Challenges in assessing voxel-wise single-subject level benefits of MB acceleration

Ritu Bhandari[1], Valeria Gazzola[1,2] and Christian Keysers[1,3]

1. Netherlands Institute for Neuroscience, KNAW, Amsterdam, The Netherlands
2. Department of Psychology, University of Amsterdam, The Netherlands

Abstract: In this technical note, we present the challenges that prevent us from directly comparing sequences with and without MB acceleration at the single subject level. Using fMRI data collected with MB1S2 (TR 2.45s), MB2S2 (TR 1.22s) and MB4S2 (TR 0.63s), we note the CNR differences in the images acquired with different sequences which leads to global mean scaling that render the direct comparison of parameter estimates meaningless. Directly comparing t-values of participants across different acquisition sequences is meaningless because of the difference in degrees of freedom (df) introduced by the higher number of volumes acquired at higher multiband. Z-transformation of the t-statics to correct for the difference in degree of freedoms suggests that sequences without MB outperform sequences with MB acceleration. However, this may be due to an excessive penalty caused by inappropriate df estimation. Thus with the current evidence presented in this and previous studies that tested the impact of MB on task related-statistics, the field lacks empirical evidence for the effects of MB on individual subject statistics. We discuss the possible alternatives such as use of Bayesian statistics.

Background: Multiband (MB) or Simultaneous multi-slice (SMS) acquisition schemes allow the acquisition of magnetic resonance imaging (MRI) signals from more than one spatial coordinate at a time. Over 40 studies with both resting state and task-based fMRI have tested the benefits of using this technique (Bhandari et al., in prep). The MB literature so far has unequivocally shown that the noise increases as the MB acceleration increases (Golestani et al., 2017). However, optimal cleaning procedures and higher statistical power due to increased sampling rate results in improved detection of functional networks (Boubela, Kalcher, Nasel, & Moser, 2014; Feinberg et al., 2010; Griffanti et al., 2014; Preibisch, Castrillón G., Bührer, & Riedl, 2015; Smitha et al., 2018). These advantages have been seen in many resting state and some task based studies. In our recent study we analysed task based fMRI data acquired in the same participants with multiple sequences and show that at group level, sequences with MB acceleration perform better in terms of voxel-wise t-values and total number of activated voxels (Bhandari et al. in prep). This was one of the only two studies (Boyacioğlu, Schulz, Koopmans, Barth, & Norris, 2015) that concluded the effects of MB on a group level, voxel-wise results. Most other studies that assessed task correlated BOLD for quantifying performance of MB acceleration, concluded their findings using subject-level summary statistics such as mean t-values within a region of interest (ROI)(Kiss, Hermann, Vidnyánszky, & Gál, 2018; McDowell & Carmichael, 2018; Sahib et al., 2018, 2016; Todd et al., 2016). This was surprizing, given that most of the fMRI studies use voxel-wise, group level statistics. Moreover, the subject-level t-values cannot be compared directly as this will bias it towards the sequence which has more samples, and therefore more degrees of freedom (df).

Whether the benefits of MB at the group level go hand-in-hand with single-subject improvements of statistics remains unclear. Here we report the difficulties we encountered in trying to address this seemingly simple question.

Acquisition: Data were acquired from 24 subjects, on a 3T Philips scanner, using a commercial version of Philips' MB implementation (based on software release version R5.4). A 32-channel head coil was used. Functional data were acquired using different acquisition sequences (table1). **Task:** Two types of stimuli were used: Complex actions (CA) showed the hand interacting with the object in typical, goal directed actions. For example, the hand of the actor reached for the lighter placed on the table, grasped it, and lit the candle with it. Complex controls (CC) stimuli had the exact same setting as the CA but the actor's hand did not interact with or manipulate the object on the table, instead, made aimless hand movements. A block was composed of three movies of the same category (CA or CC) and lasted 7s. Each fMRI session was composed of 13 blocks per stimulus category for a total of 26 blocks, presented in a randomized order. The inter-block-interval lasted between 8 – 12 s and consisted of a fixation cross on a gray and blue background similar to the stimuli background. These sessions were presented to each subject multiple times, showing the same blocks but in different order, with each session acquired with a different acquisition scheme, varying in MB factor (table 1). The order of acquisition was randomized between subjects. Importantly, the duration of the sessions was similar (~8 min), but more functional volumes were acquired during sessions at lower TRs. **Preprocessing:** Data were preprocessed using SPM12 (Wellcome Trust Centre for Neuroimaging, UCL, UK) with Matlab version 8.4 (The MathWorks Inc., Natick, USA). Briefly, functional images were slice-time corrected and then realigned to the estimated average. Anatomical images were co-registered to the mean functional image, and segmented. The normalization parameters that were generated during segmentation were used to bring all the images to the MNI space. The resampled voxel size for the functional images was 2 × 2 × 2 mm and 1 x 1 x 1 for the anatomical scans. **Subject-level GLM:** Subject level GLM included CA and CC as two separate task predictors with each predictor having 13 blocks of 7s. Regressors included the six motion parameters estimated during realignment, first five principal components of cerebrospinal fluid and five principal components of white matter (total 16 regressors). As by default, the GLM involves a global normalization step, described in the SPM manual (https://www.fil.ion.ucl.ac.uk/spm/doc/spm12_manual.pdf), section 8.7 as "SPM computes the grand mean value $g_s = \sum_{n=1}^{N} \frac{g_{ns}}{N}$. This is the fMRI signal averaged over all voxels within the brain and all time points within session s. SPM then implements "Session-specific grand mean scaling" by multiplying each fMRI data point in session s by $100/g_s$."

Table 1. Parameters used for scanning.

Parameters	TR	2.45s	1.22s	0.63s
Multiband factor (MB)		none	2	4
SENSE acceleration factor (S)		2	2	2
Acquired voxel size (mm)		2.7 isotropic	2.7 isotropic	2.7 isotropic
Flip angle in degrees		79	64	50
Number of slices		44	44	44
Acquired volumes		200	400	780
Slice gap (mm)		0.27	0.27	0.27
Field of view (mm)		216x216x130.4	216x216x130.4	216x216x130.4

Results: Figure 1A shows mean functional images per TR from one representative participant. As can be seen, the white/grey matter contrast is lower for smaller TRs. A within-subject ANOVA to test the

difference between the CNR values of different acquisition schemes confirms that the CNR decreases, as the TR becomes smaller (figure 1B). The absolute values of voxels in fMRI are in arbitrary units, and SPM by default performs a global mean scaling to bring the time series to a common scale, rendering it comparable across subjects and groups using the same acquisition scheme. However, because the white/gray matter contrast is higher at longer TRs, the gray matter gets values further away from 100 for TRs with higher white/gray matter contrast. This results in gray matter voxels having different values on average at different TRs (within subject ANOVA comparing the beta values of the constant term at the subject level GLM across TR, $q_{fdr}<0.05$; figure 1C). Directly comparing beta maps across TRs is therefore not advisable. Moreover, different TRs with constant total acquisition time would result in different degrees of freedom (due to the different number of acquired volumes) and therefore different p-values for the same t-value at the subject level, adding to the complexity of interpreting the results of direct comparisons.

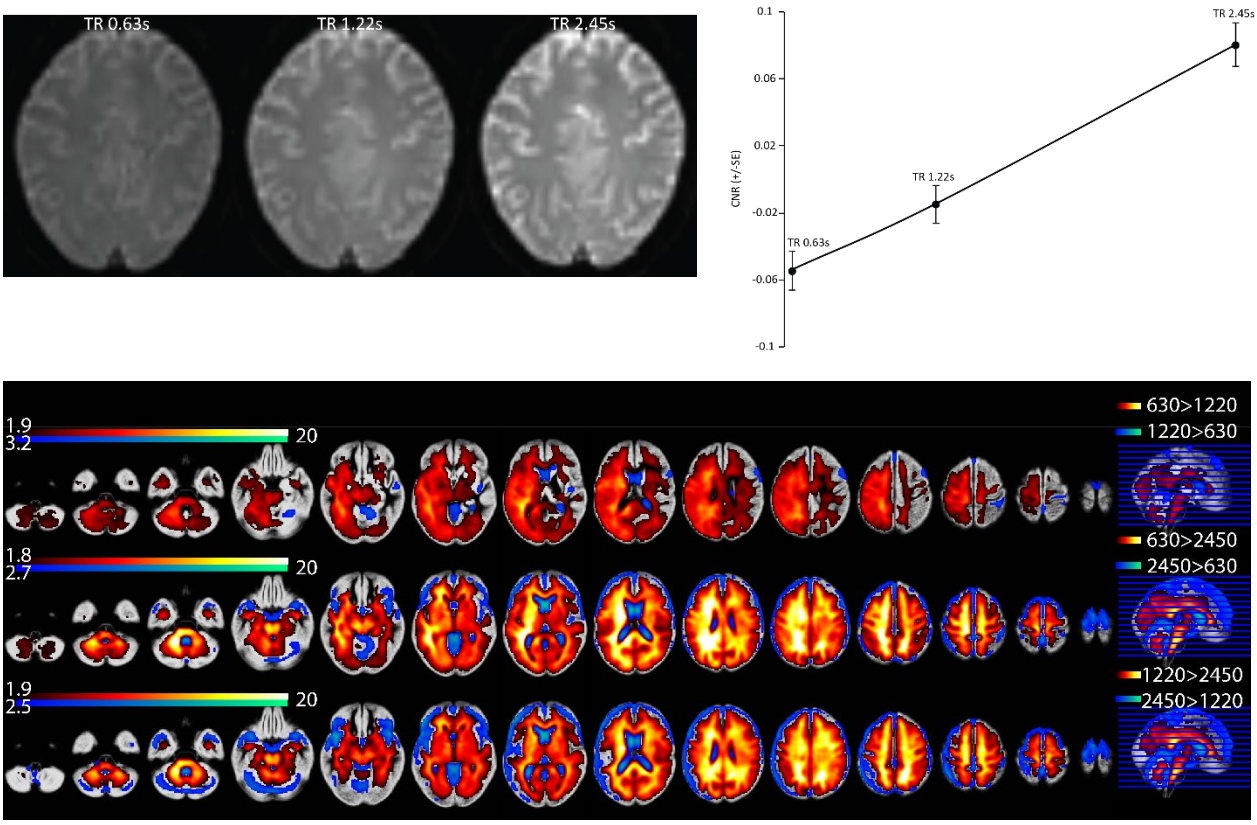

Figure 1. (A) Mean EPIs from a representative subject estimated during the realignment procedure. The images are not normalized and they correspond to z=31. **(B)** Within subject ANOVA with CNR of the white/gray matter shows a significant main effect of MB, confirming a decrease in contrast values as the TR becomes smaller. **(C)** Within-subject GLM with the beta value of the constant term shows that even after the default global normalization in SPM, the average grey matter values differ between sequences with different acceleration.

We therefore used z-values when directly contrasting different TRs in a within-subject model. Z-maps represent effect sizes in terms of a ratio between explained and unexplained variance, and are thus insensitive to scaling that affects both explained and unexplained variance similarly. To obtain z-maps, t-

maps were divided by the square root of the degrees of freedom computed by SPM after taking into account the total data points, regressors and the auto-correlation matrices. We then used a repeated measure ANOVA that compared z-values, as derived from the single subject level, across TRs for the CA-CC contrast. We found significant effects of TR in some voxels (Figure 2) in a small number of clusters, and pairwise comparisons revealed higher z-values for MB1 compared to MB4 at $q_{fdr}<0.05$ (Figure 2). There were no significant differences in the opposite direction. These results indicate that there might be no particular benefits of using MB acceleration for revealing stimulus specific activations at the subject level. This is counterintuitive in the light of the improvements of effective tSNR as well as the group level results presented in Bhandari et al., in prep.

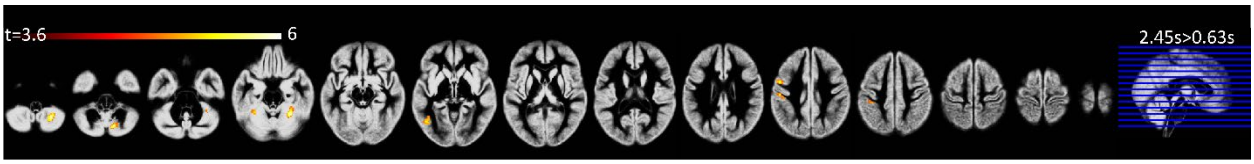

Figure 2. Pairwise contrast (for CA-CC) between the TRs show clusters with higher z-values for TR 2.45s (MB1S2) > TR 0.63s (MB4S2). Contrast not shown here did not have any significant clusters. Overlaid on the mean grey matter segment of the group. FDR q<0.05, cluster threshold 50 voxels.

To test if these results could be specific to the task-based fMRI, or are also present for resting state, we performed a pseudo-resting state analysis. Briefly, we regressed the task correlated BOLD signal from the data and used the rsn20 maps provided by Smith et al., 2009 as a predictor for a spatial GLM. This generated one time series per rsn for every TR and every subject in a dual-regression type analysis. We then performed a subject level GLM where these time series were used as predictors. The resulting t-values were then transformed to z values similarly to as explained above. A within-subject ANOVA was performed for each rsn separately. Figure 3 shows the pairwise comparisons between the TRs for one representative network (RSN 6 as described in Smith et al., 2009)). The network has significantly higher z-values for sequences with lower MB acceleration (cold colors) in the cortex compared to higher MB acceleration. We also find clusters where smaller TR show higher z-values (hot colors), but this mainly occurs in deeper brain regions, where z-values are actually negative for this component. Slower TR thus generate more extreme z-values. Within-subject analysis of the other networks also confirms this finding.

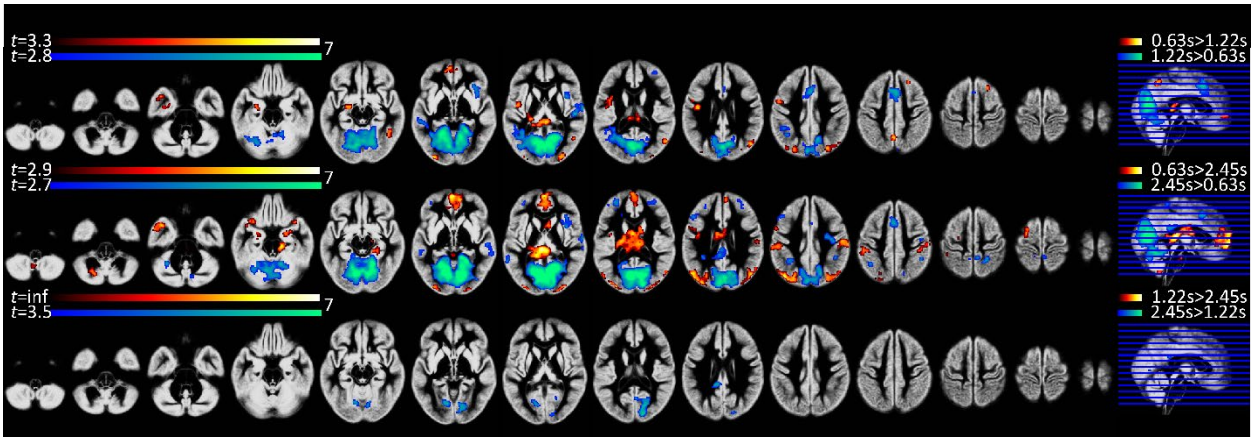

Figure 3. All significant pairwise comparisons between the TRs for RSN component 6. Significant clusters in warm colors show regions that have higher z values for a shorter TR compared to a higher TR. Clusters in cold colors show regions that are have higher z values for longer TR than shorter TR. The cold colors are regions that have positive

values in the RSN, and the clusters in warm color mostly fall in regions with negative values in RSN. Overlaid on the mean grey matter segment of the group. FDR q<0.05, cluster threshold 50 voxels.

To understand the results better, we decided to concentrate on the Action observation network. To obtain a task predictor that also has higher frequency components and is more akin to the dual regression resting state procedure, we extracted a task predictor using a dual regression pipeline but using the task based network obtained from the independent reference study as the network used in the dual regression pipeline. These findings of the within-subject ANOVA are similar to that of resting state results where the network was significantly more active for longer TRs compared to shorter TRs (Figure 4, cold colors). Pairwise comparisons show clusters that are more active for smaller TRs (Figure 4, hot colors), but these fall within voxels more active for CC than CA, and 'higher' values then mean less pronounced deactivation.

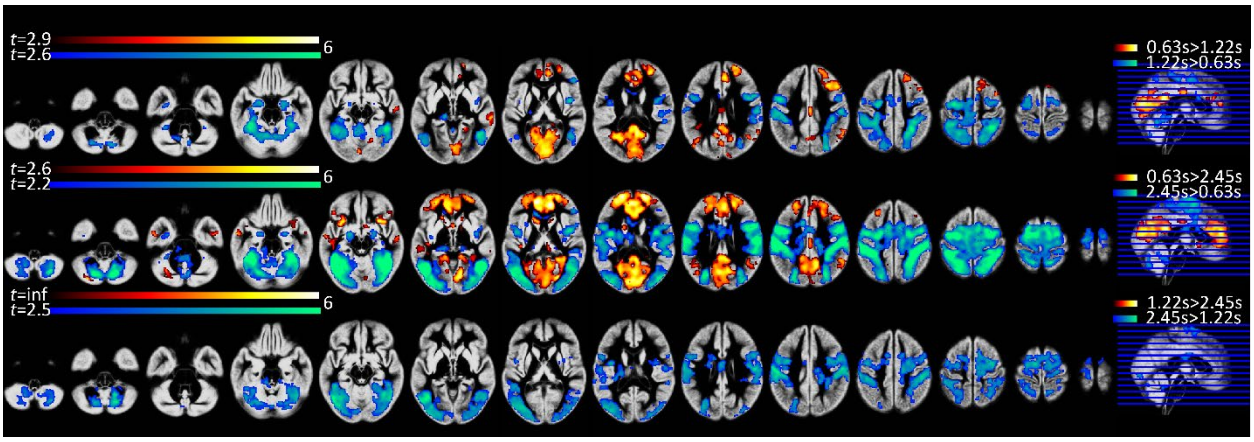

Figure 4. All significant pairwise comparisons between the TRs for the Action observation network. Significant clusters in warm colors show regions that have higher z values for a shorter TR compared to a higher TR. Clusters in cold colors show regions that are have higher z values for longer TR than shorter TR. The cold colors are regions that have positive z-values in the network, and the clusters in warm color mostly fall in regions with negative z-values in the network. Overlaid on the mean grey matter segment of the group. FDR q<0.05, cluster threshold 50 voxels.

Transforming individual-level t-statistics into z-values, and comparing them across sequences thus seem to favors the use of sequences with no MB acceleration, because they generate higher z-values. With standard GLM, resting state GLM as well as GLM with complex predictors, we consistently see that converting to z-statistics seem to penalize data from higher MB.

To validate whether the use of z-values really does make values comparable across sequences that generate different number of volumes per unit of time, we temporally down sampled the data from the multiband sequences. For MB2, we replaced every two volumes by their average. This resulted in a dataset with 200 images instead of 400 images. Similarly, we average every 4 images into a single image for TR 0.63 (MB4 S2) (the average of volumes 1, 2, 3 and 4 become volume 1; that of 5, 6, 7, and 8, volume 2 etc.) resulting in a dataset with 195 instead of 780 images. Predictors and all other regressors were averaged in the same way and a subject level GLM was performed on these downsampled dataset. The t-statics were then converted to z-statistics using the degrees of freedom for both the original and downsampled analysis. A within subject GLM for TR 1.22s (full vs average) showed that for the action observation network, the averaged dataset compared to the full dataset showed *higher* z-values. The only areas where the full dataset had higher z-values was in the regions with negative value (Figure5, row1).

This was also the case when we looked at the GLM with the TR 0.63s (Figure5, row2). This suggests that indeed the penalty of z-conversion might be too high as decreasing the statistical power by decreasing the sample size to half or one-fourth still leads to better results than the full dataset.

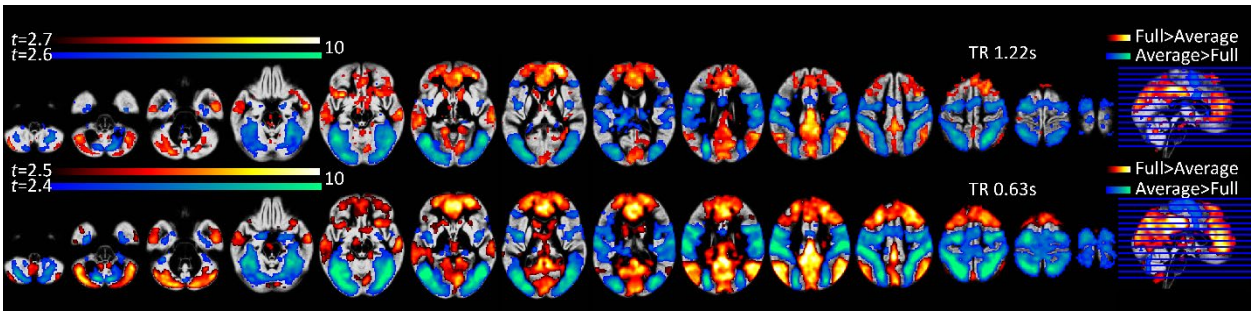

Figure 5. Functional scans were subsampled by averaging of every 2 to get 2 times less (MB2S2) and every 4 to get 4 times less (MB4S2) samples. Within-subject ANOVA with the full sample and the average sample showed that the impoverished (downsampled) data performed better than the original sample in the network of interest. Regions where the full sample shows better results correspond to the negative values in the network and signifies that the values in the average are more negative than the full sample suggesting that the average performs better for the negative values as well.

Discussion: Using the parameter estimates (Beta and con) for direct voxel-wise comparison of sequences with and without MB acceleration is not recommended, as the CNR is higher with longer TR leading to TR dependent scaling. This is due to the fact that global mean scaling, which is a default setting in SPM to bring the time series to a common scale creates systematic biases when the CNR differs between conditions, and therefore makes the interpretation of results tricky. Furthermore, t-statistics cannot be compared directly either, as for the same amount of acquisition time, many more samples are acquired for MB accelerated sequences, which results in higher degrees of freedom and therefore higher t-values, biasing analyses towards MB. Z-transformation of the t-statistics appears to be the logical way to enable a direct comparison of two sequences with different TRs (with different numbers of acquired voxels). A common way of transforming t-statistics is using the normal inverse of the student's t cumulative density. This is also the default transformation that is used by FSL (https://fsl.fmrib.ox.ac.uk). However, this transformation is not adept for transforming t-statistics with different degrees of freedom. Therefore to render the data from different MB accelerated sequences comparable, the standardization of the t-values need to be done by dividing the t-values by the square root of the degrees of freedom. Directly comparing different MB sequences using a within-subject ANOVA after this transformation, shows that sequences without MB show better statistics, for both task based and resting state analysis. This is in contrast to the previous literature, all of which points to benefits of using MB acceleration, including the group analyses of the same data we performed. To understand this discrepancy, we down sampled the MB data by averaging the scans. This resulted in a lower DOF and therefore the t-values were divided by a smaller value than the original data, for performing z-transformation. Comparing the averaged data to the original data showed that the performance of the average data exceeds the performance of the full data suggesting a high penalty posed by the z-transformation. This finding is counterintuitive, as it would suggest that impoverishing the data by subsampling would increase the ability to detect the measured phenomenon. Instead, this finding suggests that the estimation of DOF performed by SPM at the single subject may be inflated for multiband data, thereby over penalizing the z-transform for high MB.

Many of the previous studies that have assessed the effects of MB on task related statistics used summary t- or z-statistics (using the inverse of t-cumulative density) using repeated measure designs to evaluate which sequence work better (Demetriou et al., 2018; Kiss et al., 2018; Sahib et al., 2016; Todd et al., 2016). As indicated by our findings, these findings are biased towards MB and one need to be careful while interpreting these results. On the other hand, performing group-level statistics shows that the t-values detected while scanning with MB are higher than without MB. However, it offers no direct evidence for the improvement of single-subject level statistics.

In summary, with the current state of literature, we have no empirical evidence that MB acceleration improves task-based statistics at a single subject level. Moreover, while individual studies indicate that group level statistics may improve when acquired with MB acceleration, only meta-analytical results may confirm the actual benefits of MB acceleration. Furthermore, alternative comparison strategies such as Bayesian modeling which has been recently proposed as a better approach for comparing difference sequences must be explored (Zeidman et al., 2019). Finally, we appeal to the Multiband community to exercise caution while interpreting and reporting results assessing the effects of MB acceleration on task related fMRI.

Acknowledgements: This work was supported by the Netherlands Organization for Scientific Research (VIDI: 452-14-015 to V.G.), the Brain and Behavior Research Foundation (NARSAD young investigator 22453 to V.G.), the European Research Council of the European Commission (ERC-StG-312511 to C.K.) and the BIAL foundation grant (503 323 055 to V.G., C.K. & R.B.). We thank Spinoza centre for neuroimaging where the scanning was performed and the staff members of Spinoza centre.

Conflict of Interest: The authors report that there is no conflict of interest.

References:

Boubela, R. N., Kalcher, K., Nasel, C., & Moser, E. (2014). Scanning fast and slow: current limitations of 3 Tesla functional MRI and future potential. *Frontiers in Physics*, *2*, 1. https://doi.org/10.3389/fphy.2014.00001

Boyacioğlu, R., Schulz, J., Koopmans, P. J., Barth, M., & Norris, D. G. (2015). Improved sensitivity and specificity for resting state and task fMRI with multiband multi-echo EPI compared to multi-echo EPI at 7 T. *NeuroImage*, *119*, 352–361. https://doi.org/http://dx.doi.org/10.1016/j.neuroimage.2015.06.089

Demetriou, L., Kowalczyk, O. S., Tyson, G., Bello, T., Newbould, R. D., & Wall, M. B. (2018). A comprehensive evaluation of increasing temporal resolution with multiband-accelerated protocols and effects on statistical outcome measures in fMRI. *NeuroImage*, *176*, 404–416. https://doi.org/https://doi.org/10.1016/j.neuroimage.2018.05.011

Feinberg, D. A., Moeller, S., Smith, S. M., Auerbach, E., Ramanna, S., Glasser, M. F., … Yacoub, E. (2010). Multiplexed Echo Planar Imaging for Sub-Second Whole Brain FMRI and Fast Diffusion Imaging. *PLoS ONE*, *5*(12), e15710. https://doi.org/10.1371/journal.pone.0015710

Golestani, A. M., Faraji-Dana, Z., Kayvanrad, M., Setsompop, K., Graham, S. J., & Chen, J. J. (2017). Simultaneous Multislice Resting-State Functional Magnetic Resonance Imaging at 3 Tesla: Slice-Acceleration-Related Biases in Physiological Effects. *Brain Connectivity*, *8*(2), 82–93. https://doi.org/10.1089/brain.2017.0491

Griffanti, L., Salimi-Khorshidi, G., Beckmann, C. F., Auerbach, E. J., Douaud, G., Sexton, C. E., … Smith, S. M. (2014). ICA-based artefact and accelerated fMRI acquisition for improved Resting State Network imaging. *NeuroImage*, *95*, 232–247. https://doi.org/10.1016/j.neuroimage.2014.03.034

Kiss, M., Hermann, P., Vidnyánszky, Z., & Gál, V. (2018). Reducing task-based fMRI scanning time using simultaneous multislice echo planar imaging. *Neuroradiology*, *60*(3), 293–302. https://doi.org/10.1007/s00234-017-1962-4

McDowell, A. R., & Carmichael, D. W. (2018). Optimal repetition time reduction for single subject event-related functional magnetic resonance imaging. *Magnetic Resonance in Medicine*, *0*(0). https://doi.org/10.1002/mrm.27498

Preibisch, C., Castrillón G., J. G., Bührer, M., & Riedl, V. (2015). Evaluation of Multiband EPI Acquisitions for Resting State fMRI. *PLoS ONE*, *10*(9), e0136961. https://doi.org/10.1371/journal.pone.0136961

Sahib, A. K., Erb, M., Marquetand, J., Martin, P., Elshahabi, A., Klamer, S., … Focke, N. K. (2018). Evaluating the impact of fast-fMRI on dynamic functional connectivity in an event-based paradigm. *PLoS ONE*, *13*(1), e0190480. https://doi.org/10.1371/journal.pone.0190480

Sahib, A. K., Mathiak, K., Erb, M., Elshahabi, A., Klamer, S., Scheffler, K., … Ethofer, T. (2016). Effect of temporal resolution and serial autocorrelations in event-related functional MRI. *Magnetic Resonance in Medicine*, *76*(6), 1805–1813. https://doi.org/10.1002/mrm.26073

Smitha, K. A., Arun, K. M., Rajesh, P. G., Joel, S. E., Venkatesan, R., Thomas, B., & Kesavadas, C. (2018). Multiband fMRI as a plausible, time-saving technique for resting-state data acquisition: Study on functional connectivity mapping using graph theoretical measures. *Magnetic Resonance Imaging*, *53*, 1–6. https://doi.org/https://doi.org/10.1016/j.mri.2018.06.013

Todd, N., Moeller, S., Auerbach, E. J., Yacoub, E., Flandin, G., & Weiskopf, N. (2016). Evaluation of 2D multiband EPI imaging for high-resolution, whole-brain, task-based fMRI studies at 3T: Sensitivity and slice leakage artifacts. *Neuroimage*, *124*(Pt A), 32–42. https://doi.org/10.1016/j.neuroimage.2015.08.056

Zeidman, P., Kazan, S. M., Todd, N., Weiskopf, N., Friston, K. J., & Callaghan, M. F. (2019). Optimizing Data for Modeling Neuronal Responses. *Frontiers in Neuroscience*, *12*, 986. https://doi.org/10.3389/fnins.2018.00986