Supplemental Information

Classical Test Theory

It is useful at the outset to more formally define reliability using concepts derived from classical test theory (CTT). According to classical test theory, an individual's observed score, Xi, on a measure X, comprises two components: a true score component (Ti) and a random error component (Ei). Thus,

$$X_i = T_i + E_i \qquad (1)$$

The reliability of the measure X, r(X,X'), is defined as the ratio of its true score variance, Var(T), to its observed score variance, Var(X):

$$r(X,X') = \frac{\text{Var}(T)}{\text{Var}(X)} \qquad (2)$$

Since we cannot observe true scores directly, it is impossible to directly compute true score variance. Instead, we estimate it by computing the correlation between repeated measurements obtained from the same set of individuals. Because we assume that measurement error is random across instances of measurement, the only way for a measure X to be correlated with itself on second measurement occasion (X') is if the measure's true scores are correlated. Accordingly, we can estimate the numerator and denominator of the reliability coefficient by considering the correlation between the same measure on two occasions:

$$r(X,X') = \frac{\text{Covariance}(X, X')}{\text{SD}(X) * \text{SD}(X')}$$

where SD denotes the standard deviation. Inspecting this equation, it is clear that 1) the numerator estimates Var(T), since the only parts of X and X' that share variance (i.e., that covary) are their true scores, which are assumed to remain stable from test to retest; and 2) the denominator estimates Var(X), since SD(X) and SD(X') are expected to be essentially the same, and therefore their product is equal to Var(X).

The Relationship Between Generalizability Theory and Intraclass Correlation Coefficients Based on Classical Test Theory

Generalizability (G) coefficients and Dependability (D) coefficients, derived from Generalizability Theory, are both intraclass correlation coefficients (ICC). The variance components used to estimate these coefficients are derived from ANOVA models. Thus, there is not really a “qualitative difference” between calculating a G-coefficient and an ICC. However, one can distinguish the approaches based on the fact that an ICC reliability coefficient based on classical test theory does not distinguish among different sources of measurement error. Thus, in classical test theory, the ANOVA model is typically limited to a design in which Persons is crossed with a single facet of measurement error (e.g., fMRI task run or scan occasion), and the ICC is equal to the ratio of person variance to the total expected observed score variance (which is defined as person variance + residual variance in the ANOVA model). In contrast, G-theory provides for more complex ANOVA models in which multiple facets of measurement error can be considered in a single design. In estimating a G-coefficient, the investigator decides which variance components to include in the denominator of the ICC based on which facets of measurement error will contribute to between subject variability in the planned scientific study or “Decision Study.” G-theory also distinguishes between relative (i.e., rank ordering of subjects or groups) and absolute (i.e., indexing an individual’s absolute level on the measure of interest) decisions to be made based on subject scores, whereas classical test theory only considers relative decisions. Accordingly, when there is interest in using scores for absolute decisions or judgments, G-theory calls for the estimation of an ICC known as a D-coefficient, which is distinguished from a G-coefficient by its inclusion of the variance components associated with the main effects of the various measurement facets (e.g., task run, scan occasion). Finally, while the Spearman-Brown Prophecy formula can be applied to ICCs in classical test theory in order to project what reliability would be if based on a larger number of measurements associated with a single measurement facet (e.g., fMRI task runs), the G-coefficients and D-coefficients from G-theory provide for such projections to be made for multiple measurement facets simultaneously (e.g., what would my reliability be if I averaged over 8 fMRI runs and two fMRI scan occasions?).

Reliability and Measurement of Change Over Time

A measure must also have adequate test-retest reliability if it is going to be used to track changes over time in a longitudinal study design. However, it should be noted that reliability of a measure over a test-retest interval during which true change has *not* occurred does not address directly the reliability of change scores on that measure. If change scores on a measure are ultimately going to serve as the primary unit of analysis in a study, then the reliability of the change score becomes an important consideration, though the reliability of such change measures is frequently low (1-4). Based on a number of lucid reviews of this topic (5-8), several important points can be made. First, the main reason change scores suffer from poor reliability is that they typically show substantially less true score variability than their constituent scores. As true score variation approaches zero, reliability also approaches zero, even though the change score measurements may be quite precise. Second, the same reduction in true score variability that reduces the reliability of change scores is responsible for enhancing the statistical power of change scores to detect experimental treatment effects (8-12). This is because smaller change score variance translates into smaller values for the denominator, or “error term”, of the test statistics used to evaluate differences between experimental groups.

In practice, it is not always possible to evaluate the reliability of change scores, at least from a test-retest consistency perspective. This is because it is often difficult or impossible to design a reliability study in which true subject change can be repeatedly observed in the same subjects over time. Still, with an understanding of the points delineated above, and some consideration of the time interval over which true change is expected to occur, one can make use of test-retest correlations to assess the likelihood that a measure can be successfully used for its intended purpose. For example, over a time interval in which true changes of interest are expected to occur, a high correlation between time 1 and time 2 scores for a given measure would support the conclusion that the measure is reliable and stable, but the potential utility of change scores on that measure to serve as correlates or predictors of other individual difference measures will be low.

Conducting Reliability (Generalizability) Studies

The Choice of Participants: In designing a study to examine the psychometric characteristics of a potential biomarker measure (e.g., a generalizability study), there are a number of critical choices that one must make. The first is the question of who the participants should be. In assessing a measure's reliability, it is important that the participants be sampled from the population of interest so that the sample reflects the full range of variability that one expects to measure in the actual decision study. Thus, while it is sometimes tempting to assess the reliability of a measure in a sample of healthy control subjects, one cannot assume that the reliability estimated in such a sample will be applicable to a patient population. In particular, to the extent that the subject sample employed in a generalizability study under-represents the true variability, or over-represents the sources of measurement error, present in the population of interest, reliability of the measure will be underestimated. Likewise, if the subject sample employed in a generalizability study shows more true variability, or less susceptibility to measurement error, than the population targeted for study, reliability will be over-estimated. Thus, although it can be somewhat more effortful and time consuming, representative sampling from the population of interest in a generalizability study will provide more accurate estimates of the reliability that will actually affect the measurements in the substantive decision study.

Test-retest Time Frame: Another important choice is the time frame over which one measures reliability in a test-retest study. This time frame should be a long enough interval to capture the sources of random measurement error that are likely to contribute variation from occasion to occasion (e.g., scanner performance fluctuations, ambient temperature, staff member running the subject, random variation in subject vigilance). However, it should not be so long that true change in the measure is likely to have occurred (e.g., changes in brain function associated with exacerbation or remission of psychotic symptoms). As an example, think about designing a test-retest study for a measure of depression. Real and important changes in depression can occur over the course of a month. Thus, one would want to have a test-retest interval short enough to capture what one thinks are the stable parts of the construct being measured, but not long enough to be confounded by true changes that are likely to occur. In the context of imaging biomarker measures to assess cognition in schizophrenia, one would not want the test-retest interval to be so long that changes in brain function associated with exacerbation or remission of psychotic symptoms could have occurred. Thus, one may want to use a relatively short time frame in a test-retest study if you think true change could happen over a longer time frame (e.g., symptoms, etc.), even if the study you plan to conduct will occur over long time frame. If you know you have reliable measures in the short term, a researcher can then use a no-treatment or comparison control group to help interpret the source of change over a longer time span.

Reliability Estimates with Voxels Rather Than Subjects as the Objects of Measurement

All of the approaches to reliability assessment discussed above treat subjects as the objects of measurement and provide estimates of inter-subject variability relative to the total variability in the measurements. However, another possibility described in the fMRI reliability literature involves treating voxels, rather than subjects, as the objects of measurement and asks whether the relative ordering of voxel activations is preserved from one test occasion to another. This is tantamount to asking about the test-retest consistency of the pattern of activation across all of the brain voxels in an fMRI activation map. An advantage of this approach is that it permits estimation of the reliability of the activation map obtained from a single subject over two or more occasions. Once reliability coefficients are calculated for single subjects, it becomes possible to conduct statistical analyses using these coefficients as the unit of analysis. Thus, single subject reliability coefficients can be statistically compared between subject groups in order to determine if two or more groups significantly differ in their reliability. Individual differences in reliability can also be correlated with other variables representing subject characteristics (e.g., age, symptom severity) or sources of subject-specific noise (e.g., mean or maximum of a subject's movement parameters) in order to examine factors that correlate with, and possibly contribute to, inter-subject variation in test-retest activation map reliability. For some types of analytic methods, this approach may violate dependency assumptions, but it may still be an informative approximation. For example, Raemaekers reported high variability in the reliability of individual subject maps, with much of this variability explained by low signal to noise ratios (SNR) in the subjects with low reliability (13). Interestingly, these between subject differences in reliability did not appear to be related to differences in the degree to which activation in individual subjects matched the assumed hemodynamic response function (HRF), though one might have expected low SNR to also influence the goodness of fit for HRFs. Zandbelt *et al.* also found a good deal of variability in the within-subject consistency across sessions, even after subjects had been extensively practiced prior to scanning (14). These authors raised important points about the influence factors such as stress responses to scanning, caffeine, or alcohol use prior to scanning, smoking, amount of sleep, etc., might have on the variability in activation across subjects and across sessions.

Quality Assurance Considerations

Signal to Noise and Movement: Factors such as SNR, movement (which is often highly correlated with SNR), and other noise characteristics of the data (ghosting, etc.) can influence the quality of the data. There are two approaches one can take to dealing with factors that influence data quality. The first is to set some absolute threshold and only include data that meet this threshold. For example, a researcher could only include runs or subjects whose SNR is above some value and whose estimated movement is below some value. Such an approach is appealing, as it provides clear and consistent guidelines. However, the down side is knowing exactly where to set the thresholds, as many choices can appear arbitrary, and it is not always clear what thresholds best balance the needs of high quality data with data acquisition demands. An alternative approach is to *not* exclude subjects from analyses based on factors such as SNR or movement (at least above some very obvious cutoff), but to include estimates of these parameters as covariates in the statistical analyses. The advantage to this covariance approach is that it maximizes data inclusion, but at the risk of lower quality data biasing the results. Decisions about these issues are another domain in which a generalizability study can help. One can use the data from a generalizability study to help choose inclusion thresholds for quality assurance variables in a non-arbitrary fashion, by examining how factors such as SNR, movement, etc. influence estimates of reliability or activation magnitudes.

Equipment Stability: Another quality assurance consideration is the stability of the equipment that one is using to assess brain function. In the context of fMRI, this involves the stability of the scanner, the head coils, and the behavioral presentation/acquisition equipment. Fortunately, the functional Bioinformatics Research Network (fBIRN) has developed recommended tools and analysis approaches for assessing the stability of imaging equipment (15). These methods include the use of an agar phantom and a number of analyses that measure image quality characteristics. The agar phantom is designed to provide T1 and radiofrequency (RF) conductivity characteristics similar to brain tissue. The measured characteristics include signal-to-noise, signal-to-fluctuation noise, signal fluctuation and signal drift. The suggested acquisition parameters and quality assurance thresholds for 3T Siemens and GE scanners are provided at https://xwiki.nbirn.org:8443/xwiki/bin/view/Function-BIRN/What+is+the+agar+phantom+for+and+what+do+I+do+with+it. However, while these quality assurance thresholds are a useful starting point, practical experience in the Treatment

Units Research Network (TURNS) consortium has revealed that one must also have a plan for determining what level of variance around these values is typical for any given scanner. Thus, in an ideal study, each individual site would generate a reasonable collection of phantom quality assurance values collected on a weekly basis (e.g., 12-16 weeks) and compute means and standard deviations for that scanner in order to obtain an estimate of the expected variability across time. With these data, one could use scanner specific estimates (e.g., greater than 1 SD from the mean for that scanner) to track quality assurance and to detect problems with scanner performance. In addition, one would ideally also acquire agar phantom data with every subject scanned (in addition to weekly phantom scans) so as to be able to use the phantom data to identify any equipment issues that might be contributing to poor data quality in that participant. Analogous procedures are also available for assessing the quality of structural images (16, 17), including a “structural” phantom (18) and quality assurance procedures and guidelines (16, 17), much of which have been developed as part of the Alzheimer’s Disease Neuroimaging Initiative (19).

Relationship to Generalizability Theory: From the standpoint of reliability, between-subject variability in behavioral performance in fMRI studies is a source of “true score” variance (or “universe score” variance in G-theory terms) in the BOLD signal. This true score variance is captured by the Person variance in the ANOVA model used to estimate variance components for the ICC. High ICCs are achieved when the Person variance is large relative to the various error variance components. Thus, it is difficult to conceive of performance differences between subjects as a source of measurement error. In contrast, within-subject inconsistency in performance over time would be viewed as a source of measurement error. This temporal variation in performance would contribute to the Person x Occasion interaction term, which is an error variance component estimated in a test-retest reliability study. Accordingly, this interaction term implicitly contains variability due to unstable performance over measurement occasions. Beyond this, explicit inclusion of a performance variability measure, as suggested by the reviewer, would not be easily accommodated within the conceptual framework of G-theory.

Regarding equipment instability, the variance associated with such random instability is captured by the Person x Occasion interaction term estimated in a test-retest reliability study, similar to performance variability over time. However, systematic variation in scanner performance due to malfunction or slow drifts in hardware would not be evident in a typical test-retest reliability study, but could deleteriously affect scan quality over time. This is an issue for both longitudinal studies and cross-section studies in which subject recruitment and scan acquisition occur over a period of several years. To detect such drift, regular monitoring of scanner performance over time is necessary, typically by repeatedly scanning a phantom. The benefit of such monitoring is that scanner hardware problems can be detected quickly and repaired, minimizing the collection of bad data. The kind of measurement error introduced by hardware malfunction is not random; as such, it does not fit within a G-theory/reliability framework. Therefore, we believe it makes sense to separate the discussion of quality assurance considerations from the assessment of fMRI measurement reliability.

Potential Drug-Related Confounds

Include a “Control” Task: The first suggestion by Iannetti and Wise is to include a “control” task that is not expected to be influenced by the pharmacological manipulation. Should the study reveal changes in BOLD activity associated with the target cognitive task and not the control task, then one may be in a better position to argue for selective influences on neural activity. However, there are two potential problems with this approach. First, as noted by Iannetti and Wise, there may be regional specificity to some pharmacological influences on components of the BOLD response other than neural activity (20). Thus, one would not want a control task that only activated very different regions than the target task (e.g., only visual cortex versus only frontal cortex). Instead, it would be good to have a control task that activated regions that were separable from those activated in the target task, but in more similar regions (e.g., different subregions of frontal cortex, parietal cortex, etc.). A second potential problem with the control task strategy is that one may be interested in a pharmacological agent that is thought to have very wide spread effects on many aspects of cognition, and not just on very focused components. For example, if you had a drug that was designed to augment glutamatergic transmission broadly, it may be hard to find a control task that one could strongly hypothesize to *not* be influenced by changes in glutamate function.

Measuring Changes in Cerebral Blood Flow: It is certainly feasible that many pharmacological agents could alter cerebral blood flow (CBF) and that this effect could lead to a change in BOLD reactivity (21). Thus, Iannetti and Wise recommend including measures of CBF in pharmacological imaging protocols in order to directly address this potential confound (22). One method gaining increasing use in MR studies is arterial spin labeling (23, 24). Of note, simply finding CBF changes in response to a pharmacological intervention does not necessarily mean that changes in BOLD activity during cognitive challenge are solely due to drug influences on CBF. However, including CBF values as a covariate would help one to understand whether there is an influence on BOLD response over and above the influence on CBF.

Measuring Changes in Vascular Reactivity: Vascular reactivity reflects the ability of blood vessels to either dilate or contract in response to changes in physiological parameters known to modulate the brain's perfusion (20). Changes in vascular reactivity that influence BOLD responses to cognitive events, even in the absence of changes in the underlying neural activity, are one of the major potential confounds in pharmacological fMRI studies. Another way to frame this issue is to point out that the BOLD signal is the result of changes in CBF, cerebral blood volume (CBV) and cerebral metabolic rate of oxygen consumption ($CMRO_2$) (25). Thus, changes in CBF due to changes in vascular reactivity could change BOLD responses even when $CMRO_2$ does not change. A number of researchers have used a method called "calibrated" BOLD to assess CBF contributions to BOLD separately from $CMRO_2$ (25-28). In calibrated BOLD, hypercapnia (increased levels of carbon dioxide in the blood) is induced by having participants breathe air with increased concentrations of CO_2. This leads to increased vasodilatation and changes in BOLD response in the absence of changes in $CMRO_{2.}$ A number of studies have also used voluntary breathholding on the part of participants in order to induce hypercapnia, which is methodologically somewhat less complicated (25-29). However, the successful use of breathholding to induce hypercapnia is dependent on participant's compliance and accuracy in following the task instructions, a component of the process that could be more challenging in patient studies. As with measures of CBF, determining that a drug influences vascular reactivity as revealed by increased BOLD responses to hypercapnia does not preclude the possibility that the drug also has a direct influence on neural activity. However, similar to CBF measures, it would then be important to include BOLD activity related to hypercapnia as a covariate in statistical analyses of the cognitive-task related BOLD data.

Arousal, Cardiac Pulsation, and Respiration: It has long been known that cardiac pulsation and respiration effects can influence BOLD responses to cognitive challenges. Further, at least some pharmacological agents can influence factors such as arousal, which in turn could influence cardiac pulsation and/or respiration. Thus, one may need to measure and control for such changes in analyses of functional brain activity related to cognition. This can be accomplished using methods such as the RETROICOR algorithm that uses heart and respiration rate measurements in regression-based analyses to eliminate their influences on the BOLD signal related to cognitive performance (30, 31). However, it is also important to be aware of the fact that such changes in heart rate and respiration could also be correlated with the changes of interest in neural activity. Thus, removing such variance from the BOLD signal could also reduce sensitivity to drug-related changes in BOLD activity associated with cognition.

References

1. Cronbach LJ, Furby L (1970): How should we measure "change" - or should we? *Psychol Bull*. 74:68-80.
2. Linn RL, Slinde JA (1977): The determination of the significance of change between pre- and postesting periods. *Rev Educ Res*. 47:121-150.
3. Lord FM (1956): The measurement of growth. *Educ Psychol Meas*. 16:421-437.
4. Lord FM (1963): Elementary models for measuring change. In: Harris CW, editor. *Problems in Measuring Change*. Madison, WI: University of Wisconson Press, pp 21-38.
5. Rogosa DR, Brandt D, Zimowski M (1982): A growth curve approach to the measurement of change. *Psychol Bull*. 92:726-748.
6. Rogosa DR, Willett JB (1983): Understanding correlates of change by modeling individual differences in growth. *Psychometriika*. 50:203-228.
7. Zimmerman DW, Williams RH (1982): The relative error magnitude in three measures of change. *Psychometriika*. 47:141-147.
8. May K, Hittner JB (2003): On the relation between power and reliability of statistical tests: some new results. *Psychol Bull*. 94:524-533.
9. Nicewander WA, Price JM (1978): Dependent variable reliability and the power of statistical tests. *Psychol Bull*. 85:405-409.
10. Overall JE, Woodward JA (1975): Unreliability of differences scores: A paradox for measurement of change. *Psychol Bull*. 82:85-86.
11. Sutcliffe JP (1980): On the relationship of reliability to statistical power. *Psychol Bull*. 88:509-515.
12. Zimmerman DW, Williams RH (1986): Note on the reliability of experimental measures and the power of significance tests. *Psychol Bull*. 100:123-124.
13. Raemaekers M, Vink M, zandbeldt BB, van Wezel RJ, Kahn R, Ramsey NF (2007): Test-retest reliablity of fMRI activation during prosaccades and antisaccades. *Neuroimage*. 36:532-542.
14. Zandbelt BB, Gladwin TE, Raemaekers M, van Buuren M, Neggers SF, Kahn RS, *et al.* (2008): Within-subject variation in BOLD-fMRI signal changes across repeated measurements: quantification and implications for sample size. *Neuroimage*. 42:196-206.
15. Friedman L, Glover GH (2006): Report on a multicenter fMRI quality assurance protocol. *J Magn Reson Imaging*. 23:827-839.
16. Mortamet B, Bernstein MA, Jack CR, Jr., Gunter JL, Ward C, Britson PJ, *et al.* (2009): Automatic quality assessment in structural brain magnetic resonance imaging. *Magn Reson Med*. 62:365-372.
17. Kruggel F, Turner J, Muftuler LT (2010): Impact of scanner hardware and imaging protocol on image quality and compartment volume precision in the ADNI cohort. *Neuroimage*. 49:2123-2133.

18. Gunter JL, Bernstein MA, Borowski BJ, Ward CP, Britson PJ, Felmlee JP, *et al.* (2009): Measurement of MRI scanner performance with the ADNI phantom. *Med Phys*. 36:2193-2205.

19. Jack CR, Jr., Bernstein MA, Fox NC, Thompson P, Alexander G, Harvey D, *et al.* (2008): The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. *J Magn Reson Imaging*. 27:685-691.

20. Iannetti GD, Wise RG (2007): BOLD functional MRI in disease and pharmacological studies: room for improvement? *Magn Reson Imaging.* 25:978-988.

21. Cohen ER, Ugurbil K, Kim SG (2002): Effect of basal conditions on the magnitude and dynamics of the blood oxygenation level-dependent fMRI response. *J Cereb Blood Flow Metab*. 22:1042-1053.

22. MacIntosh BJ, Pattinson KT, Gallichan D, Ahmad I, Miller KL, Feinberg DA, *et al.* (2008): Measuring the effects of remifentanil on cerebral blood flow and arterial arrival time using 3D GRASE MRI with pulsed arterial spin labelling. *J Cereb Blood Flow Metab*. 28:1514-1522.

23. Paiva FF, Tannus A, Silva AC (2007): Measurement of cerebral perfusion territories using arterial spin labelling. *NMR Biomed*. 20:633-642.

24. Petersen ET, Zimine I, Ho YC, Golay X (2006): Non-invasive measurement of perfusion: a critical review of arterial spin labelling techniques. *Br J Radiol*. 79:688-701.

25. Perthen JE, Lansing AE, Liau J, Liu TT, Buxton RB (2008): Caffeine-induced uncoupling of cerebral blood flow and oxygen metabolism: a calibrated BOLD fMRI study. *Neuroimage*. 40:237-247.

26. Chiarelli PA, Bulte DP, Gallichan D, Piechnik SK, Wise R, Jezzard P (2007): Flow-metabolism coupling in human visual, motor, and supplementary motor areas assessed by magnetic resonance imaging. *Magn Reson Med*. 57:538-547.

27. Leontiev O, Dubowitz DJ, Buxton RB (2007): CBF/CMRO2 coupling measured with calibrated BOLD fMRI: sources of bias. *Neuroimage*. 36:1110-1122.

28. Thomason ME, Foland LC, Glover GH (2007): Calibration of BOLD fMRI using breath holding reduces group variance during a cognitive task. *Hum Brain Mapp*. 28:59-68.

29. Thomason ME, Glover GH (2008): Controlled inspiration depth reduces variance in breath-holding-induced BOLD signal. *Neuroimage*. 39:206-214.

30. Chang C, Glover GH (2009): Relationship between respiration, end-tidal CO2, and BOLD signals in resting-state fMRI. *Neuroimage*. 47:1381-1393.

31. Chang C, Cunningham JP, Glover GH (2009): Influence of heart rate on the BOLD signal: the cardiac response function. *Neuroimage*. 44:857-869.