Cogn Affect Behav Neurosci (2013) 13:615–626
DOI 10.3758/s13415-013-0178-2

Number of events and reliability in fMRI

Benjamin O. Turner · Michael B. Miller

Published online: 11 June 2013

Abstract Relatively early in the history of fMRI, research focused on issues of power and reliability, with an important line concerning the establishment of optimal procedures for experimental design in order to maximize the various statistical properties of such designs. However, in recent years, tasks wherein events are defined only *a posteriori*, on the basis of behavior, have become increasingly common. Although these designs enable a much wider array of questions to be answered, they are not amenable to the tight control afforded by designs with events defined entirely *a priori*, and little work has assessed issues of power and reliability in such designs. We demonstrate how differences in numbers of events—as can occur with *a posteriori* event definition—affect reliability, both through simulation and in real data.

Keywords Neuroimaging · Design and analysis · Statistical power

Recently, a great deal of attention has been paid to statistical issues in fMRI. Although several reports have focused on the validity of various results (e.g., Power, Barnes, Snyder, Schlaggar, & Petersen, 2011; Van Dijk, Sabuncu, & Buckner, 2012), the majority have expressed concerns about the reliability of various methods or results (e.g., Carp, 2012; Eklund, Andersson, Josephson, Johannesson, & Knutsson, 2012; Vul, Harris, Winkielman, & Pashler, 2009). The reliability of fMRI itself has a long history of study (e.g., McGonigle, 2012; McGonigle et al., 2000; Miller et al., 2002). However, in order to interpret this work, we need to define what we mean by reliability. Traditionally, the term "reliability" has been used in reference to a particular scale or measure designed to assess some trait or phenomenon.[1] In contrast, most previous research on fMRI reliability has focused on the replicability or trustworthiness of some particular result. However, our estimate of reliability depends on factors including the reliability of the underlying cognitive processes, the degrees of within- and between-subjects variation in brain activity, the metric that we use to measure reliability, and myriad other factors distinct from the reliability of the result *per se* (Gorgolewski, Storkey, Bastin, Whittle, & Pernet, 2013).

In the present work, we define *reliability* as the measured similarity of a pattern of results obtained under (ideally) identical conditions;[2] although there may be better names for this property, we use "reliability" because of its historical use in the field. Although we focus on reliability at the level of a single run, this will of course impact reliability at higher levels of analysis as well. Conceptually, it is important to disambiguate our ability to estimate the replicability of a result from the true replicability of that result. The former is affected by concerns such as statistical power or the proportion of activity, whereas the latter is independent of such factors, a point to which we will return in the Discussion.

For the majority of researchers, reliability is an abstract ideal—that is, a desirable property, but not one that directly impacts the interpretation of any given (nonreplication) study. For such a researcher, the problem is that there is no simple correspondence between, for example, p values and replicability (*cf.* the debate over p_{rep}; Killeen, 2005a, 2005b). Therefore, our goal for these researchers is to provide guidance as to how representative their results are of the true (unobservable) underlying result, under the assumption that results that more faithfully capture the truth ought to be more replicable.

B. O. Turner · M. B. Miller (✉)
Department of Psychological & Brain Sciences, University of California, Santa Barbara, Santa Barbara, CA 93106, USA
e-mail: m_miller@psych.ucsb.edu

[1] Of course, the reliability of fMRI has an upper bound as a measurement tool, but this is a property of the physics underlying fMRI.

[2] Interpreted at different levels, this definition encompasses several popular methods of assessing reliability, including the intraclass correlation coefficient (which assesses the stability of a pattern of scores or statistics across participants), overlap metrics (the degree to which the same decision regarding the null hypothesis was made on a voxelwise basis), and correlation (the degree of similarity of the patterns of relative activity across voxels).

A smaller subset of researchers are interested in reliability in a much more concrete way—namely, those researchers who study inter- (or intra-) individual variability (often termed "individual differences analysis"). Although these researchers will typically quantify reliability in a different way than we do here—for instance, by using an intraclass correlation (Caceres, Hall, Zelaya, Williams, & Mehta, 2009)—the stability of a result within an individual is an important consideration as well. That is, for these researchers, any result must be interpreted through the lens of the reliability of their statistical tests. Our goal relative to these researchers is to provide direct evidence of design factors that might impact their analyses—and that must be therefore accounted for as nuisance factors in these analyses—as well as to point out that measures of reliability are themselves subject to a degree of estimation error.

Although much of the recent attention paid to problems with fMRI reliability has taken the form of a focus on high false alarm rates, this is only part of the equation in determining reliability. A number of researchers have previously made this point: that the reliability of a result depends on statistical power, as well as on the intrinsic reliability of the phenomenon being studied, the degree of individual difference, and so forth (Bennett & Miller, 2010; Gorgolewski et al., 2013). Our intention is not to dismiss or diminish the seriousness of inflated false alarm rates; on the contrary, this line of research is essential to defining good practices for fMRI design and analysis. Instead, we wish to point out other, more pedestrian reasons why any particular result may fail to replicate, or more precisely, why that result may yield low reliability metric estimates—including a lack of appropriate statistical power (see also Button et al., 2013).

Although the power of a test is partially determined by factors beyond the experimenter's control, over one factor the experimenter does have some direct, *a priori* control: namely, sample size. Previous work has investigated this issue at the level of the number of scans per functional run and the number of participants per experiment (Desmond & Glover, 2002; Mumford & Nichols, 2008), as well as in terms of what number and spacing of events[3] will yield optimal power—among several other criteria that researchers may want to optimize—for different sorts of contrasts or analyses (Dale, 1999; Friston, Zarahn, Josephs, Henson, & Dale, 1999; Liu & Frank, 2004; Liu, Frank, Wong, & Buxton, 2001; Wager & Nichols, 2003).

However, it is common in cognitive neuroscientific experiments for events to be defined *a posteriori* on the basis of participant behavior, which limits the direct applicability of much of this work. For instance, consider a standard recognition memory experiment. During the retrieval phase, the experimenter can dictate on which repetition times (TRs) studied or new items occur, but events are frequently defined in the signal detection theoretic framework (Green & Swets, 1974; Macmillan & Creelman, 2005) as hits, misses, false alarms, or correct rejections—labels that depend on the participants' responses. In these situations, the number of events of a particular type is stochastic, and imbalances in the numbers of events (and therefore in event-specific contrast power) are almost inevitable.

The relationship between numbers of instances and power is not one to one. It is not the case, for instance, that the number of events factors in any direct way in the computation of degrees of freedom for statistical tests, and when the popular convolution-based general linear model (GLM) is used in an analysis (Worsley & Friston, 1995), the number of events has a nonlinear relationship even with the standard errors of the estimators. Although it is possible in some situations to analytically derive solutions for how event imbalances will affect reliability, our goal is to use simulation and simple analyses in real data to demonstrate the relationship between numbers of events and reliability across a range of situations likely to be encountered by researchers.

Method

In order to demonstrate the effects of varying numbers of events on reliability, we took two approaches. The first was a simulation analysis, which we used in place of analytic derivations in order to make the results accessible to a wider audience. The second was an analysis on a rich real data set, comprising multiple repeated measurements within individuals across a year. Because this experiment used *a posteriori* event definition, a range of numbers of events occurred for the different types, which allowed for an investigation of the impact of event number on reliability in real data. In the simulations, we focused on a direct contrast between two events to make the point that the lesser number of events primarily drives reliability. For the real data, we focused instead on event-unique contrasts (i.e., vs. baseline) for conceptual simplicity, in order to address other issues that come with using real data (e.g., how differences in the numbers of events between runs impact reliability). As an approximation, the average number of events across classes from the simulation analyses served in a role similar to the average number of events across sessions from the real-data analyses.

Simulation analysis

Our primary goal in doing the simulations was to establish estimates for reliability as a function of a number of empirically observable factors, including the extent of activation (in

[3] We will use the terminology "class" or "event type" when referring to a broad label—for instance, "hits"—and "events" or "instances" when referring to instantiations of a particular event type—for instance, a "hit" on some particular TR.

terms of proportions of active voxels), the signal-to-noise ratio (SNR), and—our primary focus, due to its being under experimental control to some degree—the number of instances of events of different classes. We additionally examined the impact of the underlying stability of the pattern of activity, which is a factor that is not directly observable empirically, but that certainly contributes to our estimates of reliability; in fact, in many instances, this is what reliability analyses are attempting to measure. We carried out two sets of simulations: The first (corresponding to the scenario described below) was meant to establish the relationship between number of events and reliability under ideal conditions, whereas the second was designed to demonstrate the impact of differences in underlying stability in more realistic circumstances. Both of these are described in more detail below.

Our simulations are framed by the following hypothetical scenario: A researcher has run an experiment comprising approximately 300 TRs, and being only interested in the direct contrast, has included 150 instances of each of two classes of events, per the recommendations of Friston et al. (1999), although without any constraints on the events' ordering, maximum stimulus onset asynchrony, or any other parameter (except that only one event can occur per TR). However, in this experiment, a participant's behavior partially determines how each event will be treated during analysis, in a manner akin to that described above for many recognition memory experiments. For both the ideal and realistic simulations, we made the simplifying assumption that under replication, the exact same number of events would be observed for each type. Deviations from this would generally serve to reduce the observed similarity (see the Results for the real-data analyses).

Generating data

To simulate fMRI data, we used a multivariate extension of the method described in Mumford, Turner, Ashby, and Poldrack (2012). Briefly, given N_1 events of Class 1 and N_2 events of Class 2, N_1 TRs were assigned to Class 1, N_2 were assigned to Class 2, and the remainder ($320 - N_1 - N_2$) were blank. In each voxel, a simulated time course was generated by constructing a boxcar whose heights for Class 1 TRs were drawn from the Class 1 distribution (see below), and likewise, Class 2 TRs had heights drawn from the Class 2 distribution. After convolving this boxcar with a double-gamma hemodynamic response function (HRF; generated using the FMRIB Software Library [FSL]), noise distributed as $N(0, \sigma^2_{wn})$ was added, generating the final observed time course for that voxel. This process was repeated independently for each voxel to yield a full volume of 10,000 voxels. No spatial smoothing was applied, because the range of SNRs chosen for the simulation (which, as we discuss below, dictated our choice of σ^2_{wn}) was based on a range observed empirically after spatial smoothing; applying smoothing to these data would therefore change the SNR from its intended value.

In order to capture randomness at multiple levels, we used a two-stage procedure for generating distributions of boxcar heights: The first stage was used to sample means for the distributions used at the second stage. To allow for varying degrees of similarity between two simulated runs, we used multivariate (specifically, bivariate) distributions, in which the two dimensions of the variable represented the two runs and the degree of covariance between the dimensions dictated their similarity. For simplicity in subscripting, we will use scalar (rather than vector) notation in the following description, except when denoting covariance matrices; note that the procedures are identical and symmetric with respect to the two runs.

For both sets of simulations, P voxels were designated as "active" and Q (that is, $10{,}000 - P$) as "inactive." For both sets of voxels, the mean of the Class 1 distribution was distributed as $N(0, \Sigma_{idl})$, where the covariance matrix Σ_{idl} encodes the within-run (across-voxel) variance of 4 and the across-run (within-voxel) covariance of 3.2. However, the Class 2 distribution depended on set membership: Letting A_m denote the sampled value for Class 1 in voxel m, for the ideal simulations, active voxels had mean Class 2 activity equal to $A_m + 3$, and inactive voxels to A_m. Because we focused on the contrast between these events, this translated to contrast values of 3 in active voxels and 0 in inactive ones. For the realistic simulations, the mean Class 2 activity within a voxel was distributed probabilistically as $A_m + N(3, \Sigma_{ac})$ in active voxels, where Σ_{ac} encodes the within-run variance of 1 and the across-run covariance of σ^2_{sim}, and as $A_m + N(0, \Sigma_{inac})$ in inactive voxels, where Σ_{inac} encodes the within- and across-run variance and covariance of 1 and 0, respectively. Finally, the actual heights of each individual boxcar were distributed as $N(\mu_{m,c}, 0.5^2)$, where $\mu_{m,c}$ denotes the value drawn from the relevant distribution of means for voxel m and class c.[4]

Parametric analysis

As described above, our primary focus was on the impact of number of events, being the parameter among those whose impact we investigated that is most directly under experimental control (even if only imprecisely, due to its dependence on participant behavior). Therefore, each simulation included variation across numbers of events, realized by holding all other elements constant (e.g., specific instantiations of noise, event onset vectors, and boxcar heights) and repeating

[4] The second simulation procedure embodied the Gaussian mixture model assumption underlying fMRI analysis techniques such as ICA; the first simulation was an extension of that logic to a noiseless case.

analyses with numbers of events per class ranging from 10 to 150 (ideal) or 100 (realistic), in steps of size 10.

In addition, we sought to establish the impacts of several other variables that affect certain measures of reliability—namely, SNR and the proportion of active voxels. SNR was varied by choosing different values of the white noise variance σ^2_{wn} from the set $\{0.5^2, 1^2, 2^2, 4^2, 8^2\}$; the proportion of active voxels P ranged from 0 to .5, spanning either the entire interval in steps of .05 (ideal) or only a subset thereof ({.05, .10, .20, .40}, realistic). Both ranges were chosen to cover the range of values likely to be encountered by researchers in real data. For each simulation, all other variables were held constant across changes in the levels of these two parameters, allowing for an inspection of the impact of each uniquely.

Lastly, for the ideal simulations, the similarity between the underlying sample of effect size means was always identical, but for the realistic simulations, we parametrically varied the similarity, testing values of σ^2_{sim} from the set $\{0.3^2, 0.6^2, 0.9^2\}$. The first set of simulations would give boundary conditions: These are theoretical upper limits on reliability (i.e., measures of reliability—the true reliability is perfect) across a range of parameters. The second set would demonstrate the relationship between true reliability and estimates of that reliability, and how this relationship depends on other variables.[5]

We fully crossed all variables within each simulation type, and conducted 50 (ideal) or 100 (realistic) simulations. To measure similarity, we used the Spearman correlation, as well as the Jaccard overlap, between the two statistical parametric mappings (SPMs) resulting from analyzing the simulated data using the standard GLM. For the former, we calculated Spearman's rho on the unthresholded *t*-statistic maps from a contrast of the two types of events for the two runs (using the MATLAB Statistics Toolbox). For the latter, we computed the overlap on the thresholded versions of the same SPMs, thresholded at $p = .05$, uncorrected for multiple comparisons. Under the null hypothesis of no true similarity, Spearman's rho should be zero, and the Jaccard overlap with an uncorrected threshold of $p = .05$ should be $.05^2/(1 - .05^2)$.= .026.

Real-data analysis

It is always important to establish that the results observed in simulations hold for real data. The real data that we used for this purpose here came from a study of recognition memory (the first part of which was published in Miller, Donovan, Bennett, Aminoff, & Mayer, 2012). In this study, 12 participants were tested five times each over the course of a year, undergoing two encoding and two retrieval scans during each test session—one pair in each of two imageability conditions—along with structural and resting-state scans. This data set allowed for the assessment of a variety of factors contributing to reliability (Bennett & Miller, 2013), in addition to an investigation of empirical questions about recognition memory (Turner, Donovan, & Miller, 2013).

Here, this data set will be used only to establish the relationship between number of events and reliability. To this end, we first analyzed each functional run with a standard GLM analysis using FSL version 4.1. Standard preprocessing, including brain extraction, motion correction, spatial smoothing with a 5-mm full-width-at-half-maximum Gaussian kernel, high-pass filtering with $\sigma = 50$ s, and grand-mean scaling, was run on each functional scan. The data were then analyzed using the standard GLM approach, with regressors for subsequently remembered or forgotten words (for encoding scans) or hits, misses, false alarms, and correct rejections (for retrieval scans), constructed by convolving a boxcar (with a 1 for any TR with the corresponding label, and 0 otherwise) with a model HRF (gamma; phase = 0 s, sigma = 3 s, peak = 6 s), in addition to the temporal derivatives of each regressor. Additional nuisance regressors were included for motion parameters (translation in x, y, z, along with rotations about each axis) and their temporal derivatives.

Per FSL's defaults, the data were prewhitened prior to analysis, and the design matrix was temporally filtered in the same way as the data. The contrasts of interest were simple unique-event contrasts for each event type, yielding a total of 12 SPMs per participant per session—two encoding and four retrieval SPMs in each of the two conditions. These SPMs were aligned to the high-resolution anatomical image from the first scanning session for each participant; there was no need to align to standard space or resample to a resolution other than the acquisition resolution, because all correlations were performed for the whole brain (without reference to particular regions of interest) and within subjects.

Next, we computed the Spearman correlation[6] between each pair of SPMs across sessions but within participants, conditions, and contrasts (see Miller et al., 2012, for related analyses of inter-, rather than intra-, individual variability in the first session of this experiment). With ten pairwise combinations

[5] Note that due to our simulation method, σ_{sim} differed from the correlation across the entire volume, and unlike those whole-volume correlations, was independent of the proportion active: The values of σ^2_{sim} only dictated the reliability within active voxels (the covariance was always zero for inactive voxels), but the active and inactive sets stayed constant. Therefore, across levels of proportions active (ignoring the trivial value of 0 for a proportion active of 0), the correlations between the true effect sizes across the entire set of voxels ranged from .29 to .54 for a nominal σ^2_{sim} of $.3^2$, from .32 to .75 for $\sigma^2_{sim} = .6^2$, and from .33 to .83 for $\sigma^2_{sim} = .9^2$.

[6] We used correlations rather than overlap because the former, being "soft" statistics, are less susceptible to minor errors in alignment. Specifically, spatial smoothing tends to cause statistics in neighboring voxels to take on similar values, which means that misalignment would cause only a small drop in correlations, but where these values fell relative to a statistical threshold would not be systematically affected by smoothing; thus, misalignment could cause changes in overlap that would be more sharp than the smooth changes likely to be seen with correlations.

across the five sessions (per subject, contrast, and condition), this resulted in a total of 1,440 correlation values. For comparison with the simulation results, we further defined each pair of scans on the basis of characteristics including SNR, proportion of activity, and numbers of events. For the last of these, we took as variables the minimum number of events across the session pair (denoted by n_{min}), along with the absolute difference in numbers of events (n_{dif}). Note that because our simulations always included the same number of events across replications, neither of these mapped directly onto the number-of-events variable from the simulation. However, n_{min} tends to be the limiting factor in determining replicability, and as we speculated above, n_{dif} might reasonably be expected to impact similarity (e.g., in active voxels, as the number of events increases, the mean and variance of the noncentral *t* from which the observed *t* statistic comes will increase as well).

For the other two variables, we had to account for the fact that unlike in the simulation, in which we could manipulate all of the variables independently, our estimation procedures for SNR and proportion active might depend on the number of events. To correct for this, we did the following: For each unique contrast, we calculated the alpha level that would result in power equal to the mean power across all contrasts, assuming equal effect sizes and noise variances for all contrasts.[7] This resulted in alpha values (after using FSL's voxelwise correction for multiple comparisons) between .0001 and .1360, which we used to set the whole-brain threshold. Next, individually for each contrast, we applied that contrast's threshold and calculated the SNR as the mean parameter value across suprathreshold voxels, divided by the mean standard deviation of the residuals for the same voxels (Mumford et al., 2012). Likewise, the proportion of active voxels was calculated as the number of suprathreshold voxels divided by the total number of voxels within the brain mask (as created by FSL). In order to relate these to the correlation for a given SPM pair, we took the average of the SNR values for each pair as the SNR value for that pair, and likewise for the proportion active. Note that these are both in the same units as in the simulation, allowing a direct comparison between the two analyses (conditioned on the success of our correction for the nonindependence between the variables).

Results

Simulation results

The results from the ideal simulation are presented in Fig. 1, which shows the effects of number of events across different levels of SNR and proportions of active voxels on reliability, as measured by Spearman's rho, in the idealized case in which the underlying amounts of activity are identical across simulated runs. Figure 2 shows the same results using Jaccard overlap as the reliability metric. Using our slightly more realistic simulation methodology yielded the results shown in Fig. 3, for Spearman's rho, and Fig. 4, for Jaccard overlap, with the error bands representing differing levels of underlying reliability. All of the figures demonstrate similar trends—namely, that the effect of number of events depends on the proportion of activity as well as on the SNR, such that the largest effects are seen for relatively higher proportions of activity and relatively lower levels of SNR. With lower proportions, the differences in numbers of events exert an influence over less of the volume, and therefore have a muted effect (although differentially for Spearman's rho and Jaccard overlap, as might be expected); as for SNR, the number of events has almost no effect at high levels, because the signal is saturated, whereas at low levels, adding increasingly more small events overcomes the noise fairly slowly.

These results are summarized in Table 1 in the form of average values of N_1 and N_2, such that reliability attained 90 % of its eventual maximum, given particular levels of SNR and proportion active, separately for each of our two simulations. This is useful because, as is obvious from Figs. 1 and 2, including additional trials produces diminishing returns, especially at higher SNRs. Table 1 also highlights the fact that as SNR improves, the dramatic flattening of the reliability curves as a function of N_1 and N_2 observable in Figs. 1 and 2 yields earlier and earlier points at which these curves reach 90 % of maximum.

Note, too, that these are mathematically nearly worst-case scenarios within our simulation framework: That is, because the path that we traversed in (N_1, N_2) space maximized imbalance between the two events (±20 because of our discrete step size), any value $(N_1 + N_2)/2 = M$ in Table 1 reflects the reliability attained with N_1 near the minimum possible, given the constraints of M and N_{total} events. In other words, if Table 1 indicates a $N_{90\%ile}$ of 80, this was calculated with $N_1 = 10$ and $N_2 = 70$; the reliability would be higher for $N_1 = N_2 = 40$, because as is shown in Figs. 1 and 2, the lesser number of events (in a contrast involving two events) is the primary limiting factor driving reliability, a point to which we shall return later.

Real-data results

Our attempts to unconfound SNR, proportion active, and number of events were at least partially successful. The mean proportion-of-similarity values from each stratum of SNR belonging to the corresponding stratum of proportions, or vice versa, ranged from .31 to .52—in other words, the variables covaried somewhat (the proportions should be

[7] The power of a *t* test depends on the effect size, the noise variance, the design matrix, and alpha; by assuming equal values for the first two and using the actual design matrices applied in the analyses for the third, it was possible to solve for the fourth for a fixed value of $1 - \beta$.

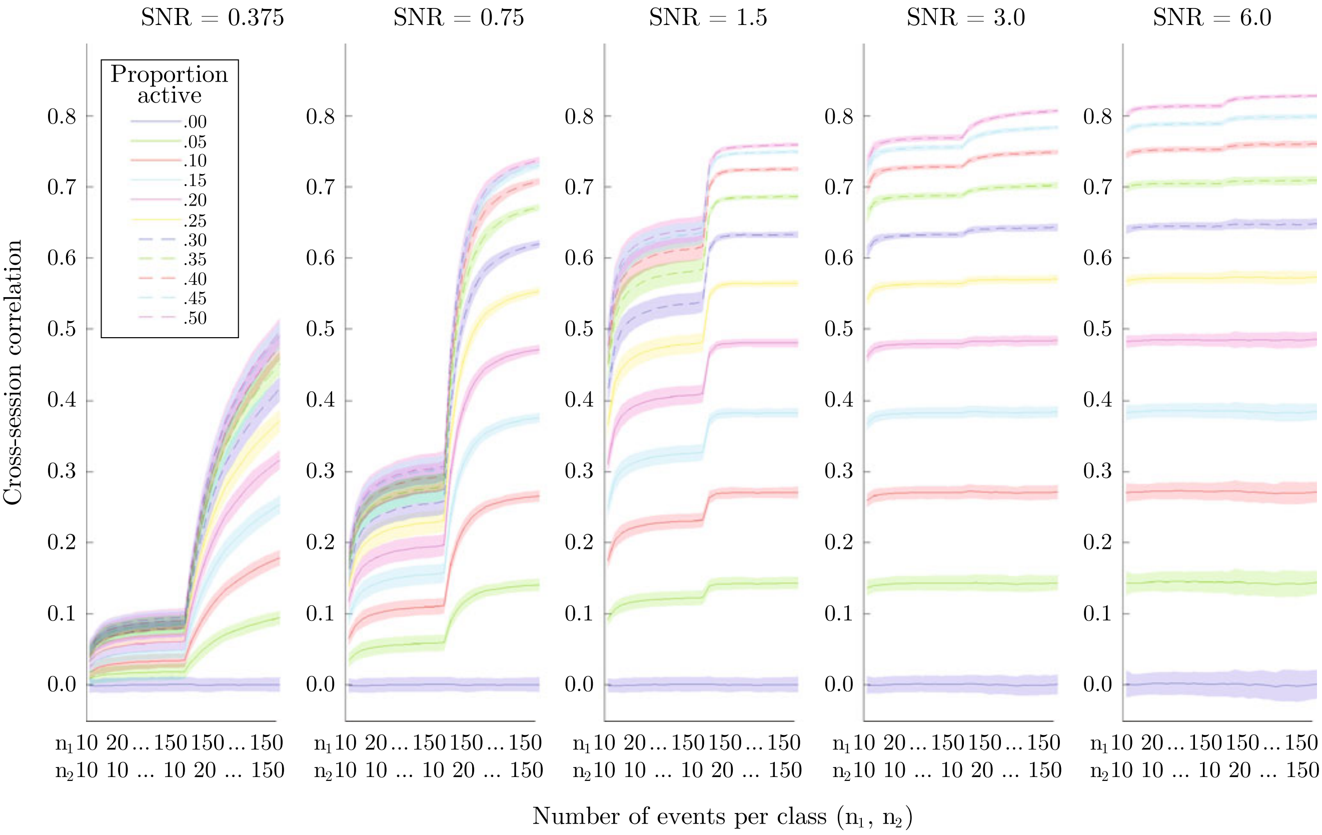

Fig. 1 Effects of signal-to-noise ratio (SNR), proportion active, and number of events on reliability for each of two simulated classes, as measured by Spearman's rho, for idealized simulations. Lines show the means, with error bounds of ±1 *SD*, from across the simulations

around .25 by chance, assuming equally sized strata), but not significantly so (all *ps* > .07 for binomial tests on shared proportions). The relationships remaining between n_{min} and the other variables were approximately equally strong. The Spearman correlation values were –.42 between n_{min} and SNR, –.34 between n_{min} and proportion active, and –.17 between n_{min} and n_{dif}.[8] Moreover, the fact that a range of values of n_{min} was represented at each stratum of these variables allowed for a relatively unbiased investigation of how the variables interact (i.e., SNR is definitionally independent of n_{min} or n_{dif} across the panes of Fig. 6, for example).

The similarities that we observed in our real data, stratified separately by SNR and proportions of active voxels, are shown in Figs. 5, 6, 7 and 8. (We decided against simultaneously stratifying by both variables, due to sample size: Although our simulations afforded us tens or hundreds of thousands of values and were balanced evenly across every combination of variables, we only had 1,440 correlations for the real data.) Although the panes showing the highest values for SNR and proportion active with n_{min} skew heavily toward pairs with low values (as is shown in the densities in the lower panels, for the reasons described above), and so should be treated cautiously, the other three panes in Figs. 5 and 6 follow the same pattern as for the simulations: no clear effect of number of events at the lowest levels of SNR or proportion active, and an increasingly stronger relationship between the two as SNR or proportion active increases. Note that the ranges tested in the simulations span a superset of the ranges seen in our real data—0.375–6.0 for SNR, and 0–.5 for proportions active in simulation, as compared with (using the 5th and 95th percentiles as robust estimates of range) 1.56–2.31 for SNR and .01–.96 for proportion active.

No result from the simulations, in which the numbers of events were always the same across the two sessions, corresponds to that shown in Figs. 7 and 8. However, the result was as anticipated: Issues of reliability aside, when the numbers of events in two scans are widely divergent, the resulting SPMs will tend to be dissimilar, particularly in this case of an upper

[8] The last relationship can be understood algebraically: As n_{min} increases, because there is an upper limit on the total number of events within a session, the range of possible values for n_{dif} shrinks. For the first two, the relatively high negative correlations are driven almost entirely by inflated SNR and proportion active values for session pairs with small numbers of events; after removing all pairs for which n_{min} fell below 20 (336 of the total number of pairs, or roughly 23 % of all pairs), these correlations dropped dramatically, to –.07 and –.08 for SNR and proportion active, respectively.

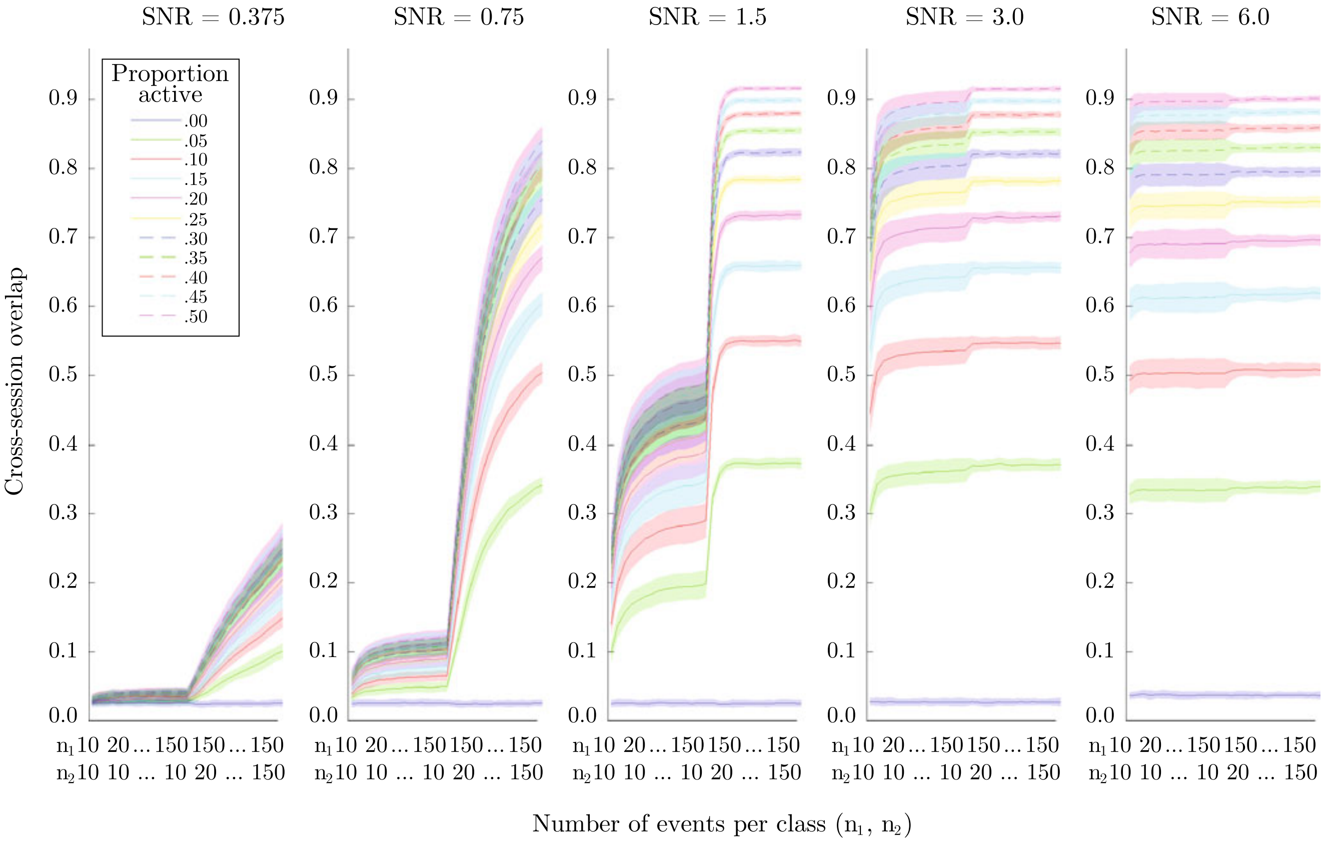

Fig. 2 Effects of signal-to-noise ratio (SNR), proportion active, and number of events on reliability for each of two simulated classes, as measured by Jaccard overlap, for idealized simulations. Lines show the means, with error bounds of ±1 *SD*, from across the simulations

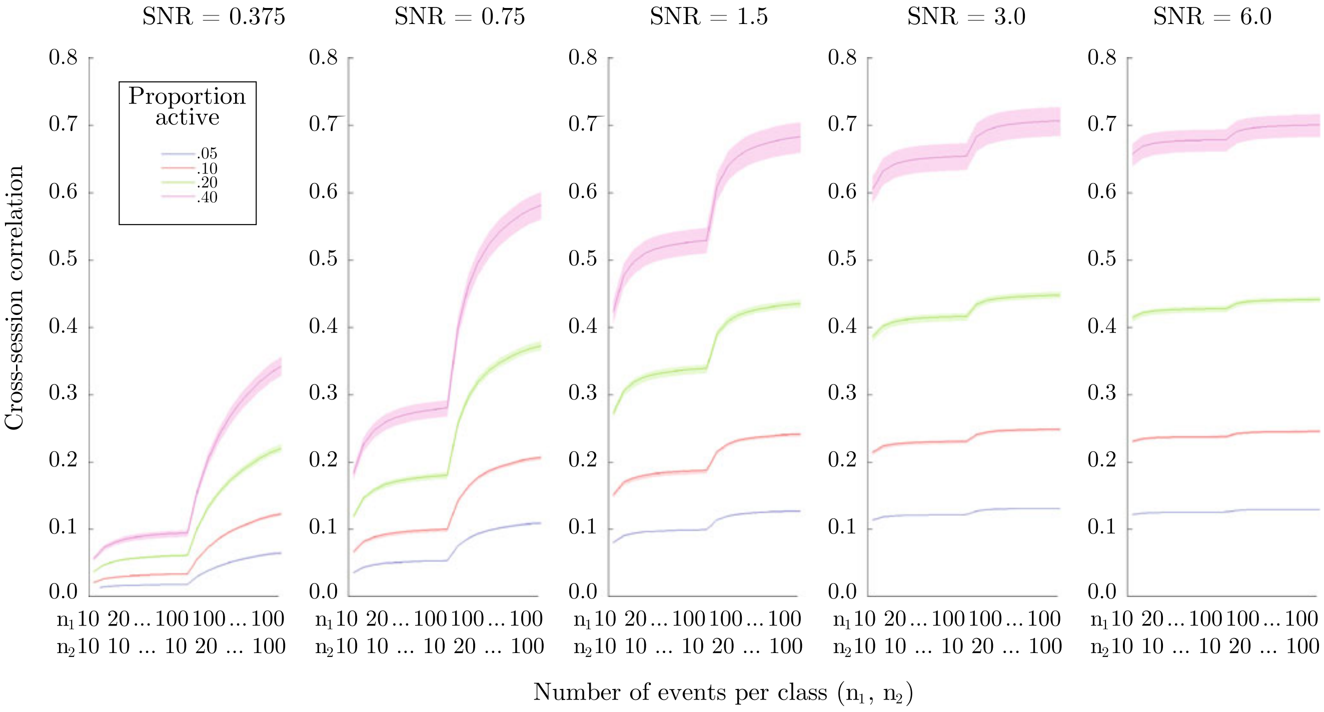

Fig. 3 Effects of signal-to-noise ratio (SNR), proportion active, and number of events on reliability for each of two simulated classes, as measured by Spearman's rho, for realistic simulations. Solid lines show the mean reliabilities given an underlying reliability of .6, and error bounds show the same given an underlying reliability of .3 (lower) or .9 (upper)

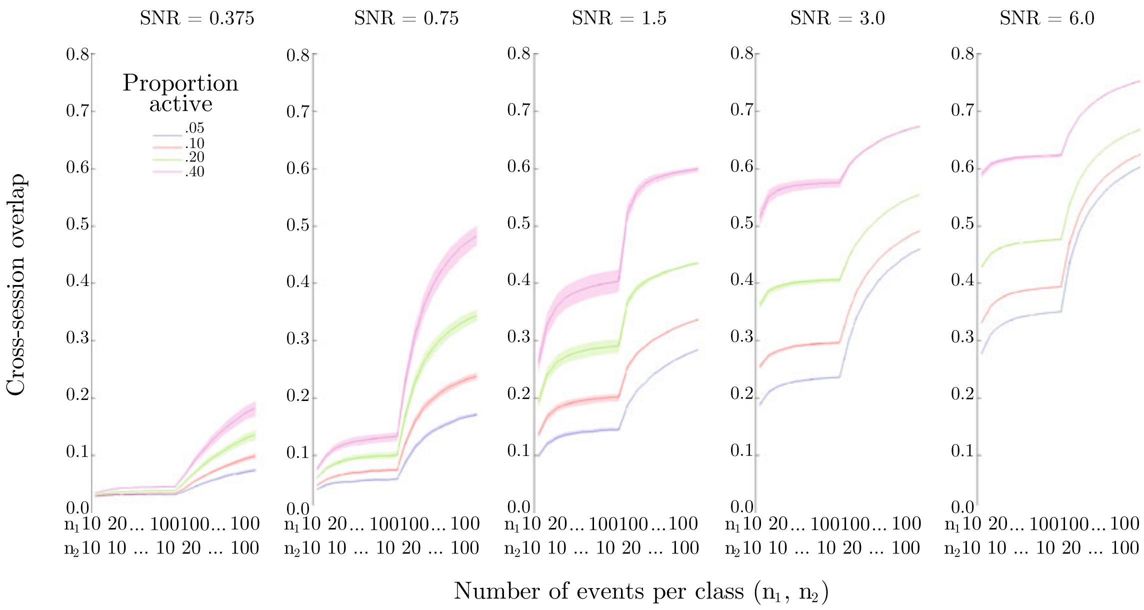

Fig. 4 Effects of signal-to-noise ratio (SNR), proportion active, and number of events on reliability for each of two simulated classes, as measured by Jaccard overlap, for realistic simulations. Solid lines show the mean reliabilities given an underlying reliability of .6, and error bounds show the same given an underlying reliability of .3 (lower) or .9 (upper)

limit on the number of events (i.e., a large disparity must result from one scan having near the maximum number of events, and the other near the minimum; in the limit of zero power, all *t*s will be drawn from the null distribution and will reflect run-specific—presumably nonreplicable—noise, whereas with high power, all *t*s will come from the noncentral *t* distribution dictated by the effect size in each particular voxel). Lending support to this interpretation is the fact that the relationships between n_{dif} and reliability are almost identical across all values of SNR and proportion active. In other words, rather than depending on these variables, the phenomenon reflects a stable relationship between n_{dif} and reliability, such that higher values of n_{dif} yield lower estimates of reliability.

Table 1 Average N_{class} values at which reliability (Spearman's rho) reached 90 % of its maximum, conditioned on signal-to-noise ratio and proportion active (given as ranges; e.g., >.05), separately for the ideal and realistic simulations

	$N_{90\%ile}$			
SNR	Ideal >.05	Realistic =.05	∈{.10, .20}	=.40
0.375	130	90	90	90
0.75	105	75	75	80
1.5	85	65	65	65
3.0	10	15	20	20
6.0	10	10	10	10

These are within-simulation maxima: The ideal simulation considered designs with up to 150 events per class, whereas the realistic simulation considered designs with only up to 100 events, so the conditional maximum levels of reliability differ between the two

Discussion

It is intuitively obvious to most researchers that the reliability of a result will depend on the number of events available for deriving that result. However, the exact relationship between the two—and how this relationship is affected by other factors such as the SNR or extent of activation—has never been investigated. This issue has grown in importance with the rise in popularity of *a posteriori* event definition, which mitigates the usefulness of some of the earlier work on "detection power" and "estimation efficiency." Our results from both simulations and real-data analyses confirm that the number of events has a strong effect on reliability, and that this effect depends on other factors in a complex way.

Moreover, the results from our analyses of real and simulated data correspond to a relatively high degree. For example, at similar levels of SNR, both show that reliability depends in similar ways on the number of events, and likewise, the proportion of activity exerts an influence in both. However, some

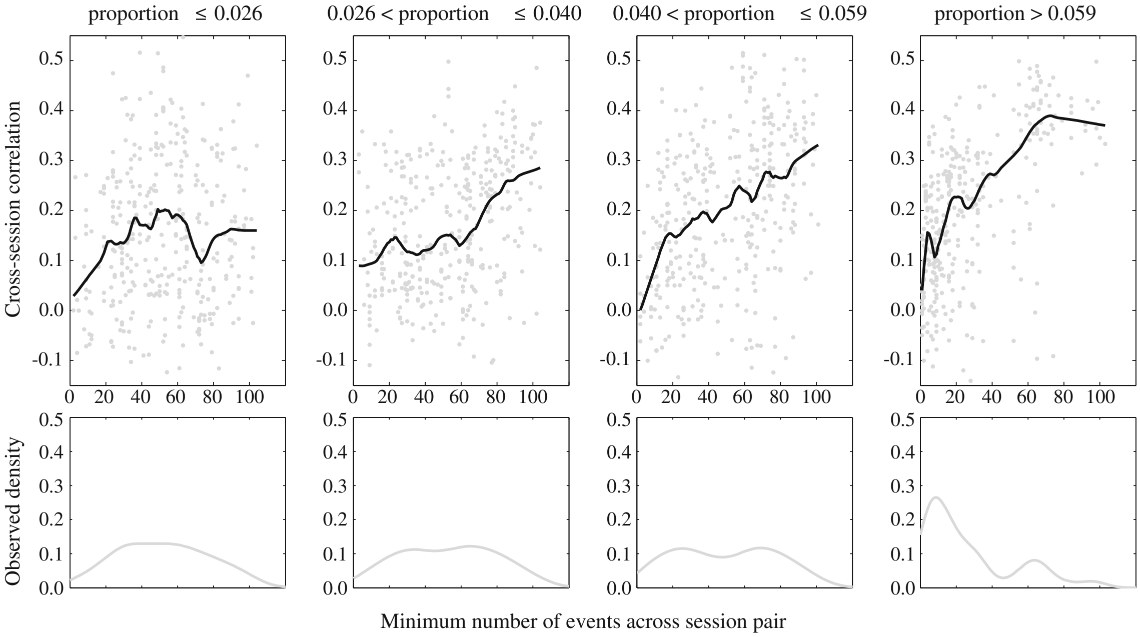

Fig. 5 Upper panels: Effects of the minimum number of events across a session pair for real data, stratified by proportions active. For this and all subsequent figures, individual pairwise correlations are shown as light gray dots; the darker lines give robust loess curves, fit continuously to 20 % of the data. Lower panels: Density estimates showing the amount of data at each point on the *x*-axis of each stratum

differences are noticeable: Most obviously, the reliabilities observed in the real data exceed what would be predicted on the basis of the simulations, given the relatively low proportions of activity that we observed. Although the mean trends from our real analyses bear a striking resemblance in form to those from the simulations, the individual pairwise reliability measures are much more variable, with values ranging from below –.1 to above .5. These differences may be due to mismatches between our simulation assumptions and the real data or to quirks in the real data.[9]

In light of these differences, although our results may be useful as a rough guide to the reliability of a particular result given the number of events, SNR, and extent of activation, they should be only a part of such a consideration. Firstly, both our simulation and real-data results depend on specific choices—for instance, on how we simulated our data or what preprocessing steps we applied in the real-data analysis. Moreover, certainly other factors contribute to reliability (Bennett & Miller, 2010; Gorgolewski et al., 2013), and although many of these were hopefully subsumed in our manipulation of SNR, other aspects are certainly specific to a particular task or contrast. Lastly, reliability measures are themselves subject to estimation error, so that even if a contrast perfectly matched the assumptions used in our simulations, it would be possible to obtain an estimated reliability above the upper bound that we reported.

These caveats notwithstanding, our results may be useful in a more qualitative way. For instance, they demonstrate the idea that the impact of number of events depends on the SNR and the extent of activity of the measured effect. In particular, outside of a certain range on these factors, number of events has almost no impact—or conversely, within a certain other range, the number of events has a marked effect (see also Table 1). In our simulations, the impact of number of events was small for proportions of activity below roughly .05, and also for SNRs above roughly 3 (although these factors interacted, so that we still found a noticeable impact for low proportions at low SNRs, and likewise for high proportions at high SNRs). Our real-data results were largely in line with these trends, although we did not observe as wide a range of SNRs or proportions of activity, and so the predicted pattern could only be partially confirmed. Likewise, without placing too much credence in the largely descriptive curves in Figs. 5 and 6,

[9] A few obvious possible differences between the two include the spatial autocorrelation in the real data—which may inflate reliability measures, or cause our measures of proportions active in the two situations to have slightly different meanings—and the all-or-none “active” versus “inactive” nature of the simulation, as opposed to the gradation that is sure to exist in real data.

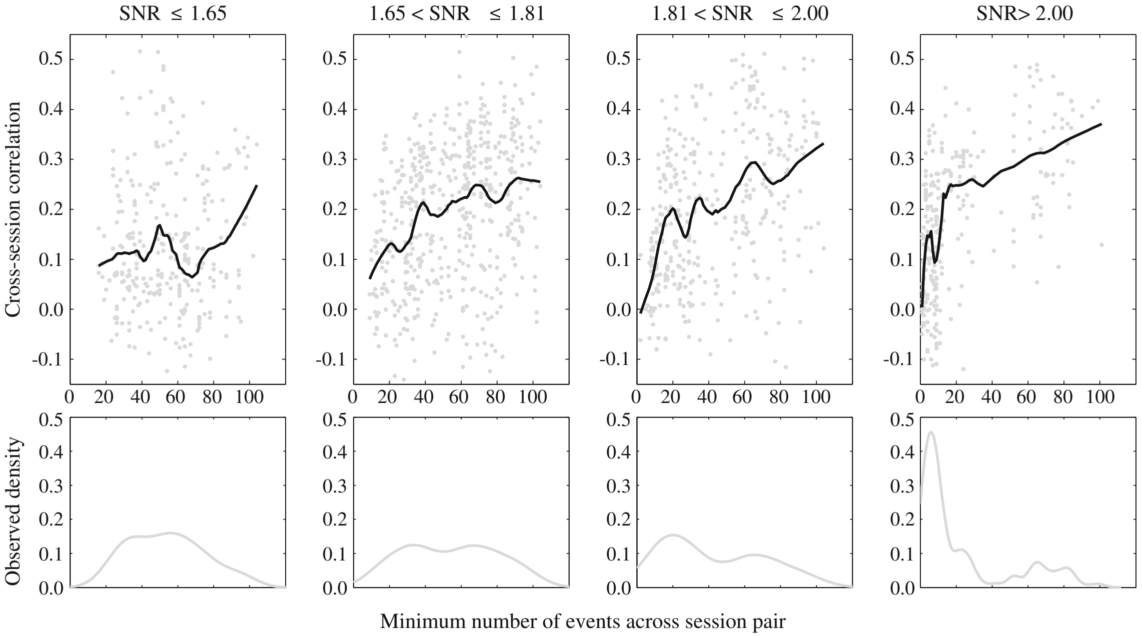

Fig. 6 Upper panels: Effects of the minimum number of events across a session pair for real data, stratified by signal-to-noise ratio (SNR). Lower panels: Density estimates showing the amount of data at each point on the *x*-axis of each stratum

the point at which the curves reach 90 % of their maxima seems to be near the values suggested by Table 1—that is, ~65 events.

Additionally, our results give an idea as to the magnitude of the impact of number of events. Although it is hard to quantify the unique impact of each factor that we investigated, due to

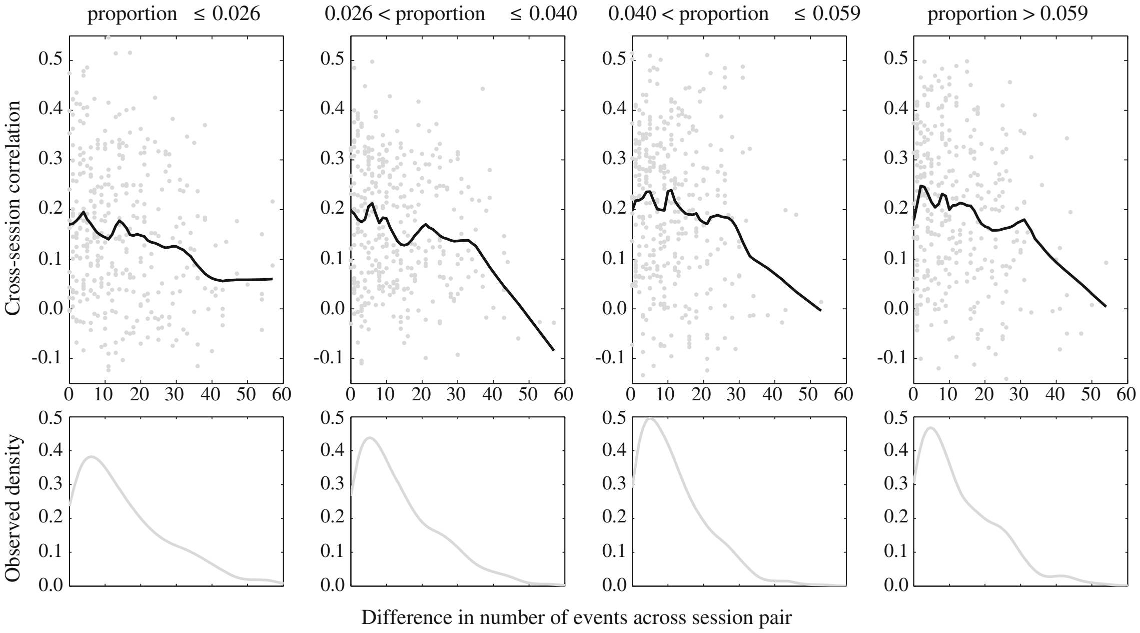

Fig. 7 Upper panels: Effects of the difference in numbers of events across a session pair for real data, stratified by proportion active. Lower panels: Density estimates showing the amount of data at each point on the *x*-axis of each stratum

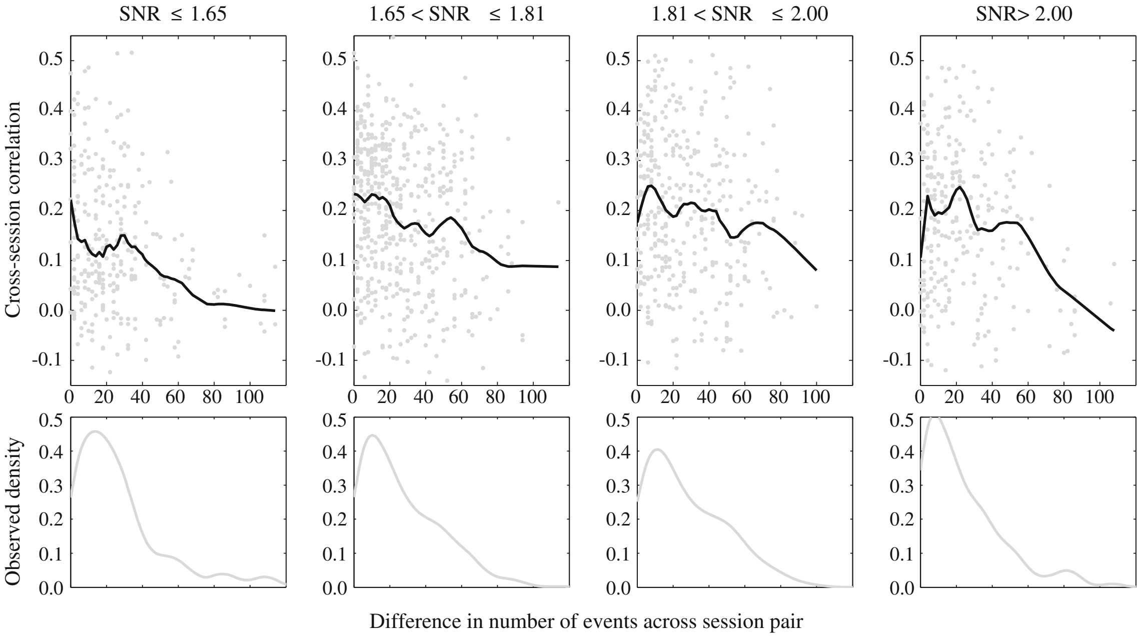

Fig. 8 Upper panels: Effects of the difference in numbers of events across a session pair for real data, stratified by signal-to-noise ratio (SNR). Lower panels: Density estimates showing the amount of data at each point on the *x*-axis of each stratum

their interactions, number of events exerted an influence on the same order of magnitude as that of SNR or proportion active (looking in realistic ranges of each variable, and likewise holding the others at realistic levels, in our simulations reliability changed by roughly .5 with a fourfold change in SNR, by .6 with a tenfold change in proportion active, and by roughly .4 with a tenfold change in number of events). This serves both to emphasize the need for the best design possible and to point out the importance of accounting for differences in numbers of events when comparing results. However, researchers should keep in mind that these are single-run reliabilities; the reliability of any result at the experiment-wide level will of course depend on the number of runs per participant, the number of participants, and so forth, as has been described elsewhere (Mumford & Nichols, 2008).

The results of our real-data analyses reveal additional impact of the number of events, one which is particularly relevant to researchers studying individual differences, or to anyone relying on single-run results: As predicted, Figs. 7 and 8 show that reliability decreases as a function of the difference in numbers of events between two scans. Unlike the effect of number of events above, this effect appears to be largely independent of proportion of activity and SNR, and so might constitute more of a general principle to which researchers must always be sensitive. As we discussed in the Results, this phenomenon may be due here to the fact that a large difference necessarily means that one run had many events and the other had near zero. In other words, the difference in the numbers of events *per se* is not what matters, but rather, the difference in the two scans in terms of where they lie on the power function. That is, two hypothetical scans with 600 and 700 events will probably not be as different as two scans with 10 and 110. However, our results do not speak to the exact cause of this effect, so we merely highlight the issue and recommend that researchers proceed with caution in such situations.

Given the increasingly common use of *a posteriori* event definition, the role played by the number of events underlying a result has gained new importance. Researchers can no longer rely strictly on earlier work guiding experimental design for achieving certain levels of estimation efficiency or contrast power. Not only this, but our results demonstrate that even under ideal conditions, and even with sufficient numbers of events, the reliability of a single-run-level SPM may be quite low. Of course, most researchers operate with group-level results (but cf. Miller et al., 2009, for a discussion of the remarkable amount of dissimilarity possible between group- and individual-level results), but if one event is systematically more or less common than another, these effects on reliability will carry up to the group level. Therefore, any result, whether or not it is principally concerned with intra- and interindividual reliability *per se*, must be considered through the lens of the number of events on which that result is based.

Author note This research was supported by the Institute for Collaborative Biotechnologies through Grant No. W911NF-09-D-0001 from the U.S. Army Research Office. The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

Bennett, C. M., & Miller, M. B. (2010). How reliable are the results from functional magnetic resonance imaging? *Annals of the New York Academy of Sciences, 1191,* 133–155.

Bennett, C. M., & Miller, M. B. (2013). *Differences in reliability of block versus event-related designs*. Manuscript submitted for publication.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14,* 365–376. doi:10.1038/nrn3475

Caceres, A., Hall, D. L., Zelaya, F. O., Williams, S. C. R., & Mehta, M. A. (2009). Measuring fMRI reliability with the intra-class correlation coefficient. *NeuroImage, 45,* 758–768.

Carp, J. (2012). The secret lives of experiments: Methods reporting in the fMRI literature. *NeuroImage, 63,* 289–300.

Dale, A. M. (1999). Optimal experimental design for event-related fMRI. *Human Brain Mapping, 8,* 109–114.

Desmond, J. E., & Glover, G. H. (2002). Estimating sample size in functional MRI (fMRI) neuroimaging studies: Statistical power analyses. *Journal of Neuroscience Methods, 118,* 115–128. doi:10.1016/S0165-0270(02)00121-8

Eklund, A., Andersson, M., Josephson, C., Johannesson, M., & Knutsson, H. (2012). Does parametric fMRI analysis with SPM yield valid results? An empirical study of 1484 rest datasets. *NeuroImage, 61,* 565–578.

Friston, K. J., Zarahn, E., Josephs, O., Henson, R. N. A., & Dale, A. M. (1999). Stochastic designs in event-related fMRI. *NeuroImage, 10,* 607–619.

Gorgolewski, K. J., Storkey, A. J., Bastin, M. E., Whittle, I., & Pernet, C. (2013). Single subject fMRI test–retest reliability metrics and confounding factors. *NeuroImage, 69,* 231–243.

Green, D. M., & Swets, J. A. (1974). *Signal detection theory and psychophysics (Rev. ed)*. Huntington, NY: Krieger.

Killeen, P. R. (2005a). An alternative to null hypothesis significance tests. *Psychological Science, 16,* 345–353.

Killeen, P. R. (2005b). Replicability, confidence, and priors. *Psychological Science, 16,* 1009–1012.

Liu, T. T., & Frank, L. R. (2004). Efficiency, power, and entropy in event-related fMRI with multiple trial types: Part I. Theory. *NeuroImage, 21,* 387–400.

Liu, T. T., Frank, L. R., Wong, E. C., & Buxton, R. B. (2001). Detection power, estimation efficiency, and predictability in event-related fMRI. *NeuroImage, 13,* 759–773.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide*. Mahwah, NJ: Erlbaum.

McGonigle, D. J. (2012). Test–retest reliability in fMRI: or how I learned to stop worrying and love the variability. *NeuroImage, 62,* 1116–1120.

McGonigle, D. J., Howseman, A., Athwal, B., Friston, K., Frackowiak, R., & Holmes, A. (2000). Variability in fMRI: An examination of intersession differences. *NeuroImage, 11,* 708–734.

Miller, M. B., Donovan, C. L., Bennett, C. M., Aminoff, E. M., & Mayer, R. E. (2012). Individual differences in cognitive style and strategy predict similarities in the patterns of brain activity between individuals. *NeuroImage, 59,* 83–93.

Miller, M. B., Donovan, C.-L., Van Horn, J. D., German, E., Sokol-Hessner, P., & Wolford, G. L. (2009). Unique and persistent individual patterns of brain activity across different memory retrieval tasks. *NeuroImage, 48*(3), 625–635.

Miller, M. B., Van Horn, J. D., Wolford, G. L., Handy, T. C., Valsangkar-Smyth, M., Inati, S., . . . Gazzaniga, M. S. (2002). Extensive individual differences in brain activations associated with episodic retrieval are reliable over time. *Journal of Cognitive Neuroscience*, *14*, 1200–1214.

Mumford, J. A., & Nichols, T. E. (2008). Power calculation for group fMRI studies accounting for arbitrary design and temporal autocorrelation. *NeuroImage, 39,* 261–268.

Mumford, J. A., Turner, B. O., Ashby, F. G., & Poldrack, R. A. (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *NeuroImage, 59,* 2636–2643. doi:10.1016/j.neuroimage.2011.08.076

Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2011). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *NeuroImage, 59,* 2142–2154.

Turner, B. O., Donovan, C.-L., & Miller, M. B. (2013). *Dissociating memory processes by their differences in neural stability.* Poster presented at the annual meeting of the Cognitive Neuroscience Society.

Van Dijk, K. R. A., Sabuncu, M. R., & Buckner, R. L. (2012). The influence of head motion on intrinsic functional connectivity MRI. *NeuroImage, 59,* 431–438.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science, 4,* 274–290. doi:10.1111/j.1745-6924.2009.01125.x

Wager, T. D., & Nichols, T. E. (2003). Optimization of experimental design in fMRI: A general framework using a genetic algorithm. *NeuroImage, 18,* 293–309.

Worsley, K. J., & Friston, K. J. (1995). Analysis of fMRI time-series revisited—again. *NeuroImage, 2,* 173–181.