Correspondence: fMRI replicability depends upon sufficient individual-level data

Derek Evan Nee

Department of Psychology and College of Medicine, Florida State University, Tallahassee, FL 32306-4301

Main Text: 1159 words

Number of Figures: 2 (2 Supplemental)

Corresponding Author:

Derek Evan Nee
Florida State University
1107 W Call St
Tallahassee, FL 32306-4301
nee@psy.fsu.edu

The reproducibility of task-based functional magnetic resonance imaging (fMRI), or lack thereof, has become a topic of intense scrutiny[1,2]. Relative to other human techniques, fMRI has high costs associated with data collection, data storage, and data processing. To justify these costs, the inferences gained from fMRI need to be robust and meaningful. Hence, although collecting large, sufficiently powered datasets may be costly, this is favorable to collecting many insufficiently powered datasets from which reliable conclusions cannot be drawn. However, it can be difficult to determine *a priori* how much data needs to be collected. Although power analyses can help[3], accurately calculating power itself requires an appropriate estimate of the expected effect size, which can be hard to obtain if previous studies had insufficient data to produce reliable effect size estimates. Furthermore, mechanistic basic science explores novel phenomena with innovative paradigms such that extrapolation of effect sizes from existing data may not be appropriate.

In light of these issues, many studies rely on rules-of-thumb to determine the amount of data to be collected. For example, Thirion et al[4] suggested that twenty or more participants are required for reliable task-based fMRI inferences. Turner et al[5] recently pointed out that such recommendations are outdated, and set out to empirically estimate replicability using large datasets. The authors found that even datasets with one-hundred or more participants can produce results that do not replicate, suggesting that large sample sizes are necessary for task-based fMRI.

It is typical for considerations of power in task-based fMRI to focus on sample size. This is because between-subject variability tends to dominate within-subject variability, such that sampling more subjects tends to be a more effective use of scanning time that scanning individuals for longer[3,4]. For example, Mumford and Nichols[3] suggested that scanning time per individual was maximally cost effective at between four to eight minutes. Perhaps drawing from such observations, large task-based fMRI data collections such as the Human Connectome Project (HCP) have used batteries of tasks wherein each task is scanned on the order of ten minutes[6]. However, using data from the HCP and other data of similar scanning durations, Turner et al[5] demonstrated that task-based fMRI can be unreliable.

With the rising popularity of resting-state fMRI, investigators have examined the duration of resting-state data needed for reliable parameter estimates. Some have suggested that parameter estimates are stable after 5-10 minutes of resting-state scans[7]. On the surface then, it would seem that both the resting-state and task-based literatures have converged on similar amounts of individual-level data. However, parameters estimated from rest use the entire (cleaned) data time-series, while task-based fMRI splits the time-series into composite cognitive events. For example, in a rapid event-related design, there may be approximately 4-6 seconds of *peak signal* attributable to a given transient event of interest (e.g. a choice reaction). If twenty such events exist in a ten-minute task run, that amounts to less than two minutes of signal attributable to that task event. In such cases, it is likely that parameter estimates would benefit from additional measurements at the individual level.

To examine the impact of individual-level measurements on task-based fMRI replicability, I re-analyzed data from a recently published pair of datasets[8,9]. Each dataset estimated five contrasts-of-interest spanning main effects and an interaction in a 2 x 2 x 2 factorial design. Given the observations of Turner et al[5], the sample sizes employed (n=24) should produce low replicability. Previously, I suggested the reproducibility in these data were good[8,9], but I did not compute the replicability measures calculated by Turner et al[5], so it is possible that the results were not as reproducible as I believed them to be. On the other hand, ~one-two hours of task data were collected for each individual, which could have facilitated

reliability. To examine this matter, I computed the replicability measures of Turner et al[5] on randomly sub-sampled independent datasets on five contrasts-of-interest. I varied the amount of individual-level data from ~ten minutes (one task run) to ~one hour (six task runs). I also varied the sample size from sixteen to twenty-three individuals with sixteen matching the minimum examined by Turner et al[5] and twenty-three being the maximum that can be split into independent groups in the forty-six participants examined. All data and code are available at https://osf.io/b7y9n.

Figure 1 shows the results at n=16. When only one run is included for each individual, the replicability estimates all fall in the ranges reported by Turner et al[5]. However, reproducibility markedly improved with more data at the individual-level. While there are some indications of diminishing returns after four runs, there were clear benefits to more scans at the individual level. Figure 2 reports the results at n=23, which again show clear benefits to reproducibility when more than one run is collected. For example, the mean peak replicability with two runs (~65%) matches observations in Turner et al[5] at n=64. These data suggest that in some cases, scanning more at the individual-level nets more reproducibility than scanning more individuals. Furthermore, no contrast in Turner et al[5] approached perfect replicability with any combination of measure, sample size, and threshold, whereas multiple combinations produced near perfect replicability for the Contextual Control contrast with as little as six runs at n=16 (Supplemental Figure 1). In the most striking such case, I find nearly 90% of the peaks replicate on average with four runs at n=23 (Supplemental Figure 2), which again exceeded the observations of Turner et al[5] even at the largest sample size (n=121). These data paint a much more reliable picture of task-based fMRI at modest sample sizes when individuals are adequately sampled.

These observations raise the question of how much individual-level data are needed. Mumford and Nichols[3] found optimal cost effectiveness at six minutes of task with a simple on/off block design. Hence, three minutes of task effectively contributed to the contrast numerator and three minutes to the denominator. In the event-related design studied here, there were approximately thirty seconds of scans per parameter of interest per task run. Hence, six runs were necessary for this task to reach the same scans per contrast parameter as in Mumford and Nichols[3]. Six runs at n=23 can produce near perfect replicability (Figure 2, Contextual Control contrast), or poor replicability (Figure 2, Verbal contrast). Thus, the amount of individual-level data necessary for replicability will depend upon the phenomenon of interest. Furthermore, there appear to be diminishing returns after approximately four task runs, which may owe to the duration of time participants can remain attentive and still. Therefore, I draw the following conclusions:

1) Replicability cannot be judged on sample size alone.
2) The amount of data at the individual-level is a critical determinant of replicability.
3) 5-10 minutes of a task will be sufficient for only the simplest designs with large effect sizes.
4) Aiming for at least three minutes of scans per contrast parameter is a good starting point.
5) Split scans across multiple sessions if >45 minutes of task is needed.

Methods

Full details of the participants, task, preprocessing, and modeling can be found in my previous reports[8,9]. Briefly, the task manipulated two forms of cognitive control (contextual control, temporal control) and stimulus domain (verbal, spatial) in a 2 x 2 x 2 factorial design. Five contrasts from the factorial design were included in this report: contextual control, temporal control, temporal control x contextual control, verbal (> spatial), and spatial (> verbal). On each block, participants performed a sequence-matching task

in a given stimulus domain. Then, sub-task phases orthogonally manipulated the cognitive control demands. In the original report, we examined stimulus domain (verbal>spatial, spatial>verbal) across all trials. But here, I use only the sub-task phases so that all contrasts have the same amount of data at the individual level. A separate contrast estimate was created for each individual and each run. I included data from 46 participants, excluding participants in the original reports that did not complete all of the task runs. 23 participants performed 12 scanning runs and 23 participants performed 6 scanning runs, wherein each scanning run took approximately 10 minutes to complete. Data and code are available at https://osf.io/b7y9n.

Following the procedures of Turner et al[5], replicability was determined by pairwise comparison of group-level t-statistic maps. For each analysis, the data were randomly split into two independent groups 500 times. Analyses varied the number of runs included at the individual level (1, 2, 4 or 6) by randomly selecting a subset of the data, and also the number of individuals (16 or 23). Extra-cranial voxels were masked out and voxels for which t-statistics could not be computed (i.e. due to insufficient signal across participants) were discarded prior to computations of replicability.

The first analysis examined the voxel-wise correlation of t-statistics across all voxels. Subsequent analyses examined Jaccard overlap on thresholded t-statistic maps where the Jaccard overlap indicates the proportion of results that replicate. Although Turner et al[5] utilized both positive and negative activations for their Jaccard overlap calculations, here I use only positive activations given that two of the contrasts are the inverses of one another. Following Turner et al[5], Jaccard overlap was computed at the voxel-level by first thresholding the complete group dataset and determining the number of significant voxels, *v*, at a voxel-wise threshold. This map represented the "ground truth." Then, in each pair of sub-sampled datasets, the conjunction of the top *v* voxels was divided by their union to determine the proportion of replicated voxels.

The voxel-level procedure does not attempt to control false-positives for each group analysis. Therefore, low replicability in this measure might be anticipated by the inclusion of false-positives. So, Turner et al[5] also performed family-wise error correction using cluster-level thresholding in each group map, and calculated the number of overlapping voxels passing correction. However, cluster-level correction allows for cluster-level, but not voxel-level inference. That is, the cluster is the unit of significance rather than the voxels within the cluster. Noting the number of overlapping voxels therefore does not capture the essence of whether a cluster has replicated or not. Therefore, I modified the procedure to determine the number of overlapping clusters rather than voxels. A cluster was deemed to have replicated if at least half of the voxels of that cluster were present in the replicate. Half is an arbitrary number intended to safeguard against trivial overlap. Finally, Turner et al[5] examined peak overlap determined by whether the peak of a given cluster was also significant in the replicate. This is likely to be an important practical metric of replicability given that replication attempts will often examine a small radius around the peak of a previous report.

As in Turner et al[5] each Jaccard overlap was performed at both a conservative threshold (depicted in the main text) and liberal threshold (depicted in the supplemental material). The liberal/conservative thresholds were as follows: voxel-level: $p < 0.00025/0.00000025$; cluster-level: $p < 0.05$ height, 1019 voxel extent/$p < 0.01$ height, 300 voxel extent, each achieving alpha < 0.01 according to 3dClustSim in AFNI. Interestingly, although it has been reported that liberal cluster-forming thresholds have inflated false

positives[10], which would be expected to harm replicability, replicability measures improved at the more liberal thresholds, which was also observed in Turner et al[5] to some extent.

References

1. Button, K. S. *et al.* Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14,** 365–376 (2013).

2. Szucs, D. & Ioannidis, J. P. A. Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biol.* **15,** e2000797 (2017).

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

4. Thirion, B. *et al.* Analysis of a large fMRI cohort: Statistical and methodological issues for group analyses. *Neuroimage* **35,** 105–20 (2007).

5. Turner, B. O., Paul, E. J., Miller, M. B. & Barbey, A. K. Small sample sizes reduce the replicability of task-based fMRI studies. *Commun. Biol.* **1,** (2018).

6. Barch, D. M. *et al.* Function in the human connectome: task-fMRI and individual differences in behavior. *NeuroImage* **80,** 169–189 (2013).

7. Van Dijk, K. R. A. *et al.* Intrinsic functional connectivity as a tool for human connectomics: theory, properties, and optimization. *J. Neurophysiol.* **103,** 297–321 (2010).

8. Nee, D. E. & D'Esposito, M. The hierarchical organization of the lateral prefrontal cortex. *eLife* **5,** (2016).

9. Nee, D. E. & D'Esposito, M. Causal evidence for lateral prefrontal cortex dynamics supporting cognitive control. *eLife* **6,** (2017).

10. Eklund, A., Nichols, T. E. & Knutsson, H. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proc. Natl. Acad. Sci. U. S. A.* **113,** 7900–7905 (2016).

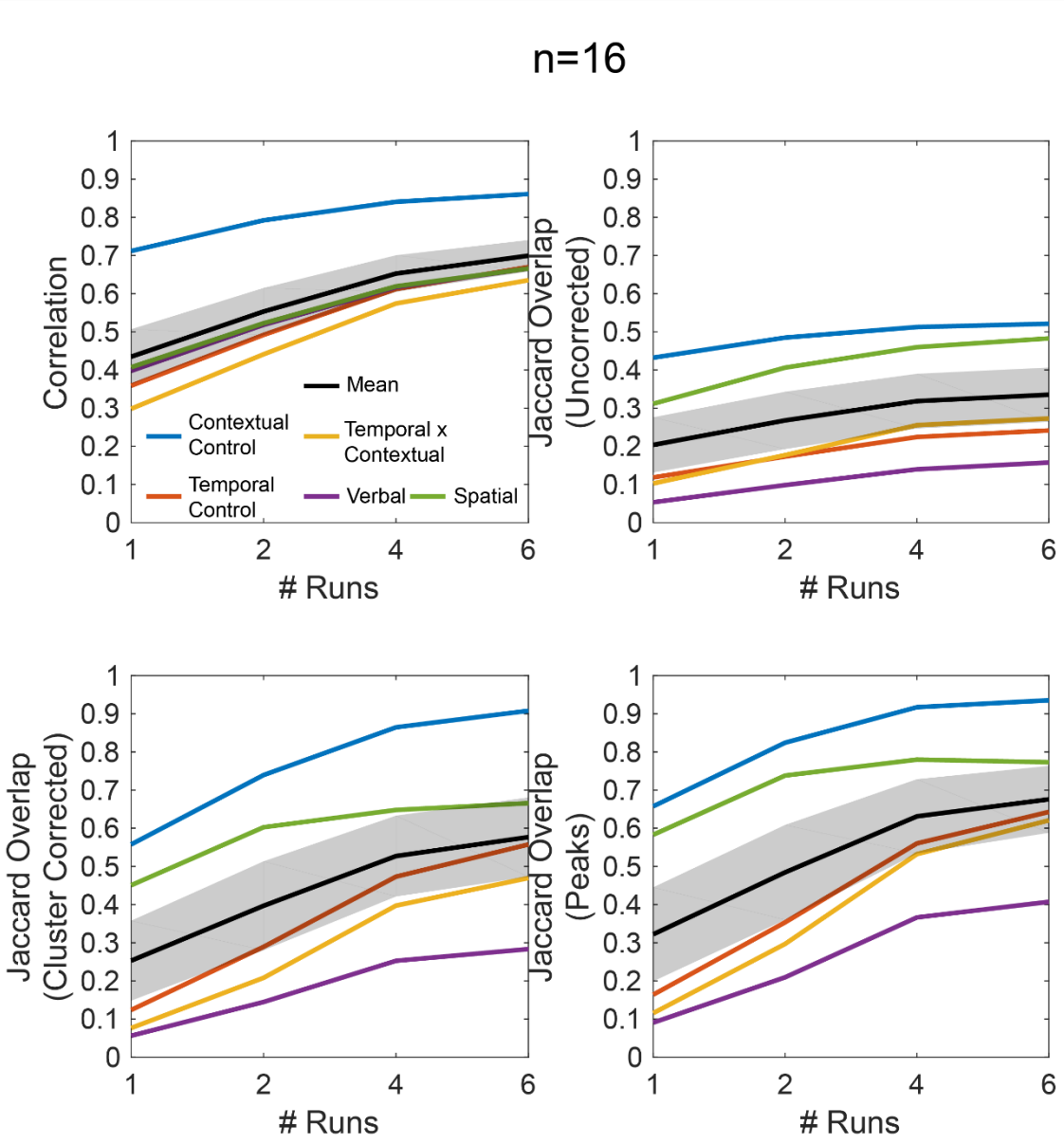


Figure 1. Replicability estimates at n=16. Metrics correspond to those used in Turner et al[5]. Jaccard Overlaps were calculated using conservative thresholds comparable to those reported in Turner et al[5].

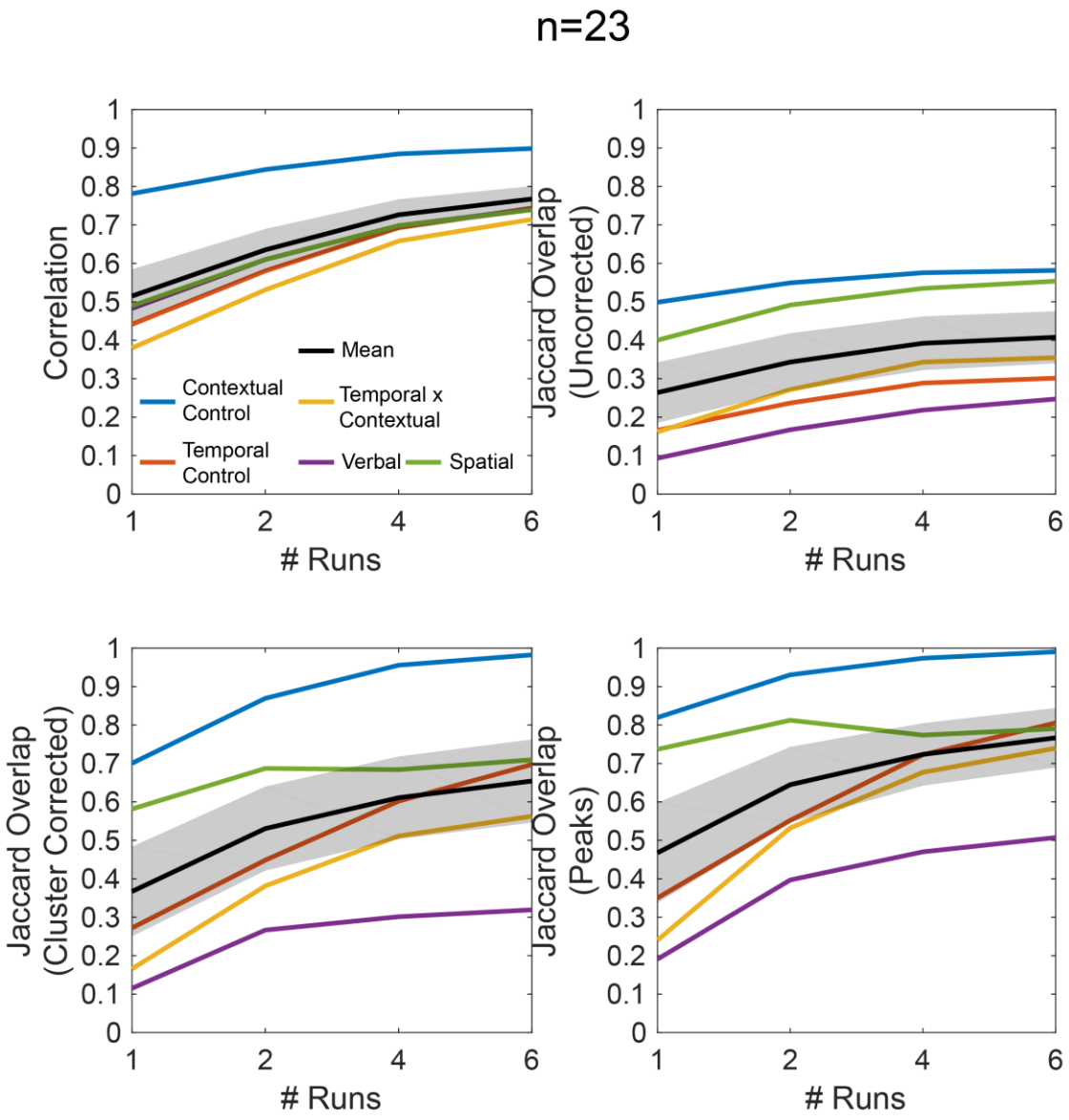
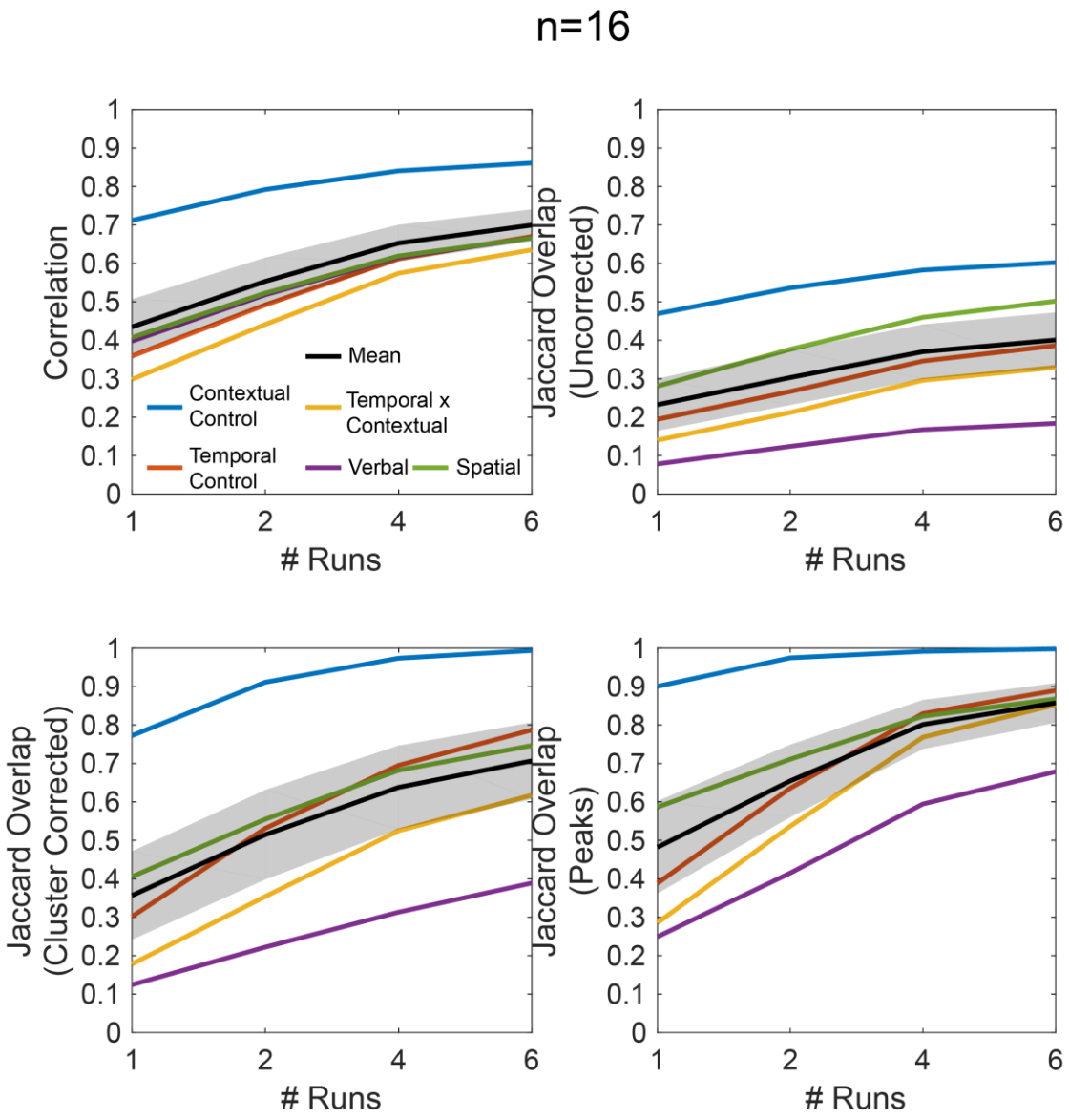
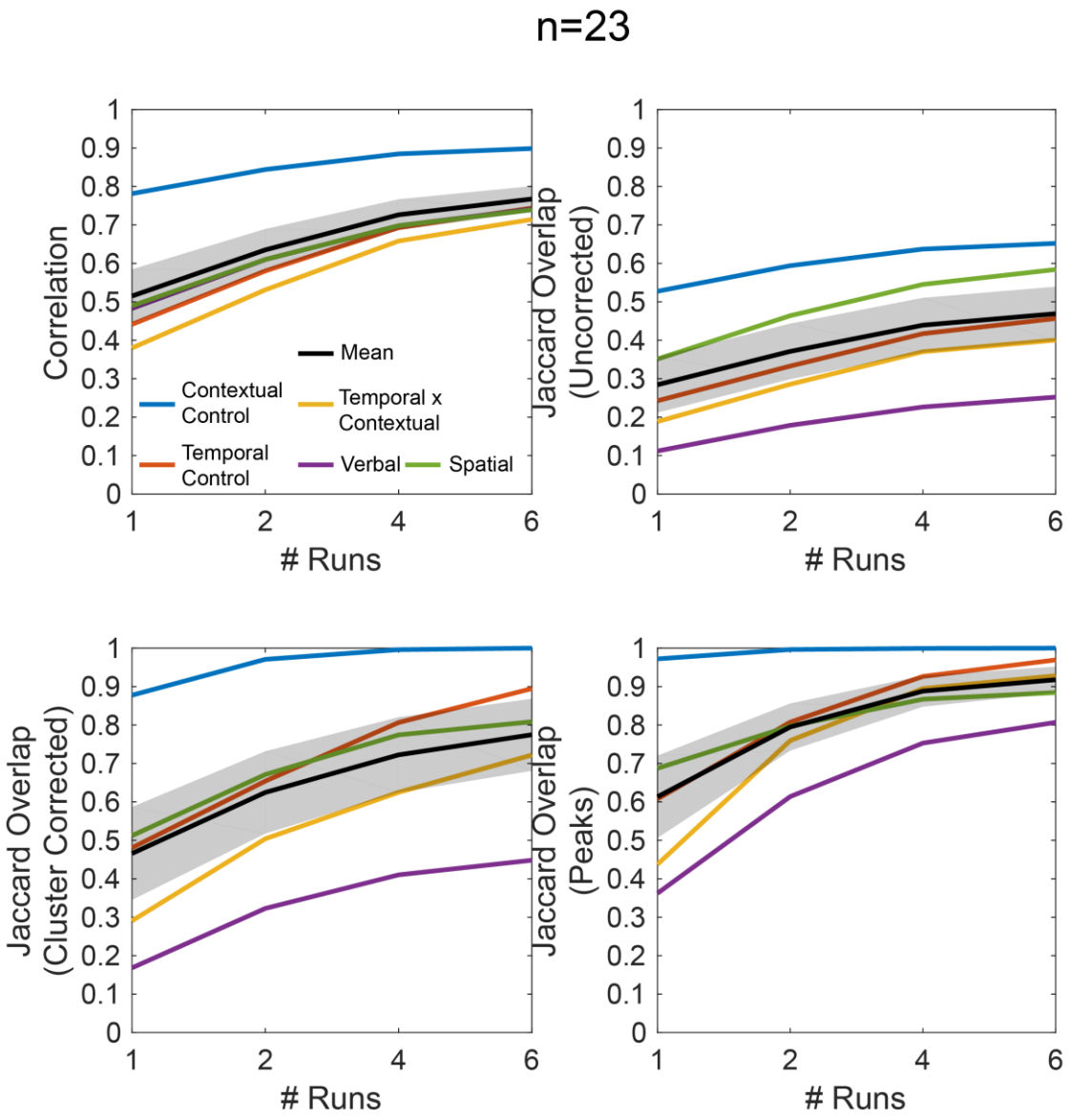


Figure 2. Replicability estimates at n=23. Other details match Figure 1.



Supplemental Figure 1. Details are identical to Figure 1, but Jaccard Overlap was computed using liberal threshold comparable to those reported in Turner et al[5].



Supplemental Figure 2. Replicability estimates at n=23 with liberal thresholding. Other details match Supplemental Figure 1.