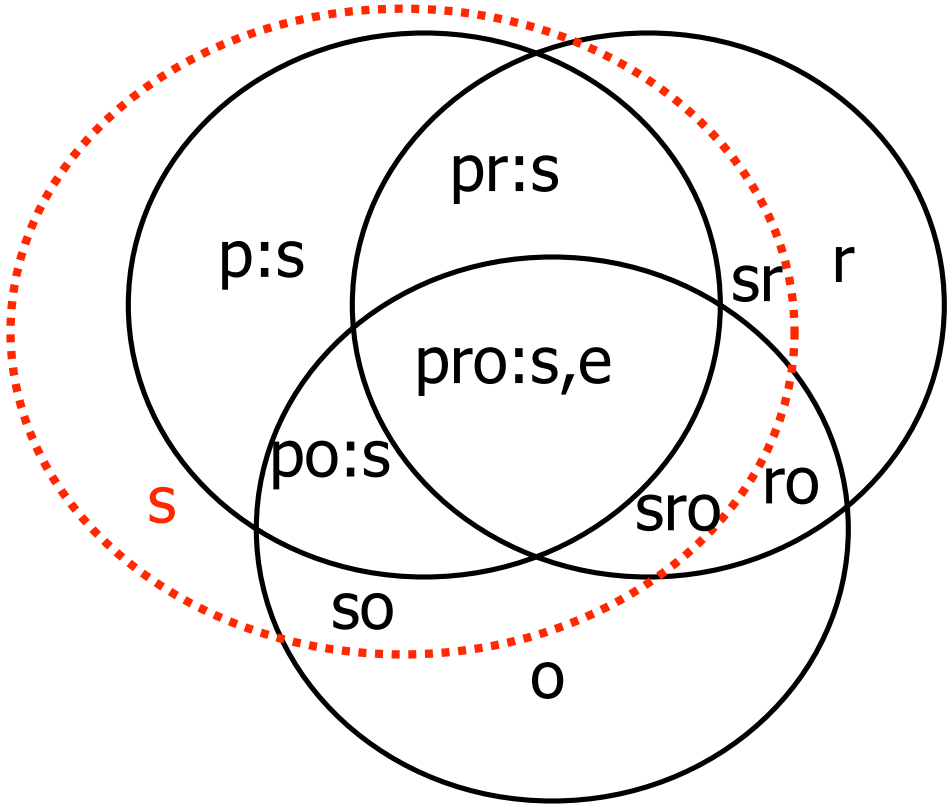
This is the reliability study design when considered within each site.

fBIRN G-Study Design



Sources of Variance for 3 facet mixed design

(Person : Site) x Run x Occasion



Estimable Variance Components from ANOVA Model:

p:s
s
o
r
pr:s
po:s
so
sr
ro
sro
pro:s, e

Generalizability vs. Dependability Coefficients

- Only Relative Comparisons Among Persons Important

  Scores have relative meaning; scores have meaning in relation to each other---------> Generalizability Coefficient

  Uses Relative Error Estimate (omit variance due to main effects of run and occasion)

  Site? Part of True Variance or Error Variance (in this design, can't tell. So, conservatively, treated it as error variance).

- Absolute Estimates Important

  Scores have absolute meaning ------> Dependability Coefficient

  Uses Absolute Error Estimate (includes all variance components except Persons:Site)

Various interpretations of Site and implications for D and G coefficients

Site as a facet in the domain of "instrumentation"

-Site variance contains method variance related to specific scanner, task presentation idiosyncracies, etc.

$$D\_Coefficient = \frac{\mathbf{p:s}}{s + \mathbf{p:s} + o/n\_o + r/n\_r + so/n\_o + sr/n\_r + po:s/n\_o + pr:s/n\_r + or/(n\_r*n\_o) + sor/(n\_r*n\_o) + por:s/(n\_r*n\_o)}$$

$$G\_Coefficient = \frac{\mathbf{p:s}}{\mathbf{p:s} + so/n\_o + sr/n\_r + po:s/n\_o + pr:s/n\_r + or/(n\_r*n\_o) + sor/(n\_r*n\_o) + por:s/(n\_r*n\_o)}$$

Decision study design

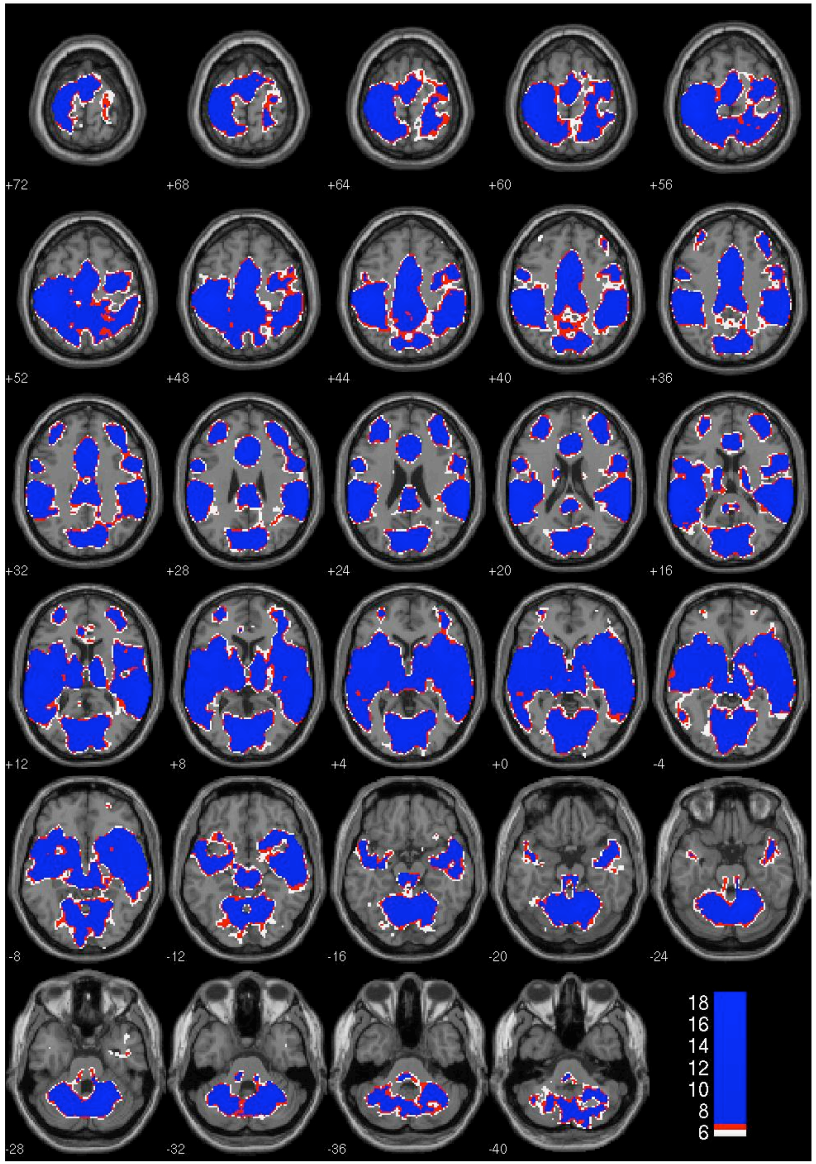
n_o = number of occasions = 1 or 2

n_r = number of runs = 4

Targets-Standards in Healthy Controls

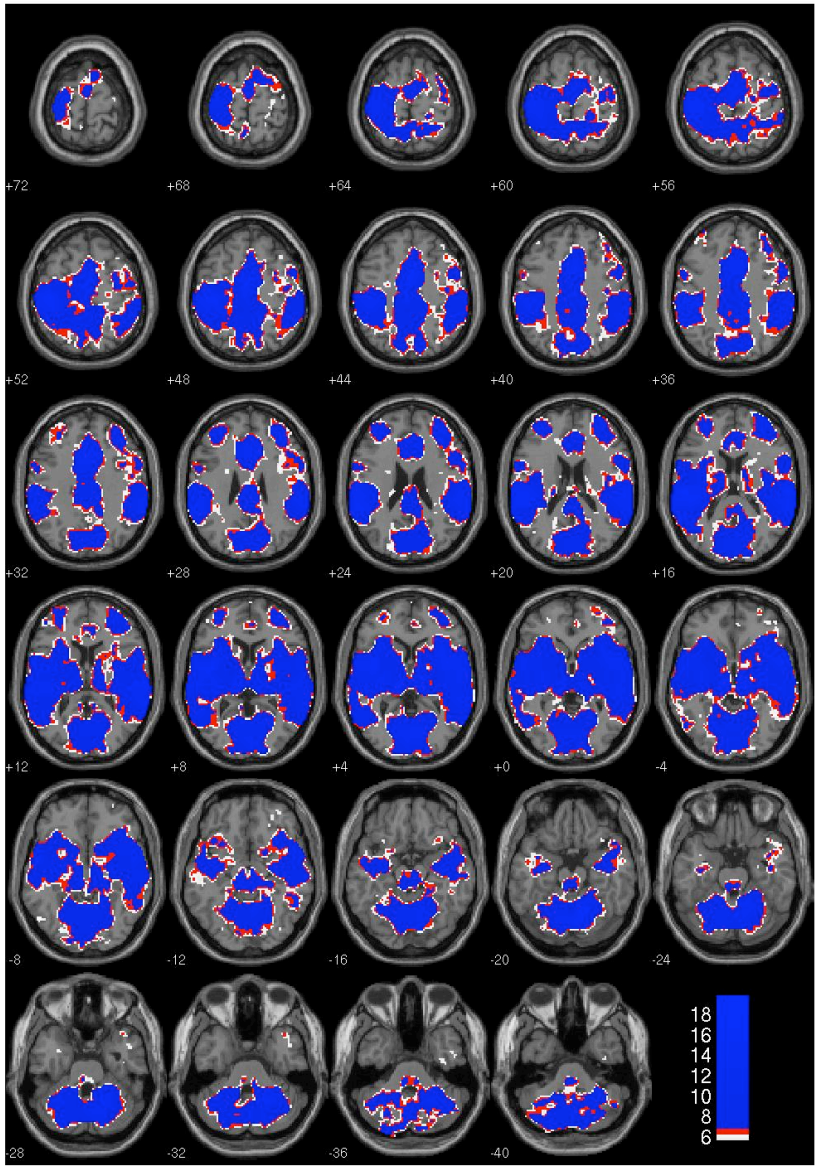
Time 1

Analysis: testretest unweighted visit copes Group: controls t1 cope1 QA1234 Group Contrast: t-test	p=0.01 (FWE) => T=5.6756 (white) p=0.001 (FWE) => T=6.2786 (red) p=0.0001 (FWE) => T=6.8539 (blue) max T=19.0302



Time 2

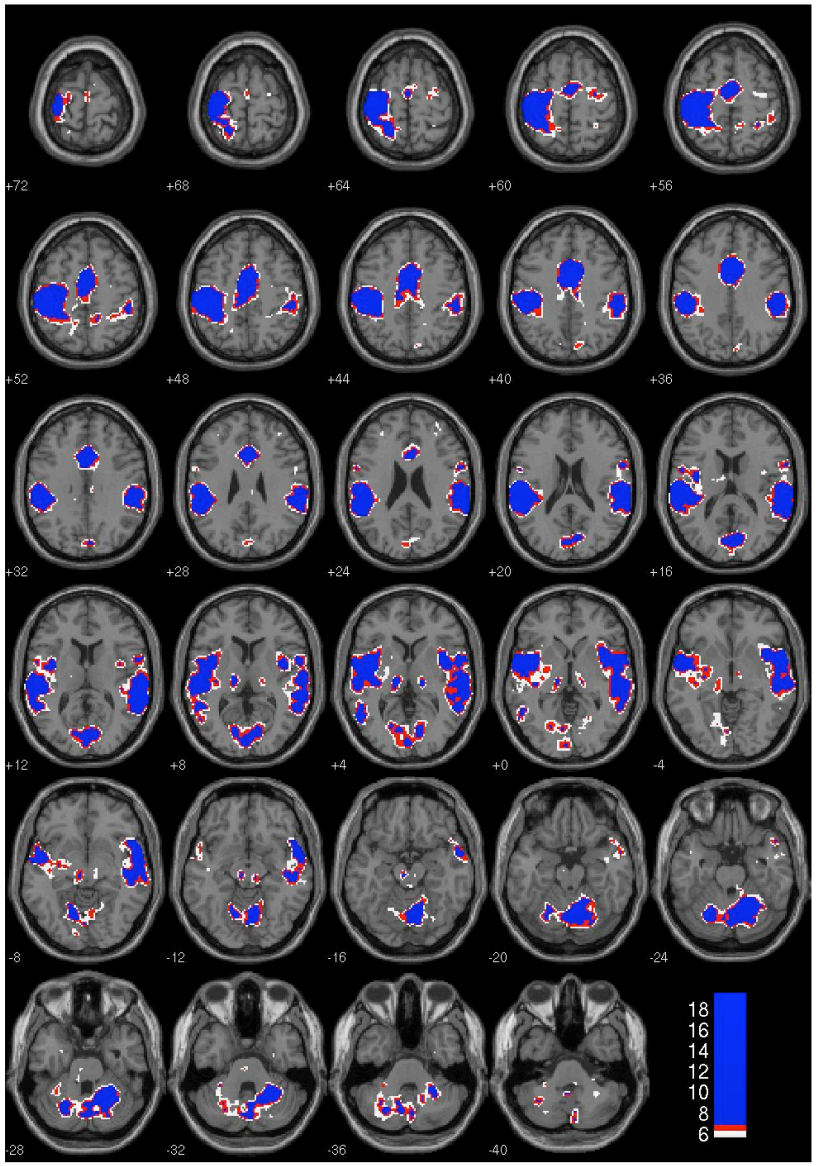
Analysis: testretest unweighted visit copes Group: controls t2 cope1 QA1234 Group Contrast: t-test	p=0.01 (FWE) => T=5.7145 (white) p=0.001 (FWE) => T=6.3155 (red) p=0.0001 (FWE) => T=6.8896 (blue) max T=19.8887



Targets-Standards in Schizophrenia Patients

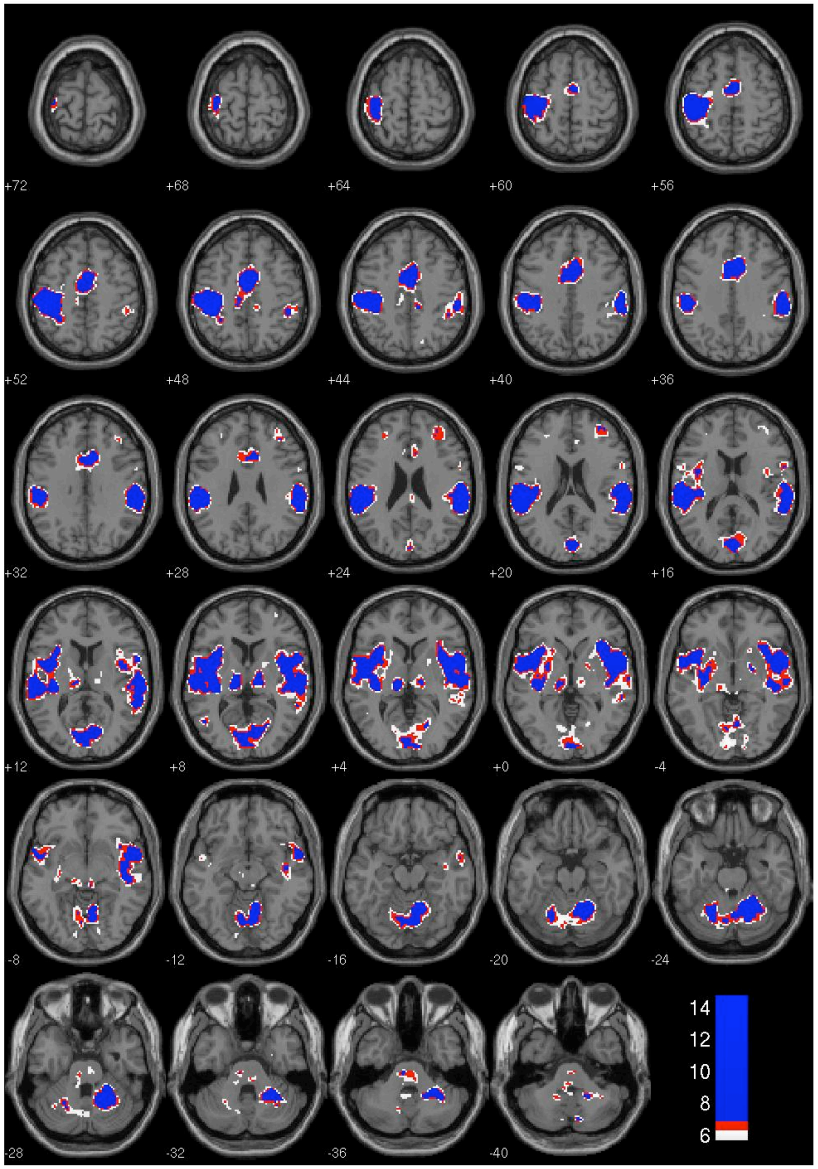
Time 1

Analysis: testretest unweighted visit copes Group: patients t1 cope1 QA1234 Group Contrast: t-test	p=0.01 (FWE) => T=5.7027 (white) p=0.001 (FWE) => T=6.3219 (red) p=0.0001 (FWE) => T=6.9148 (blue) max T=19.6028



Time 2

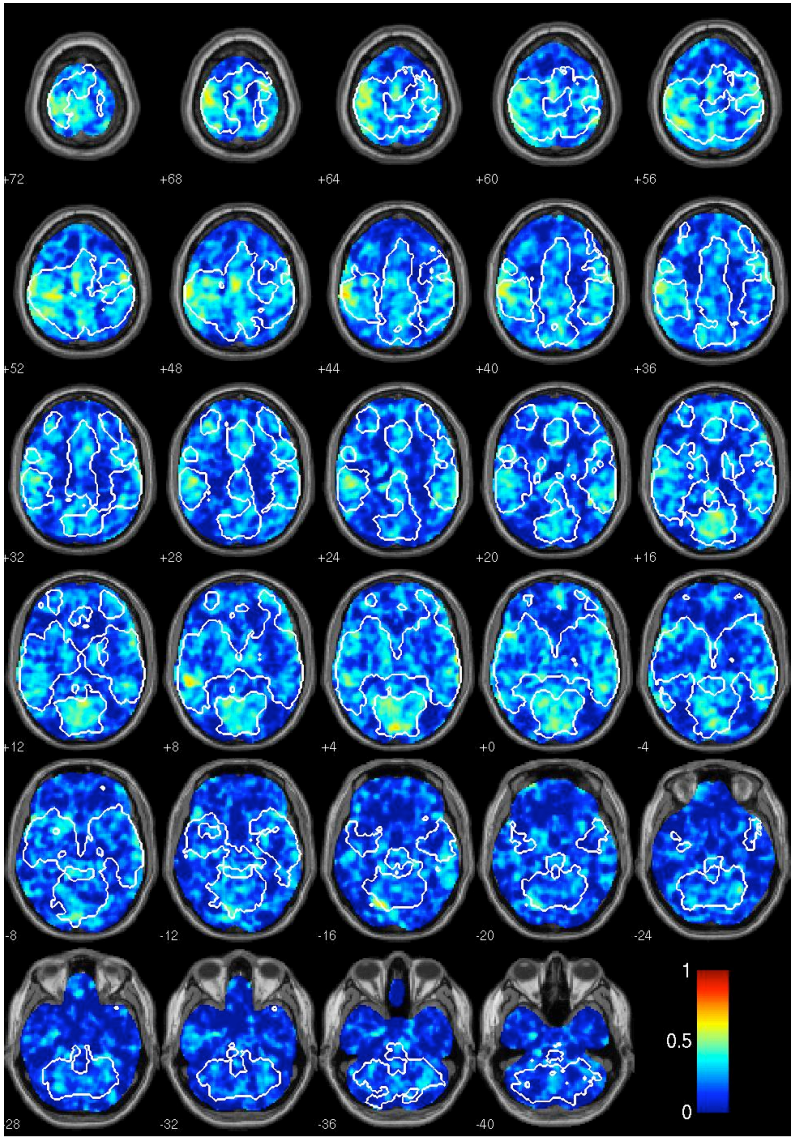
Analysis: testretest unweighted visit copes Group: patients t2 cope1 QA1234 Group Contrast: t-test	p=0.01 (FWE) => T=5.7143 (white) p=0.001 (FWE) => T=6.333 (red) p=0.0001 (FWE) => T=6.9256 (blue) max T=14.783



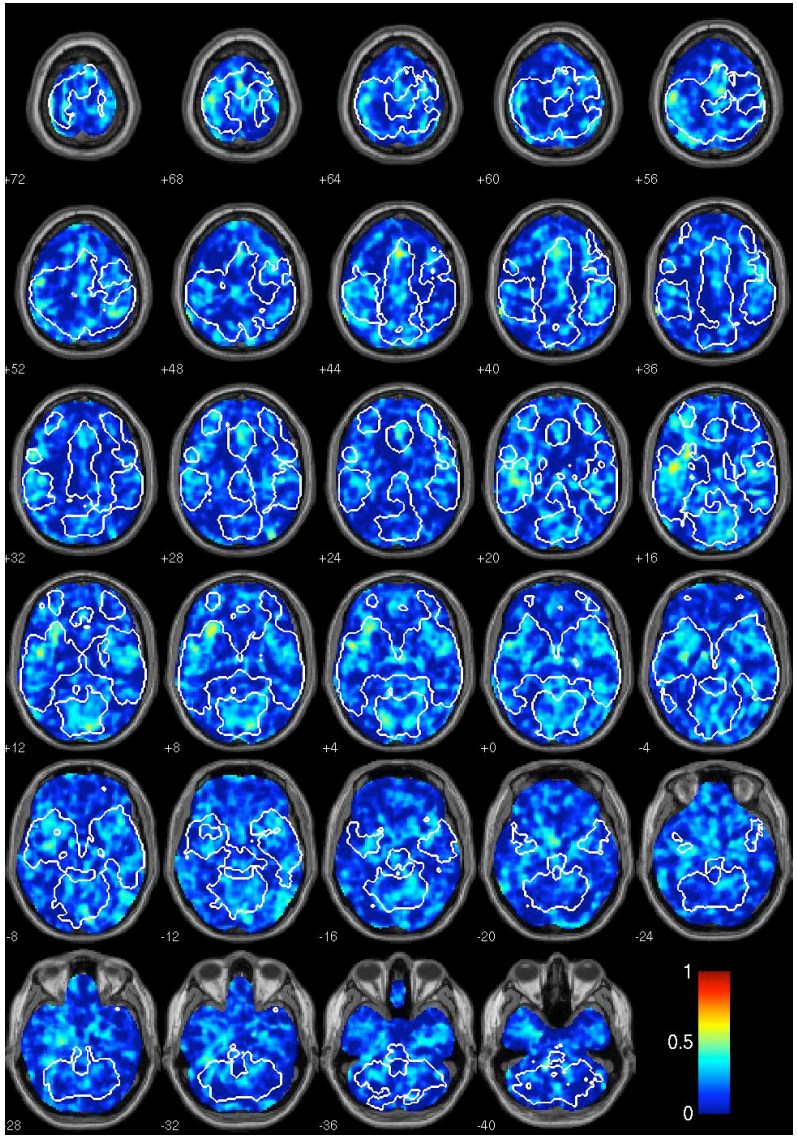
Targets-Standards G-Maps

(with Site in error term,
D-study with 4 runs, 1 occasion)

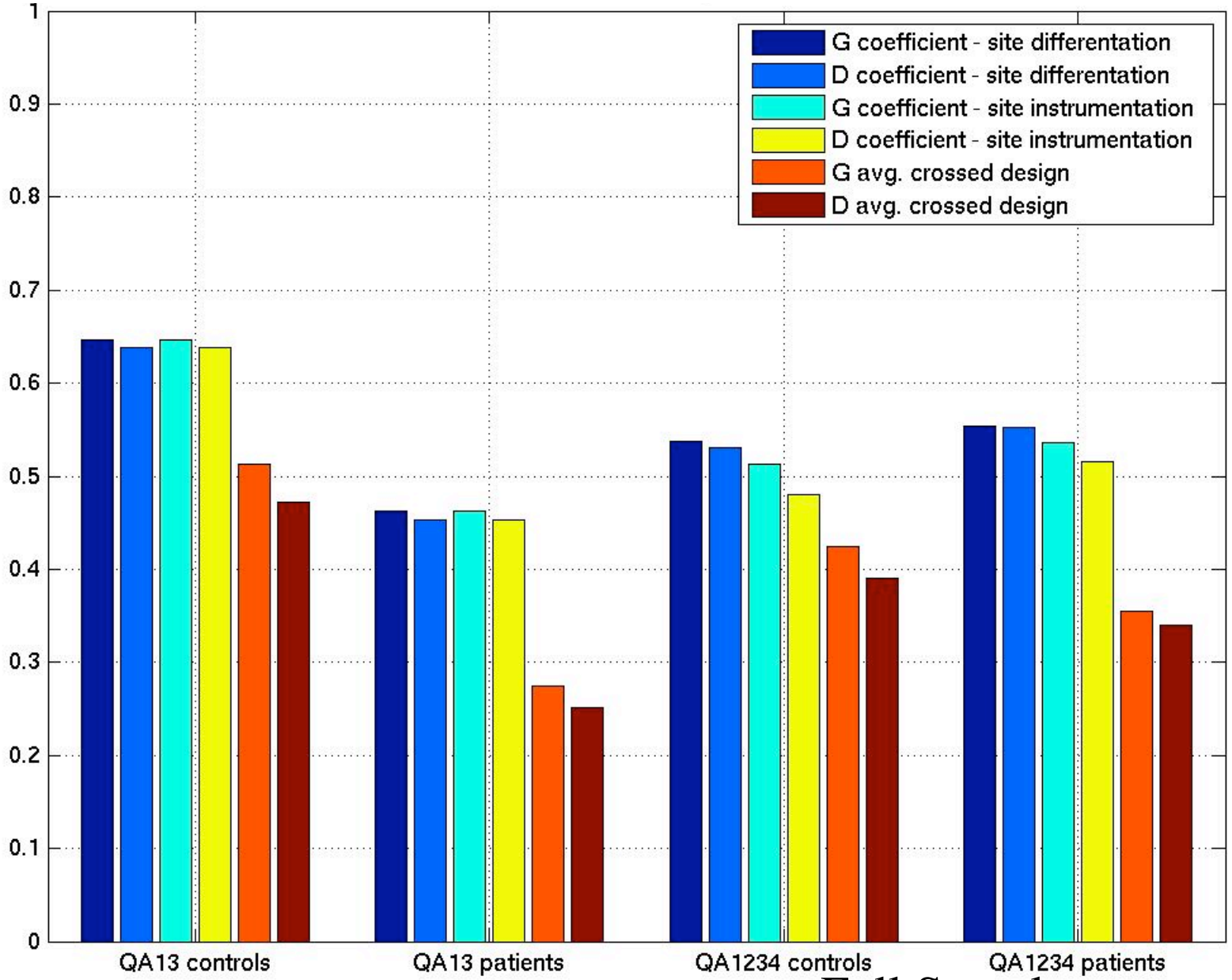
Controls



Patients

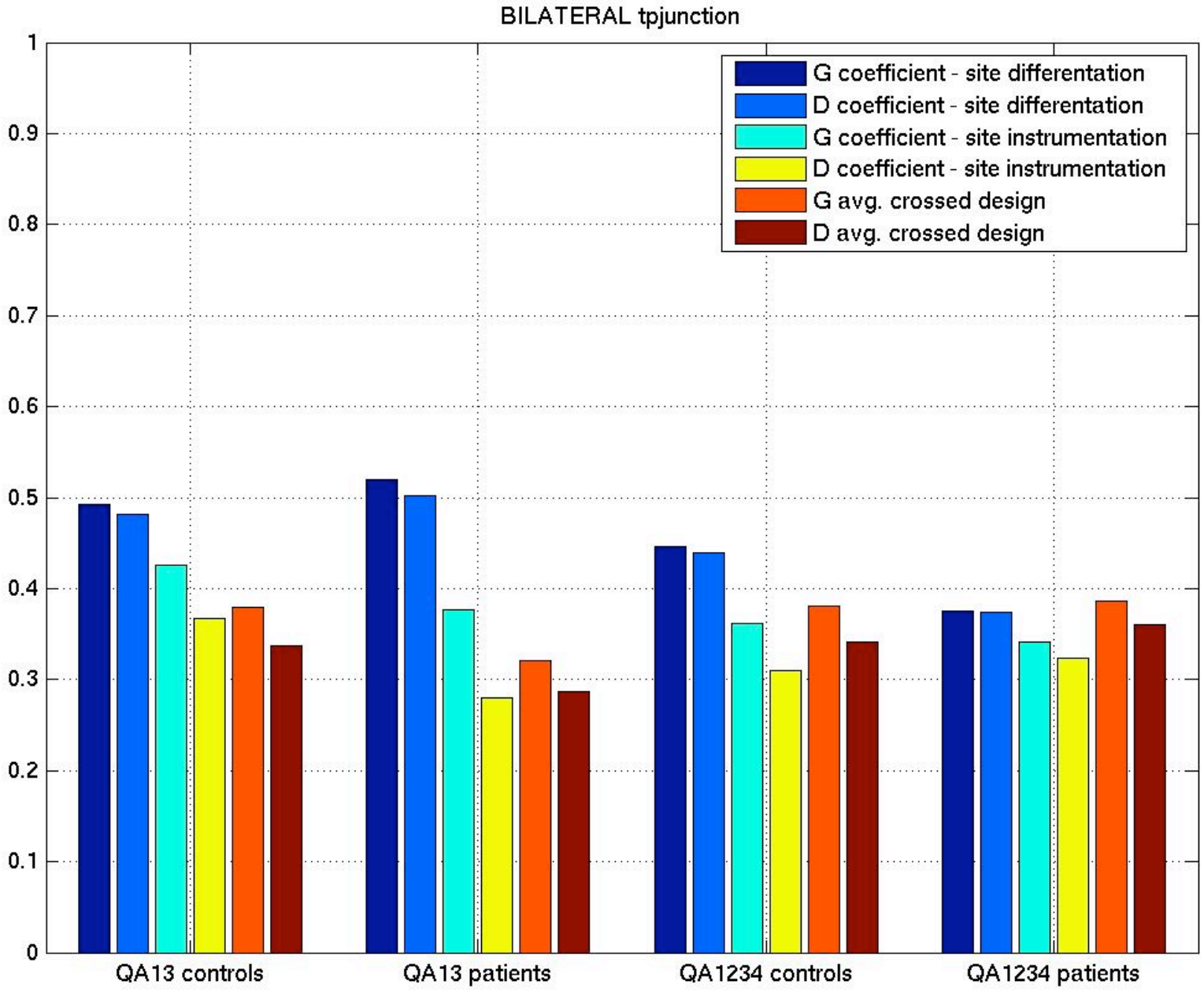


G-Coefficients for Targets-Standards Activation in DLPFC ROI



Better Data Quality Sub-Groups

Full Samples

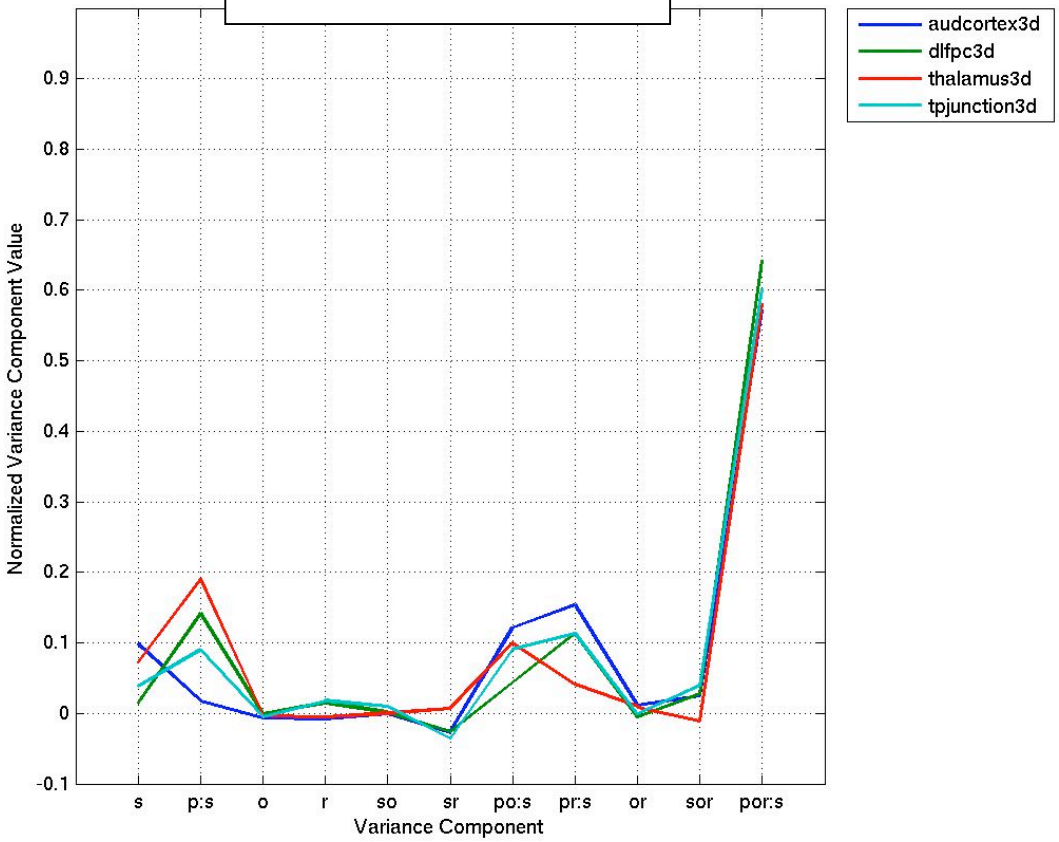


Better Data Quality Sub-Groups

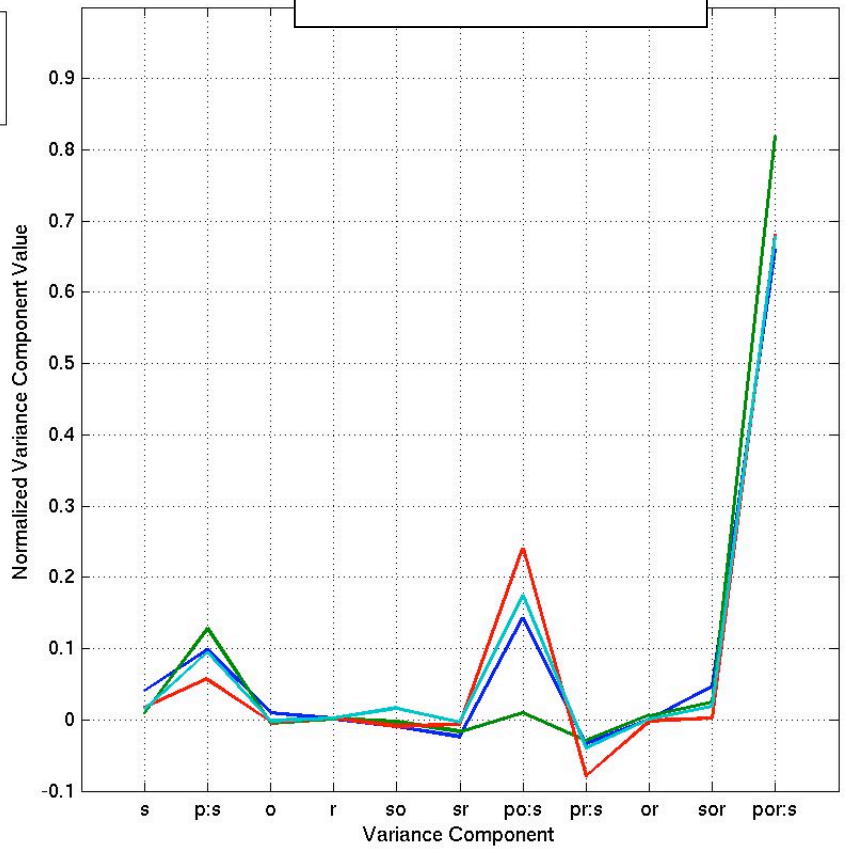
Full Samples

ROI Variance Component Estimates

Controls



Patients

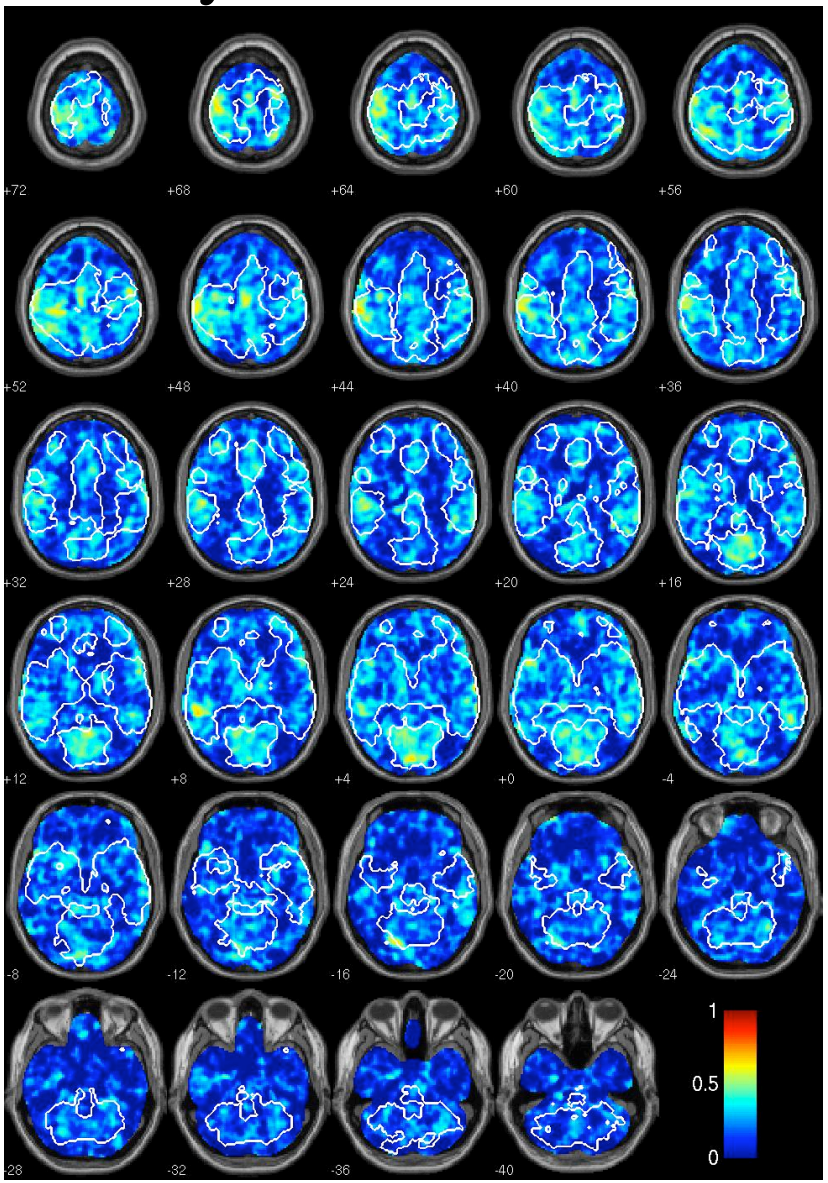


• Relative to controls, patients tend to show less Person variance, more Person x Occasion variance (except for DLPFC), less Person x Run variance, and more residual variance.

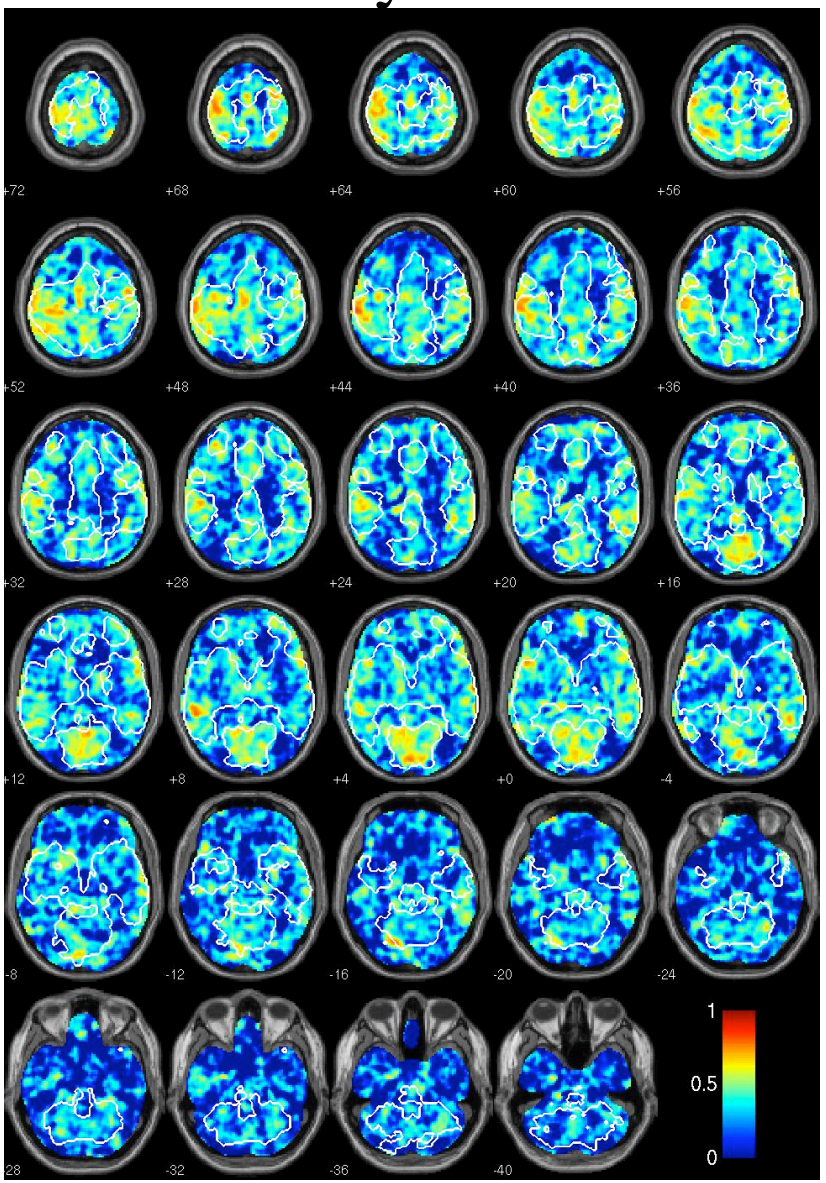
Targets-Standards G-Maps in Controls

(with Site in error term,
D-study with 4 runs)

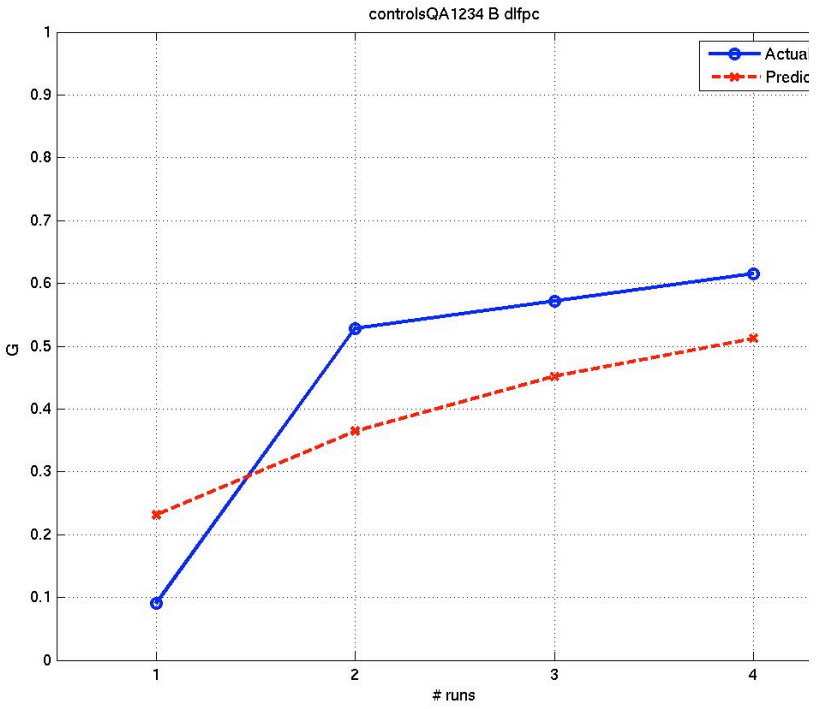
D-study: 1 Occasion



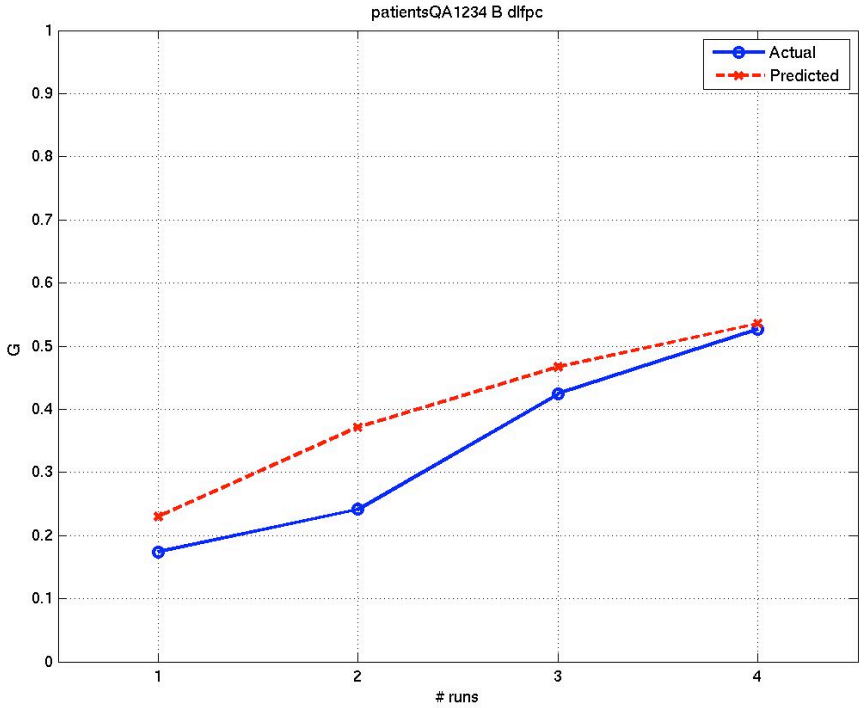
D-study: 2 Occasions



Actual vs. Predicted G-Coefficients in DLPFC based on number of runs in Controls and Patients



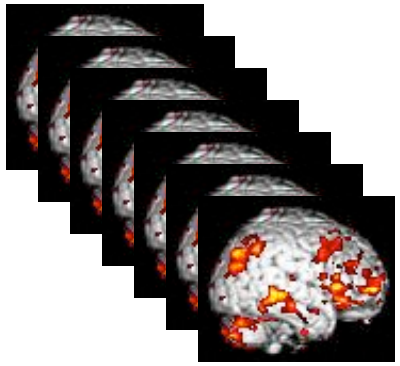
Healthy Controls



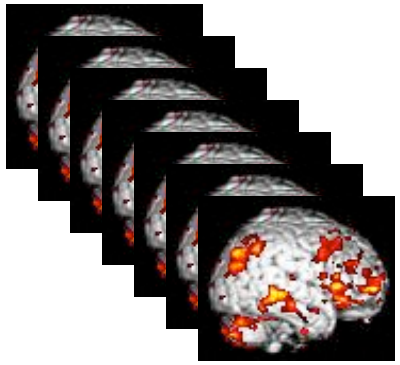
Schizophrenia Patients

Intra-subject, cross-voxel, test-retest reliability

Subjects (Persons)
session 1



Subjects (Persons)
session 2



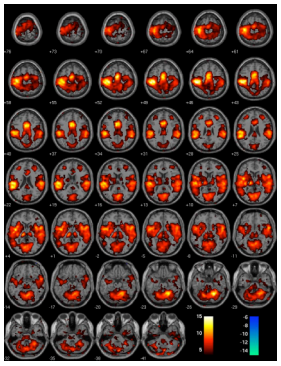
For each voxel

	S1	S2
P1		
P1		
.		
Pn		

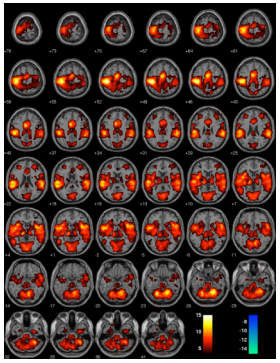
→ **ICC MAP**
ICCs for each voxel across subjects

Person 1

session 1



session 2

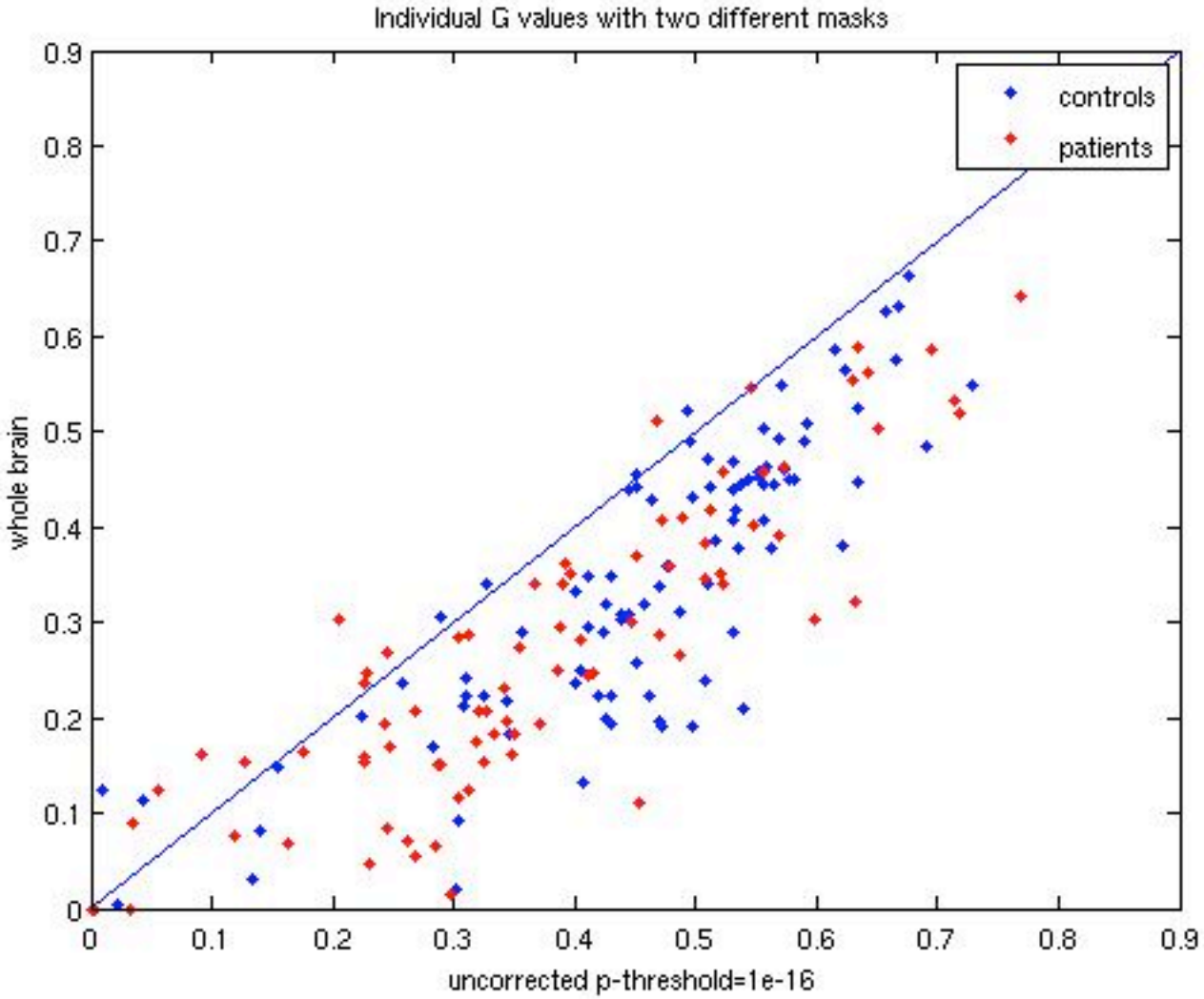


For each subject

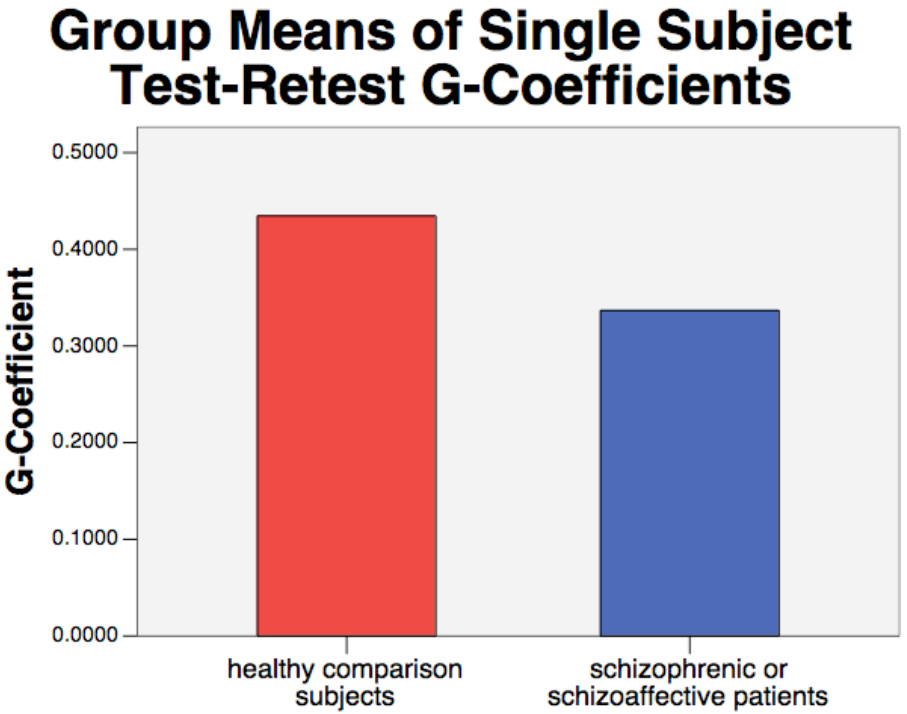
	S1	S2
V_{xyz1}		
V_{xyz2}		
.		
V_{xyzn}		

→ **Intra-subject, cross-voxel, ICC**
ICC for each subject across voxels

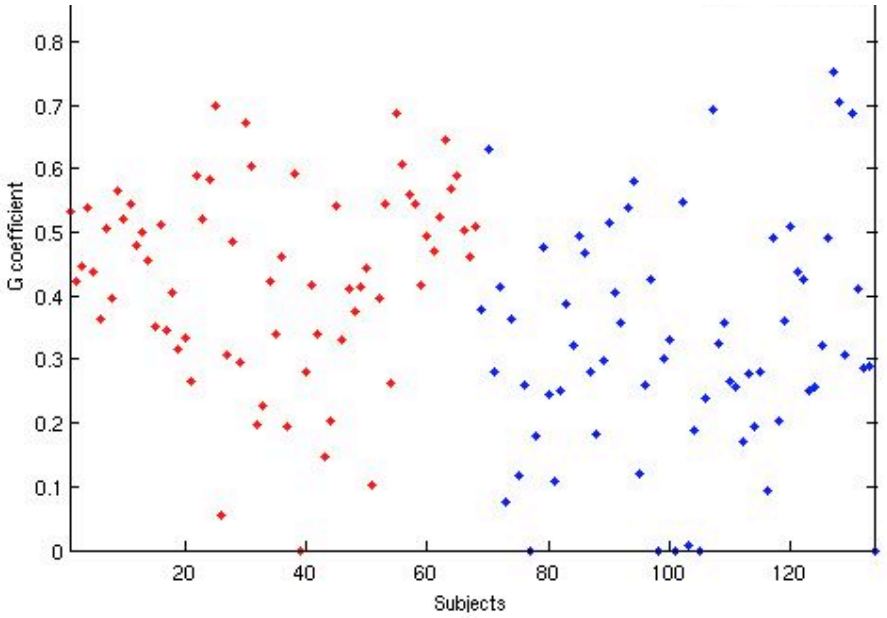
Intra-Subject Time 1 vs. Time 2 G-Coefficients for Targets - Standards based on Whole Brain vs. Union Activation Masked Brains



For Cope 1, Targets-Standards



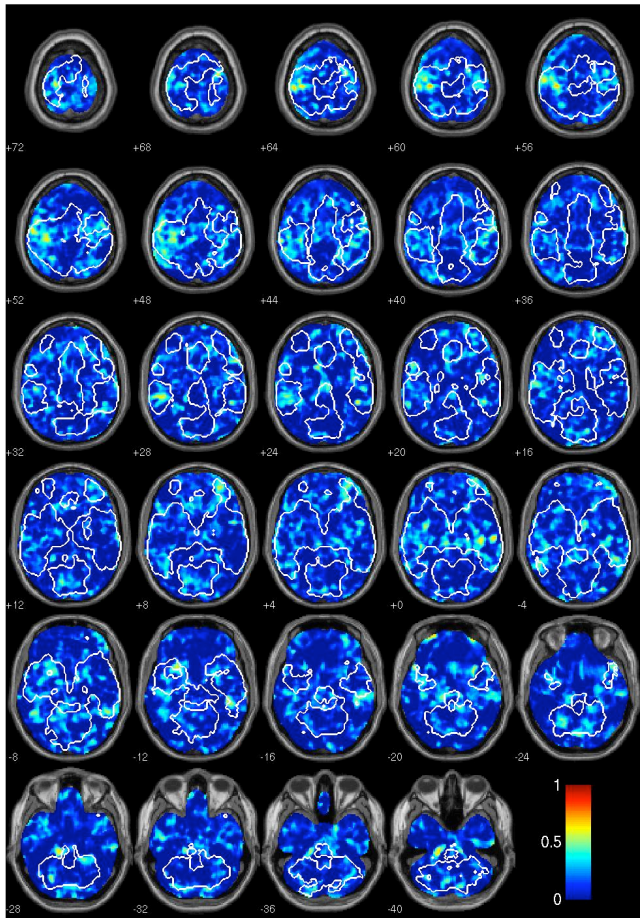
Group Effect F (1,132)=14.75, p<.001



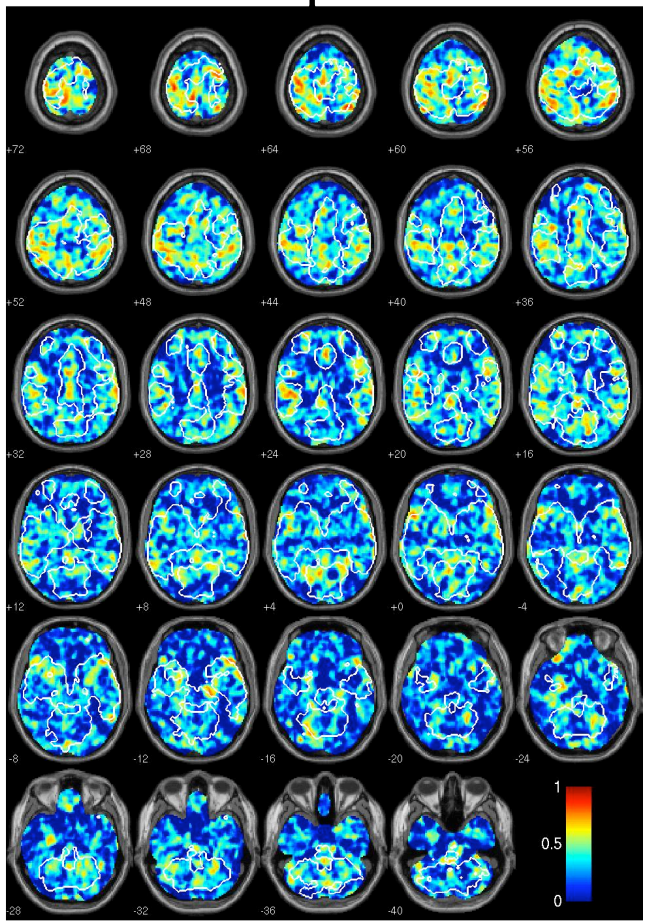
Patients n=66
Controls n=68

G-Coefficient Maps in bottom and top 25% Intra-subject G Coefficient Groups (Healthy Controls)

Bottom 25%



Top 25%



High vs. Low Reliability Subjects
(Based on Intra-Subject, Cross-Voxel G-Coefficients)

Top 25% of Subjects

One-Sample T-test Maps for Mean of T1 & T2 COPE 1 Images* (Targets-Standards)

P<.001, FDR

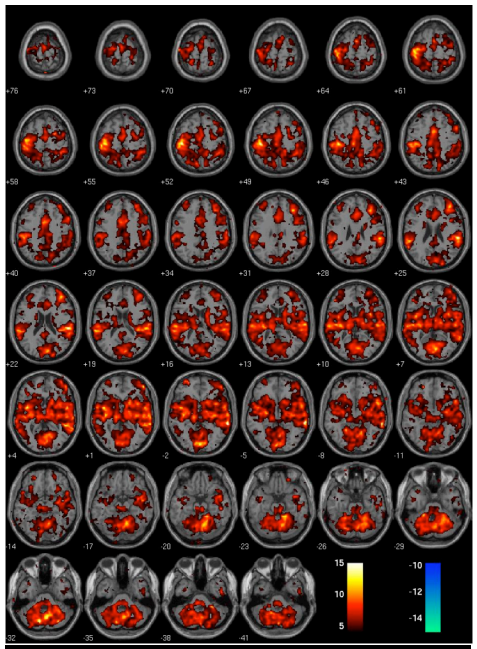
*Variance weighted

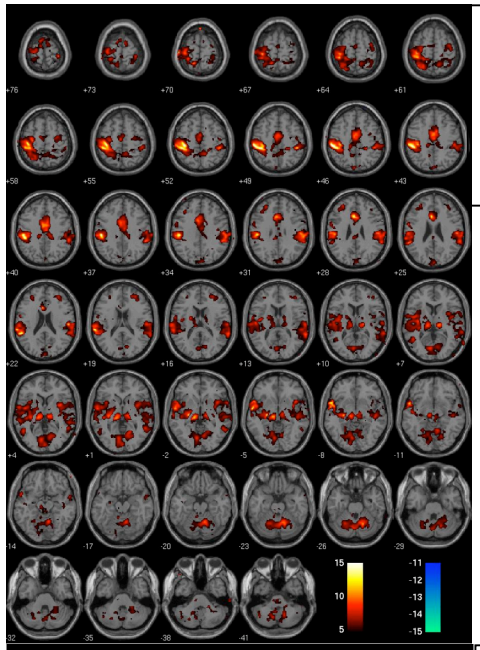
Bottom 25% of Subjects

Healthy Controls (n=17)

Schizophrenia Patients (n=17)

NC > SZ

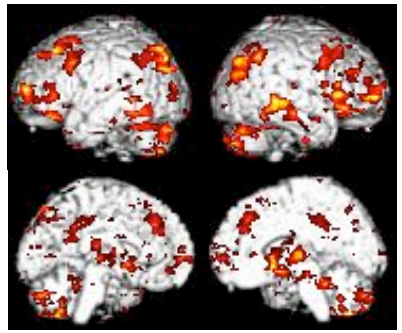


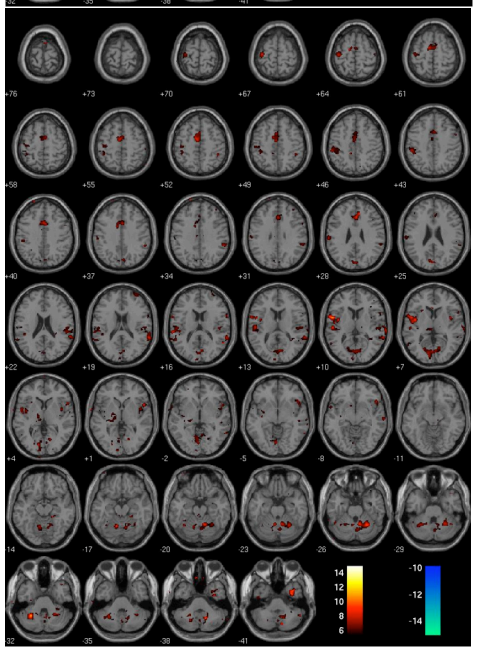


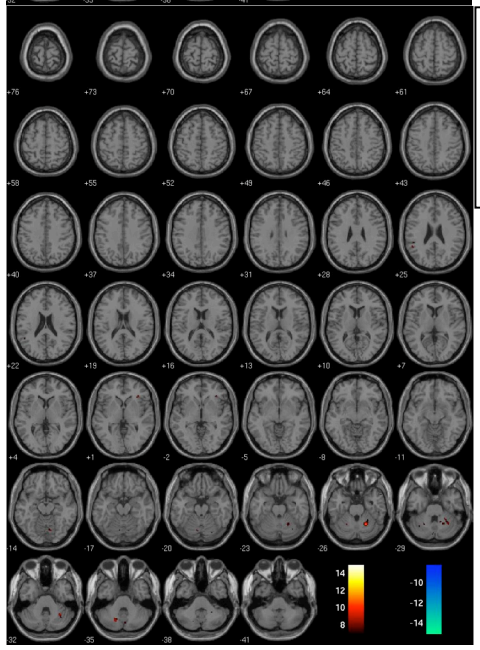
2 Sample t-test

NC > SZ

P<.005, UNC, ext. 6



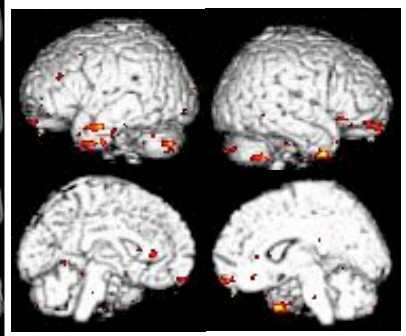




2 Sample t-test

NC > SZ

P<.005, UNC, ext. 6

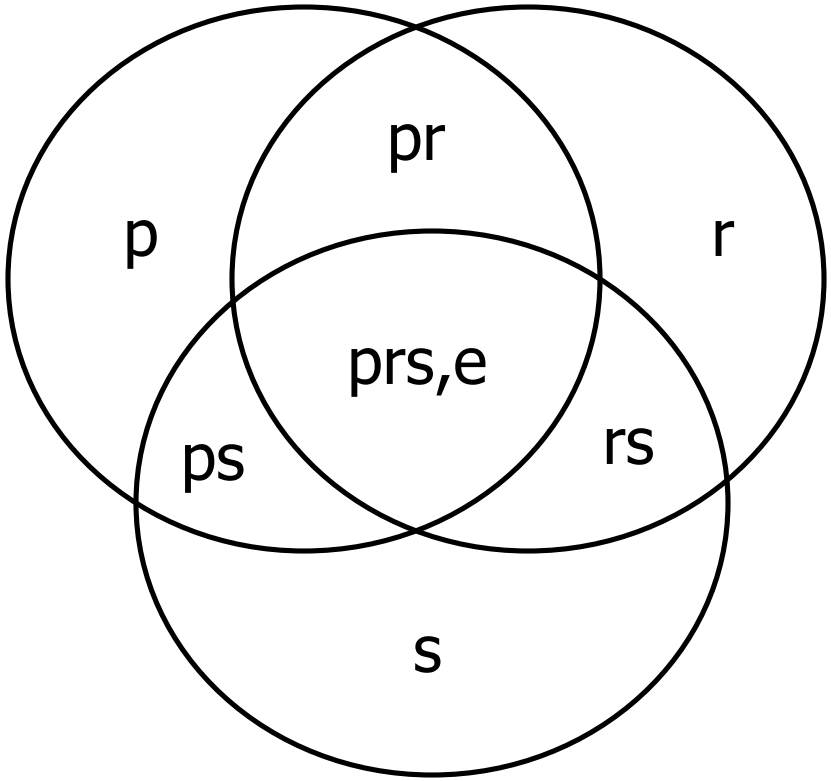


FBIRN
East Coast Traveling Subjects Study (n=18 Healthy Controls)

FBIRN Working Memory Emotional Distraction Task

Sources of Variance for 2 facet crossed design: ECTS

Person x Run x Site



SITES:
Yale x 2
Harvard
UNC/DUKE

Estimable Variance Components from ANOVA Model:
P
r
s
ps
pr
rs
prs, e

Emotional Working Memory Task: Emotional Distraction

Scramble Picture Baseline

Encode

Maintain
Emotional or Neutral Distractors

Forced Choice Memory Probe

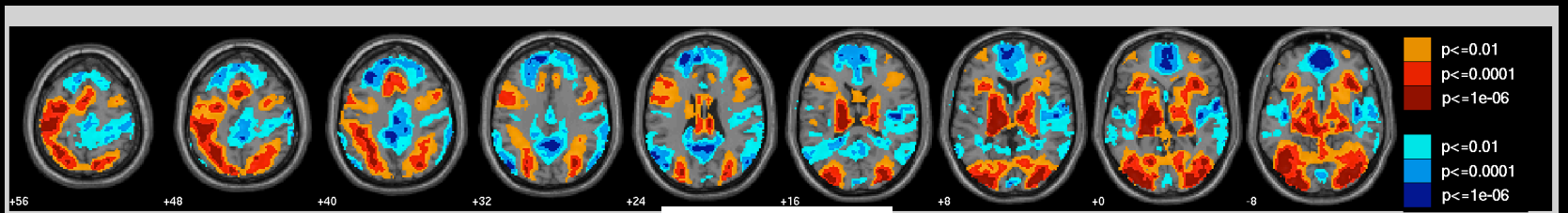
Scramble Picture Baseline

Stimulus Item

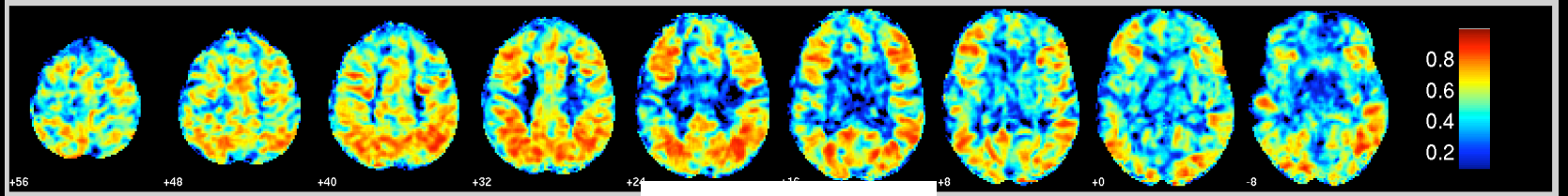
Eight pictures presented during the encode period.

Eight picture pairs presented during the forced choice period.

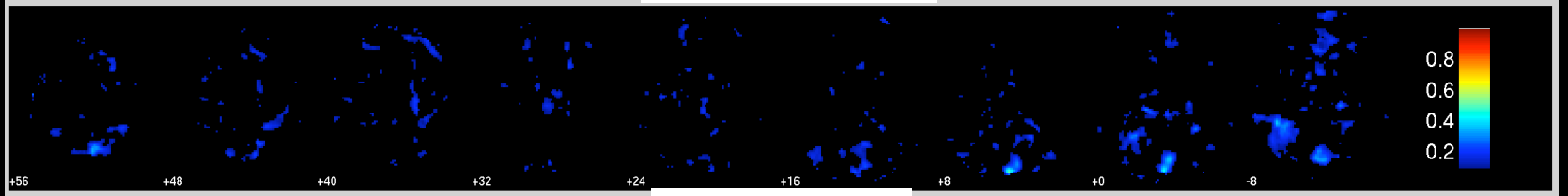
Average Probes vs. Scramble Pictures Contrast



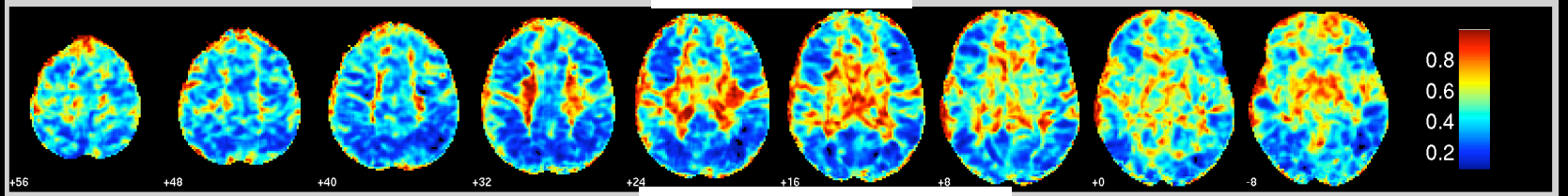
Activation Map



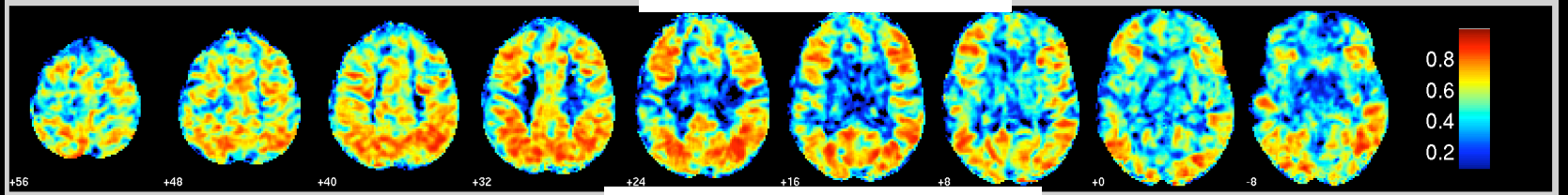
Person Variance Map



Site Variance Map



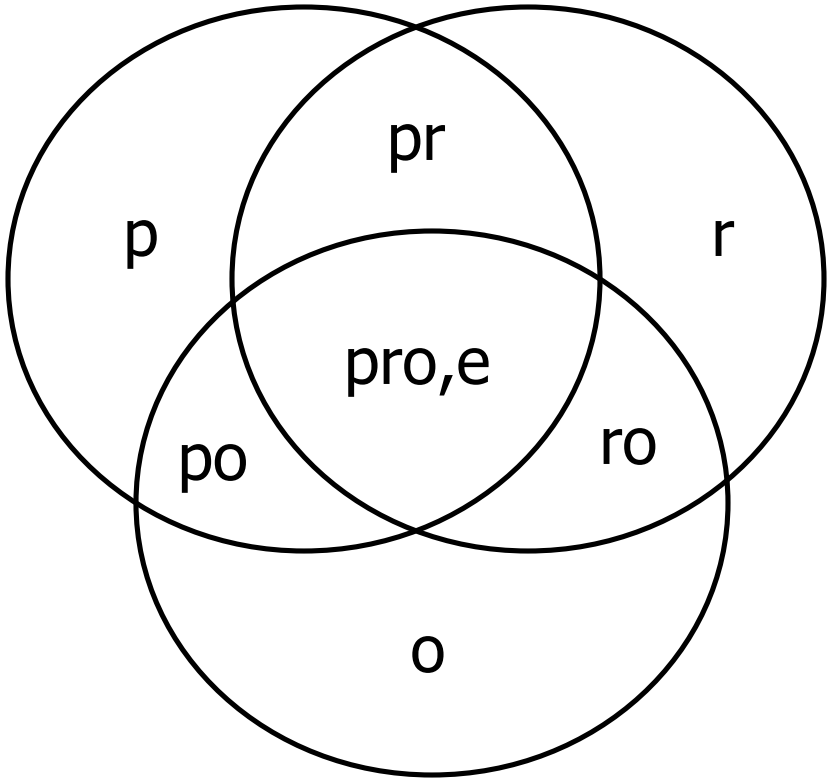
Person x Site Variance Map



Generalizability Coefficient Map

Sources of Variance for 2 facet crossed design: Yale Data

Person x Run x Occasion



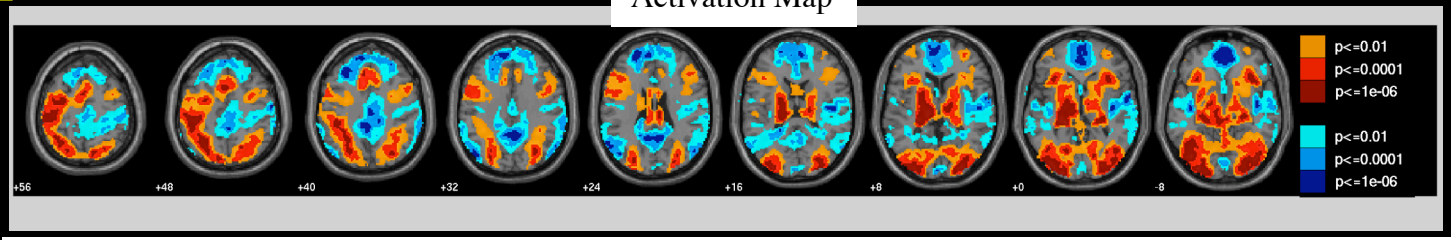
Estimable Variance Components from ANOVA Model:

p
o
r
pr
po
ro
pro, e

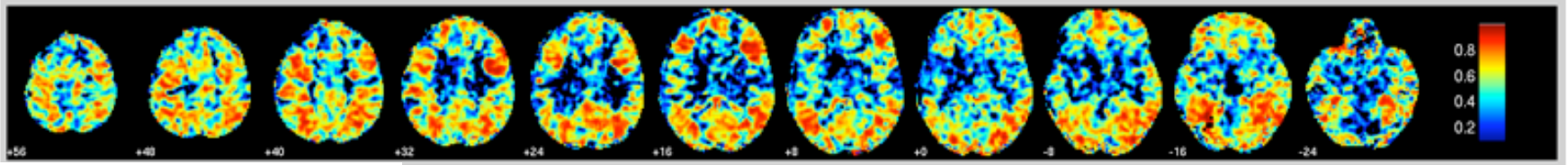
This is the reliability study design when considered within each site.

Average Probes vs. Scramble Pictures Contrast: Yale Time 1 vs. Time 2

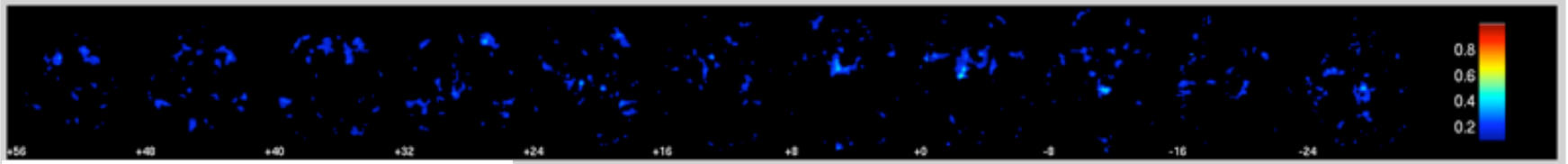
Activation Map



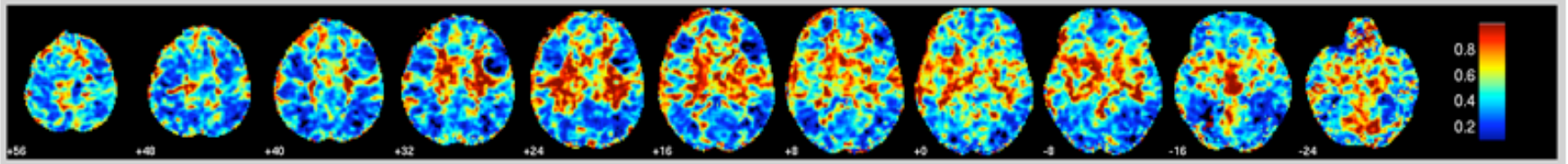
Person Variance Component



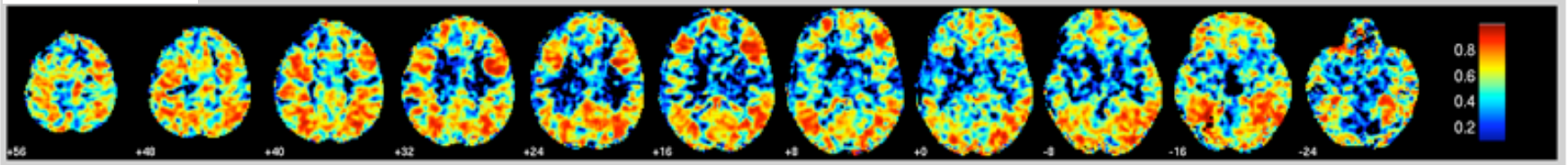
Occasion Variance Component



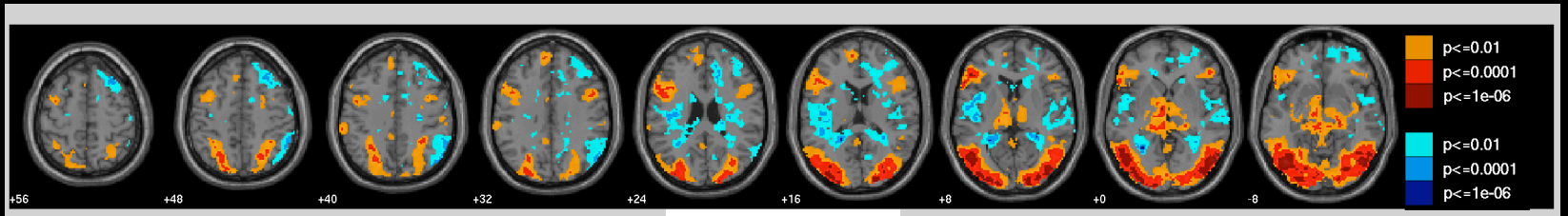
Person x Occasion Variance Component



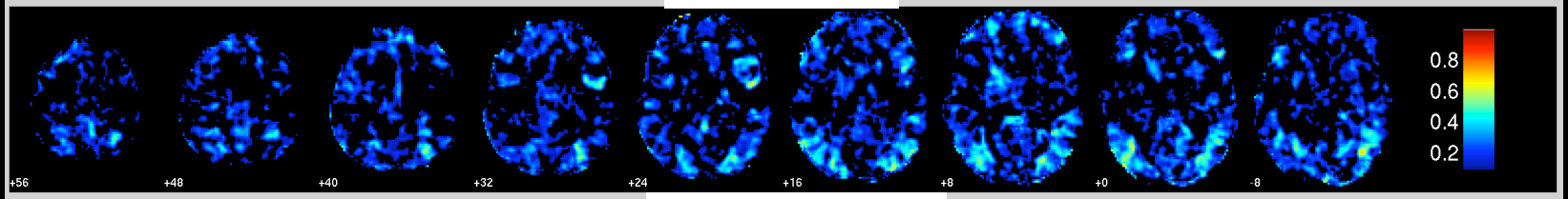
G-Coefficient



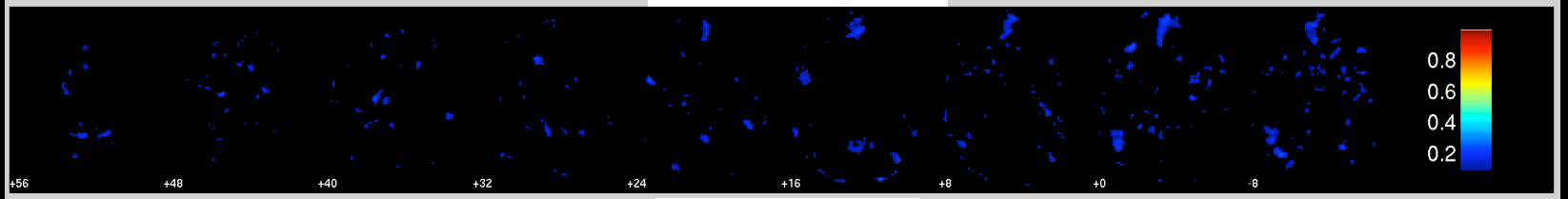
Emotional vs. Neutral Distractors Contrast



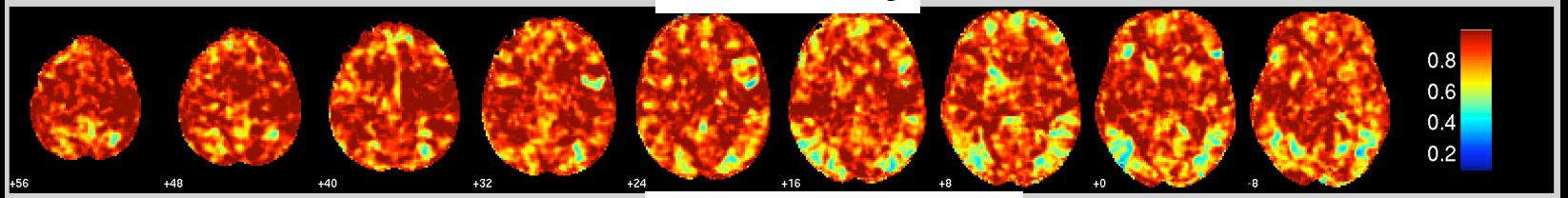
Activation Map



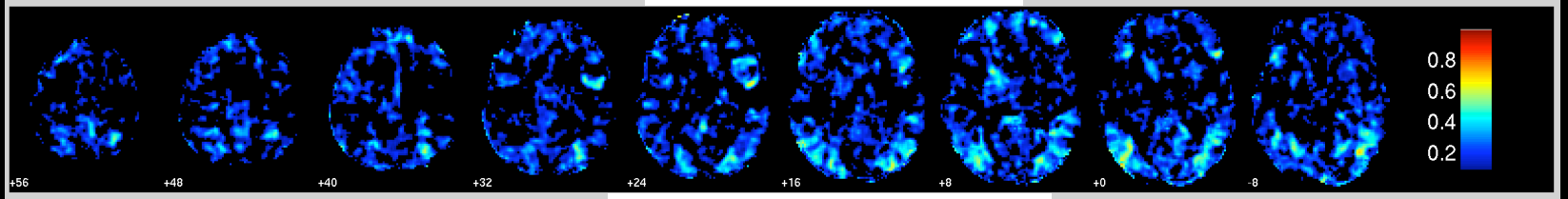
Person Variance Map



Site Variance Map

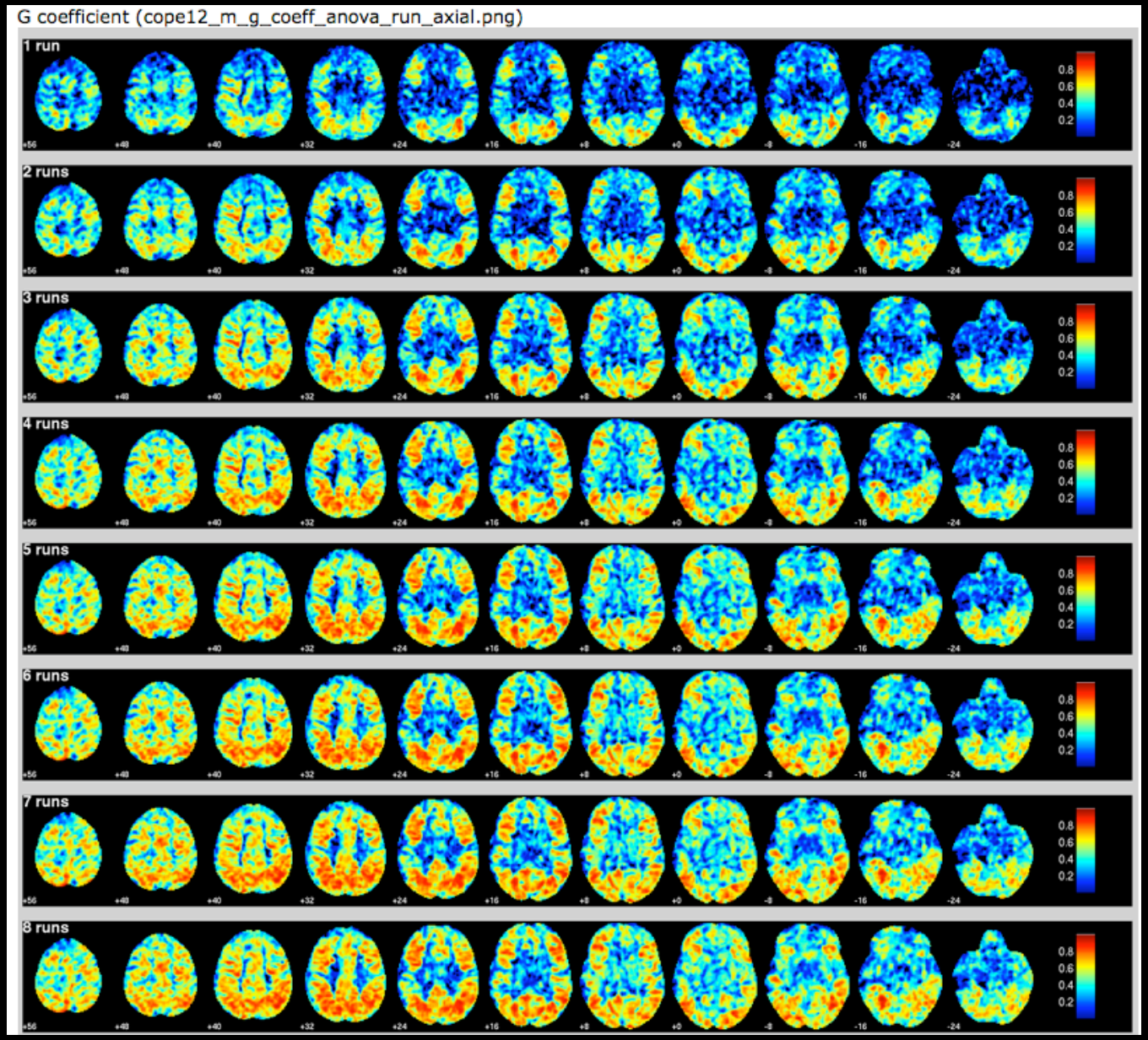


Person x Site Variance Map

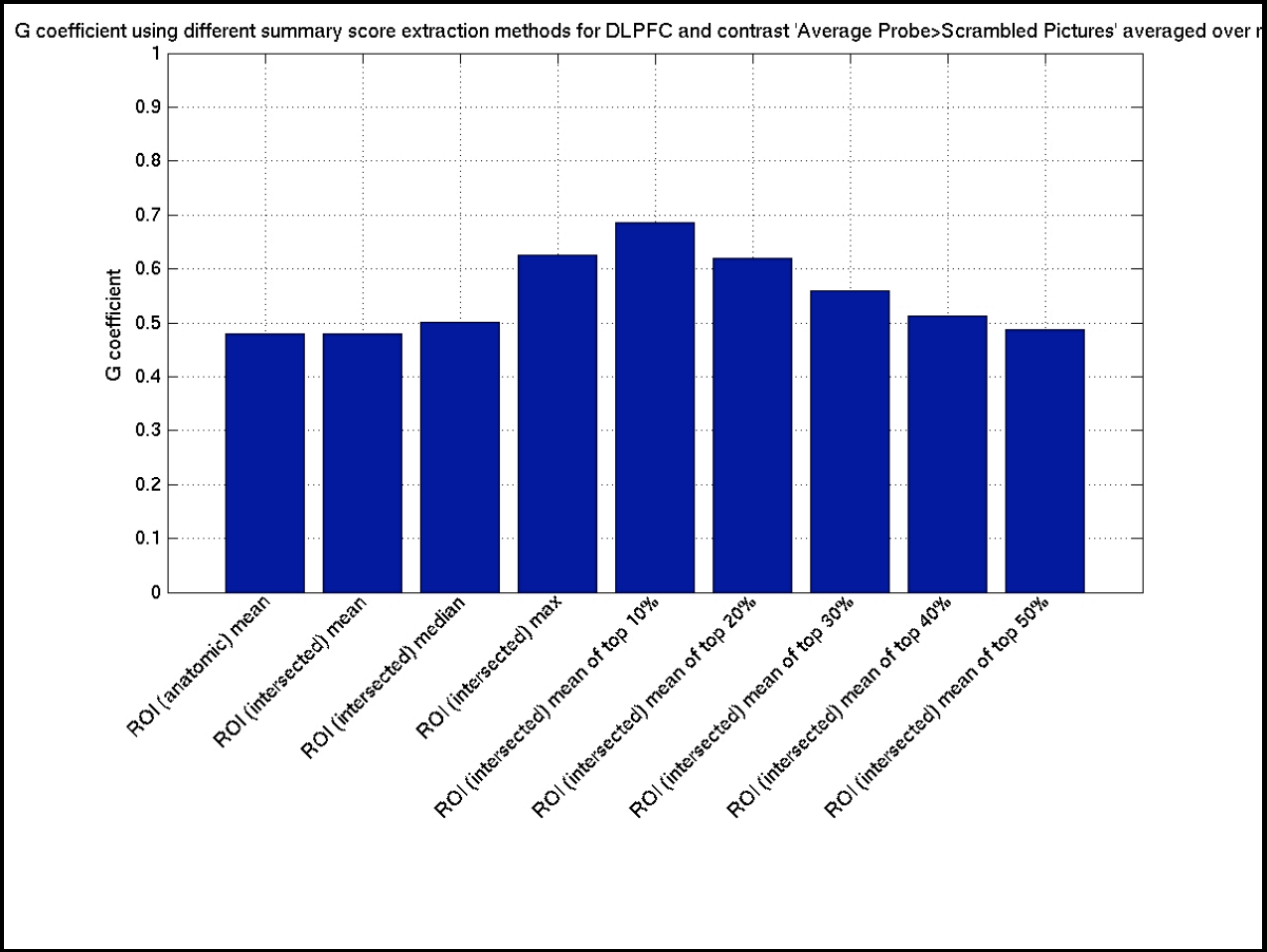


Generalizability Coefficient Map

G-Coefficients for Average Probes vs Scramble Pictures: Increase with Number of Runs



G-Coefficient for Average Probe vs. Scramble Picture Baseline ROI: DLPFC



Conclusions I

- Task Related fMRI can be as reliable as other measures used in psychiatric research (including clinical ratings).
- Reliability is specific to
  - Task
  - Contrast within task
  - Region activated
  - Type of measure extracted (magnitude, extent of activation in ROI, peak in ROI, mean in ROI, etc)
  - Population studied (Sz patients are less reliable than controls).

Conclusions II

- Reliability can be increased by increasing the number of runs (to a point).
- Reliability can be increased by averaging over two or more fMRI occasions (not very practical).
- FMRI can be reliable across sites:
  - Many steps taken to improve reliability, including limiting to 3 T field strength magnets, standardization of methods.
  - fMRI reliability can be improved by reducing noise in the data (effects of data quality not shown).
  - Removal of true variance unrelated to questions of interest can reduce reliability but improve criterion validity.

Acknowledgements

Kasper Jorgensen at UCSF
Greg Brown UCSD
FBIRN Consortium