Title: Improving effect size and power with multi-echo fMRI and its impact on understanding the neural systems supporting mentalizing

Authors: Michael V. Lombardo[1,2,3]*, Bonnie Auyeung[3,4], Rosie Holt[3], Jack Waldman[3], Amber Ruigrok[3], Natasha Mooney[3], Edward T. Bullmore[5], Simon Baron-Cohen[3], & Prantik Kundu[6]*

Affiliations:

1 Department of Psychology, University of Cyprus, Cyprus
2 Center for Applied Neuroscience, University of Cyprus, Cyprus
3 Autism Research Centre, Department of Psychiatry, University of Cambridge, UK
4 Department of Psychology, School of Philosophy, Psychology, and Language Sciences, University of Edinburgh, UK
5 Brain Mapping Unit, Department of Psychiatry, University of Cambridge, UK
6 Section on Advanced Functional Neuroimaging, Departments of Radiology & Psychiatry, Icahn School of Medicine at Mount Sinai, USA

Corresponding Authors: Michael V. Lombardo (mvlombardo@gmail.com) and Prantik Kundu (prantik.kundu@mssm.edu)

Abstract

Functional magnetic resonance imaging (fMRI) research is routinely criticized for being underpowered due to characteristically small sample sizes. Additionally, fMRI signals inherently possess various sources of non-BOLD noise that further hampers ability to detect subtle effects. Here we demonstrate that multi-echo fMRI data acquisition and denoising can increase effect size and statistical power for block-design experiments, allowing for novel insights by detecting effects that are typically obscured in small sample size/underpowered studies. Application of this method on two different tasks within the social cognitive domain of mentalizing/theory of mind demonstrates that effect sizes are enhanced at a median rate of 25-32% in regions canonically associated with mentalizing. For non-canonical cerebellar areas that have been largely less focused on by the field, effect sizes boosts were much more substantial in the range of 43-108%. These cerebellar areas are highly functionally connected at rest with neural systems typically associated with mentalizing and the resting state connectivity maps largely recapitulate the topology observed in activation maps for mentalizing. Power simulations show that boosts in effect size enable ability to conduct high-powered studies at traditional sample sizes. However, cerebellar effects will remain underpowered at traditional sample sizes and without the multi-echo innovations we describe here. Thus, adoption of multi-echo fMRI innovations can help address key criticisms regarding statistical power and non-BOLD noise and enable potential for novel discovery of aspects of brain organization that are currently under-appreciated and not well understood.

Introduction

A common criticism of neuroscience research in general[1] and functional MRI (fMRI) in particular[2], is that studies are characteristically statistically underpowered. Low statistical power by definition means that a study will have less of a chance for detecting true effects, but also means that observed statistically significant effects are less likely to be real, and that such effects will be more susceptible to the biasing impact of questionable research practices[1,3]. This problem is important given the emergent 'crisis of confidence' across many domains of science (e.g., psychology, neuroscience), stemming from low frequency of replication and the pervasive nature of questionable research practices[1,3,4].

Low statistical power can be attributed to small sample sizes, small effect sizes, or a combination of both. The general recommended solution to this problem is to increase sample size, increase scan time within-subjects, or both. These recommendations are pragmatic mainly because these are variables that are completely within the control of the researcher when planning a study. While these recommendations are important to consider[2,5-8], other considerations such as dealing with substantial sources of non-BOLD noise inherent in fMRI data, also need to be evaluated before the field assumes increasing sample size or scan time to be the primary or only means of increasing statistical power. These considerations are especially poignant when mandates for large-N studies and increased within-subject scan time are practically limiting due to often cited reasons such as the prohibitively high costs for all but the most well-funded research groups or in situations where the focus is on studying sensitive, rare, and/or less prevalent patient populations and where increasing scan time is impractical (e.g., children, neurological patients).

On the issue of non-BOLD noise variability, it is well known that fMRI data are of variable quality. Indeed, fMRI data can be quite poor and of variable quality and this can significantly hamper ability to achieve accurate and reproducible representations of brain organization. It is widely understood that the poor sensitivity of fMRI often arises from high levels of subject motion (often task correlated), cardiopulmonary physiology, or other types of imaging artifact[9]. It is key to underscore that these artifacts are problematic because they are often inadequately separable from the functional blood oxygenation level dependent (BOLD) signal when using conventional fMRI methods. Given an advance in fMRI methodology that allows enhanced detection and removal of these artifacts, the situation regarding statistical power and sample size may change markedly. Such advances could create viable experimental alternatives or supplements to the recommendation for increasing sample size/scan time to boost statistical power, and concurrently make for a situation that can more reliably enable discovery of subtle but potentially key aspects of typical and atypical brain function.

In this study, we seek to make the situation of statistical power and sample size in fMRI research more favorable by revisiting the methodological implementation of task-based fMRI around the precise removal of non-BOLD artifact. We have applied a new approach that integrates the fMRI data acquisition innovation of multi-echo EPI with the decomposition method of independent components analysis (ICA), towards principled removal of non-BOLD signals from fMRI data. Our fully integrated implementation is called multi-echo independent components analysis or ME-ICA[10]. ME-ICA utilizes multi-echo fMRI to acquire both fMRI signal time series *and their NMR signal decay,* towards distinguishing functional BOLD from non-BOLD signal components based on their respective and differentiable signatures in the decay domain. Critically, BOLD and non-BOLD signal domains are readily differentiable in data analysis of the echo time (TE) domain - irrespective of overlap of signal patterns in the spatial and temporal domains. BOLD-related signals specifically show linear dependence of amplitude on TE, whereas non-BOLD signal amplitudes demonstrate TE-independence.

In ME-ICA, a multi-echo (ME) specific high-dimensional ICA is applied to ME-fMRI datasets, producing a set of signal components explaining high proportions of total dataset variance (85-98% depending on imaging parameters). Component-level metrics of the "physical" measures of TE-dependence and TE-independence are computed for each component and used to determine functional BOLD vs. non-BOLD origin. The non-BOLD components are then removed from the data, after which in task-related fMRI, statistical modeling of task-related effects can proceed as they typically would in any other study. It is important to underscore here that ME-ICA leverages additional information from multiple readouts of T2* signal at various echo times and isolates and removes non-BOLD signal in a manner that is completely blind to task design. Therefore, ME-ICA acts as a principled bottom-up denoising tool that does not need information about task design and stands in contrast to other denoising methods whereby task-design information is necessary and utilized (e.g., GLMdenoise[11]) or in circumstances where ICA is employed but cannot take advantage of multiple echo readouts and model BOLD-related and non-BOLD characteristics of the data (e.g., FIX[12]).

In past work we have applied ME-ICA to problems of non-BOLD noise and seed-based connectivity estimation within resting state fMRI data[10,13-16]. However, this study is the first to apply ME-ICA to the problem of task-related fMRI studies and in particular, the study will demonstrate the performance of ME-ICA in improving effect size and power. As proof-of-principle ME-ICA is applied in two paradigms that tap the social-cognitive function/domain of mentalizing and theory of mind. Theoretically, because ME-ICA is a powerful denoising method and rooted in principled biophysical differentiation of BOLD and non-BOLD sources of variation, we expect there to be substantial improvements in effect size estimated from task-based fMRI paradigms. This study aims to evaluate the magnitude of such ME-ICA-related improvements, specifically for second-level random-effects group analyses employing one-sample t-tests. We will evaluate any such improvements from brain areas typically considered 'canonical' for the

domain of mentalizing and also from areas that are 'non-canonical' for that domain. Finally, we characterize many practical implications of our effect size estimation for conducting power simulations for future study planning (e.g., cost of acquiring sample sizes for achieving sufficient power).

To evaluate the specific value of ME-ICA, effect size and statistical power are compared between data that employ ME-ICA denoising versus T2*-weighted optimal combination of echoes (TSOC) and denoising via the general linear model (GLM) with motion parameter time courses regressed out (TSOC+MotReg). This is an ideal comparison because the two datasets are identical in preprocessing pipelines and both use T2*-weighted optimally combined data across echo times (i.e. TSOC) and only differ in the final step of the ME-ICA pipeline, where it diverges by additionally running ME-ICA specific processing such as ICA and TE-dependence analysis for the purpose of denoising. It is notable that the TSOC data is at least a fair comparison to conventional single-echo fMRI acquisition without acceleration, and is likely superior with respect to decreased thermal noise and mitigated signal dropout in high-susceptibility areas such as orbitofrontal cortex and inferior temporal regions. Therefore, our comparison here of ME-ICA to TSOC+MotReg is a much more conservative comparison of the benefits of ME-ICA, relative to comparing ME-ICA to conventional single-echo EPI acquisition.

In the following analyses we examine two different mentalizing tasks (i.e. the 'SelfOther' and 'Stories' tasks; see Methods for description) and compare effect sizes across ME-ICA and TSOC+MotReg pipelines for specific regions of interest that are present in the NeuroSynth[17] meta-analysis for the feature 'mentalizing' (Fig 2A & 3A). Regions of interest are split into two classes. The first class comprises 'canonical' regions: those that are regularly identified and heavily focused on as important in the literature[18-24]. The second class of regions comprises what we call 'non-canonical' regions: mainly localized in the cerebellum. These regions are largely less focused on by the field, although some recent meta-analytic evidence has argued for their importance[25,26] and here we show using NeuroSynth meta-analyses that these regions are also identified. For all regions we estimate effect size and also conduct power simulations to inform the effects that ME-ICA would have on future study planning for achieving 80% power, as well as demonstrating any beneficial effects of ME-ICA in terms of monetary savings over and above TSOC+MotReg.

Results

ME-ICA Effects on the Raw Time Series

Before touching on quantitative comparisons of effect size and power due to ME-ICA, it is helpful to convey properties of the images and time series acquired with ME acquisition, as well as the effect on the time series from ME-ICA denoising. ME sequences capture the decay of EPI images and (time series) with increasing TE, shown in Fig 1A. ME data poignantly highlight the problem of susceptibility artifact (i.e. signal

dropout) in areas such as ventromedial prefrontal cortex (vMPFC) - it is made clear from Fig 1A that signal dropout occurs at longer TEs, as affected regions have short T2* due to proximity to air-tissue boundaries. Additionally, gray/white signal contrast increases over longer TE due to T2* differences between these tissue types. The T2*-weighted optimal combination (TSOC) implements a matched-filter of TE images yielding a new image time series with optimized contrast (TE~T2*) and mitigation of susceptibility artifact by weighting towards the early TE in areas with short T2*. In Fig 1B we present time series from vMPFC, posterior cingulate cortex/precuneus (PCC), and right cerebellum, demonstrating the effect of optimal combination on time series, and then the effect of removing non-BOLD noise using ME-ICA in exposing block-design activation. Each regions depicts the time series from the middle echo around 31ms, which is characteristic of most single-echo EPI studies conducted using 3T MRI. In addition we show the time series from TSOC data, ME-ICA isolated BOLD signals, and non-BOLD signals removed from data. Overlaid on all comparisons are the blocked time courses from the task. It is particularly apparent that ME-ICA recovers task-based block fluctuations while much of the middle echo, TSOC, and non-BOLD isolated signals carrying complex artifacts including drifts, step changes, and spikes.

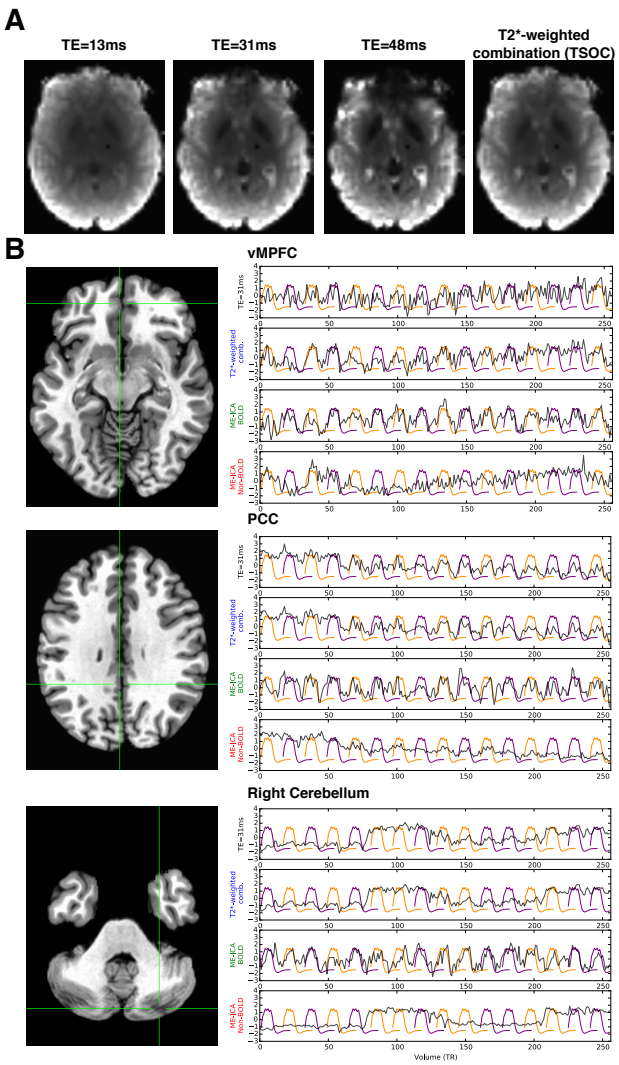

Fig 1: Multi-Echo Signal Characterization. Panel A shows the signal decay captured in multi-echo EPI images, for a single representative volume. With longer TE, gray/white contrast increases. Susceptibility artifact (e.g. dropout) also increases, as these regions have short T2* due to proximity near air-tissue boundaries. The T2*-weighted optimal combination (TSOC) implements a matched-filter of TE images yielding a new image with optimized gray/white contrast and mitigation of susceptibility artifact. Panel B shows comparisons of time courses across three regions of interest in mentalizing: ventromedial prefrontal cortex (vMPFC), posterior cingulate cortex/precuneus (PCC), and right cerebellum. Each comparison shows time courses (before model-based filtering) of middle TE data (black), combined data (blue), BOLD signals isolated on the basis of TE-dependence (green),

and non-BOLD signals removed from the data (red). Purple and orange lines represent modeled mentalizing and physical blocks respectively.

ME-ICA Boosts Effect Size Estimation in Canonical Mentalizing Regions

In the SelfOther task, we found that of the 'canonical' regions for mentalizing, all but RTPJ show substantial ME-ICA-related boosts in effect size compared to using TSOC+MotReg data. Effect size boosts ranged from as small as around 14.42% for an area like LTPJ to as large as a 111.93% and 76.73% increases for traditionally harder-to-image regions like vMPFC and temporal pole. The median ME-ICA effect size percentage increase was 32.93% across all canonical mentalizing regions (Supp Table 1). Since the estimated effect sizes here are point estimates of the true effect size in the population, we ran bootstrap resampling to determine 95% confidence intervals around effect size estimates and effect size boosts (see error bars in the effect size and effect size % increase bars graphs; Fig 2B and 2C) and we have also quantified the percentage of bootstrap resamples where the ME-ICA-related boost was greater than 0. Nearly all regions showed ME-ICA effect size increases on 98-100% of the 1000 bootstrap resamples (Fig 2B).

In the Stories task, we found largely similar results as the SelfOther task, but with some subtle differences. Quantitatively the effect size percentage increase due to ME-ICA ranged from as small as around 13.46% for an area like LTPJ to as large as a 67.07% increases for a traditionally harder-to-image region like temporal pole. The median ME-ICA effect size percentage increase was 25.64% across all canonical mentalizing regions (Supp Table 1). With bootstrapping we also found that such boosts were fairly robust. All 'canonical' regions except for vMPFC, showed effect size boosts from ME-ICA in more than 92% of the 1000 bootstrap resamples (Fig 2C). It is important to point out one subtlety that in this task, there is less sensitivity for effects in vMPFC compared to the SelfOther task, but more sensitivity for effects in areas like RTPJ (e.g., large effect size).

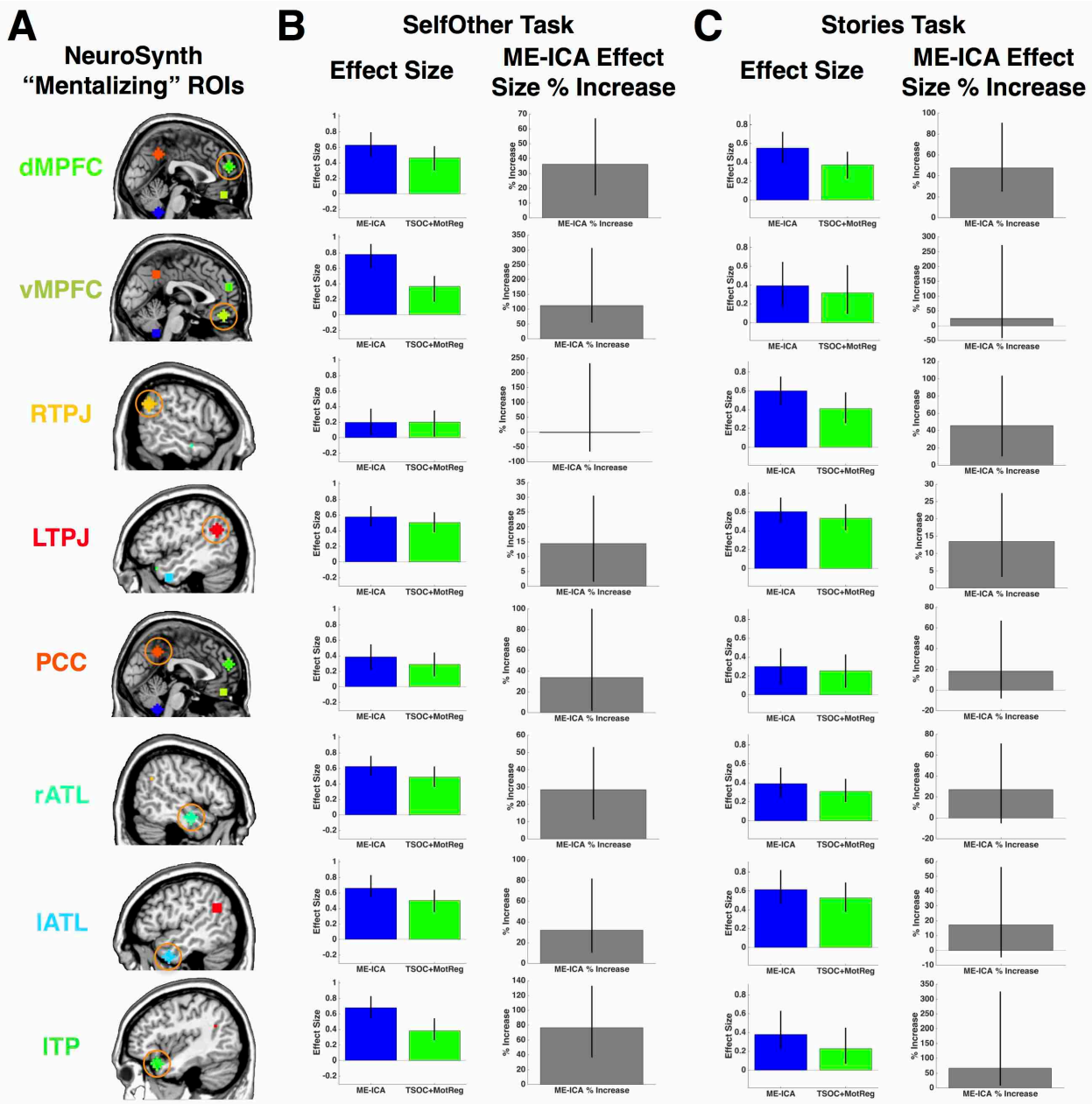

Fig 2: Effect size in canonical mentalizing regions. This figure shows effect size estimates and ME-ICA effect size percentage increases from each canonical mentalizing region extracted from 8mm spheres around the peaks in the NeuroSynth 'mentalizing' map (A) when using ME-ICA (blue) or TSOC+MotReg data (green). Panel B shows results from the SelfOther task and panel C shows results from the Stories task. All error bars are 95% bootstrap confidence intervals. Effect sizes are expressed in standard deviation units and are analogous to Cohen's d.

ME-ICA Boosts Effect Size Estimation in Cerebellar Regions

In contrast to 'canonical' regions, we also examined 'non-canonical' areas in the cerebellum that are relatively neglected in the literature on mentalizing and theory of mind, but which appear in prior meta-analyses[25] and are also present in NeuroSynth (Fig 3A). One reason for examining these non-canonical cerebellar areas is because ME-ICA potentially has the ability to reveal new effects of importance, via removing non-BOLD noise variability. If such effects occurred in the context of mentalizing, some initial good candidates would be cerebellar regions that van Overwalle and colleagues have recently highlighted might be important from meta-analytic inference[25]. Thus, in these datasets we have the opportunity to investigate the hypothesis that one reason cerebellar regions are not heavily focused on may be due to typically low effect size (and by extension, low statistical power) when ME-ICA is not employed.

For the SelfOther task, we find that ME-ICA allows for very prominent cerebellar effect size boosts from 51.66% to 108.19% increases, and within all regions, more than 94% of the 1000 bootstrap resamples showed a greater than 0 boost from ME-ICA (Fig 3B). In practical terms, effect sizes in TSOC+MotReg for each of these regions were small (e.g., 0.25 to 0.35), but after ME-ICA, the effect sizes were in a range that are characteristic of most canonical regions (e.g., 0.38 to 0.73). Cerebellar regions showed similar effects in the Stories task. That is, right and left hemisphere cerebellum (but not medial cerebellum) showed ME-ICA effect sizes increases of 73.96% and 43.90% respectively, and such positive effect size increases were observed on 99-100% of the 1000 bootstrap resamples (Fig 3C). In practical terms, right and left cerebellar regions possessed effect sizes after ME-ICA that were considerable (e.g., 0.43 and 0.52) and similar to effects observed in 'canonical regions, whereas in TSOC+MotReg the effect sizes were small (e.g., 0.24 to 0.36). In addition to these insights in targeted meta-analytically defined cerebellar ROIs, the robust presence of cerebellar activations are also apparent in whole-brain analyses across both tasks (Fig 4) (Supp Table 2).

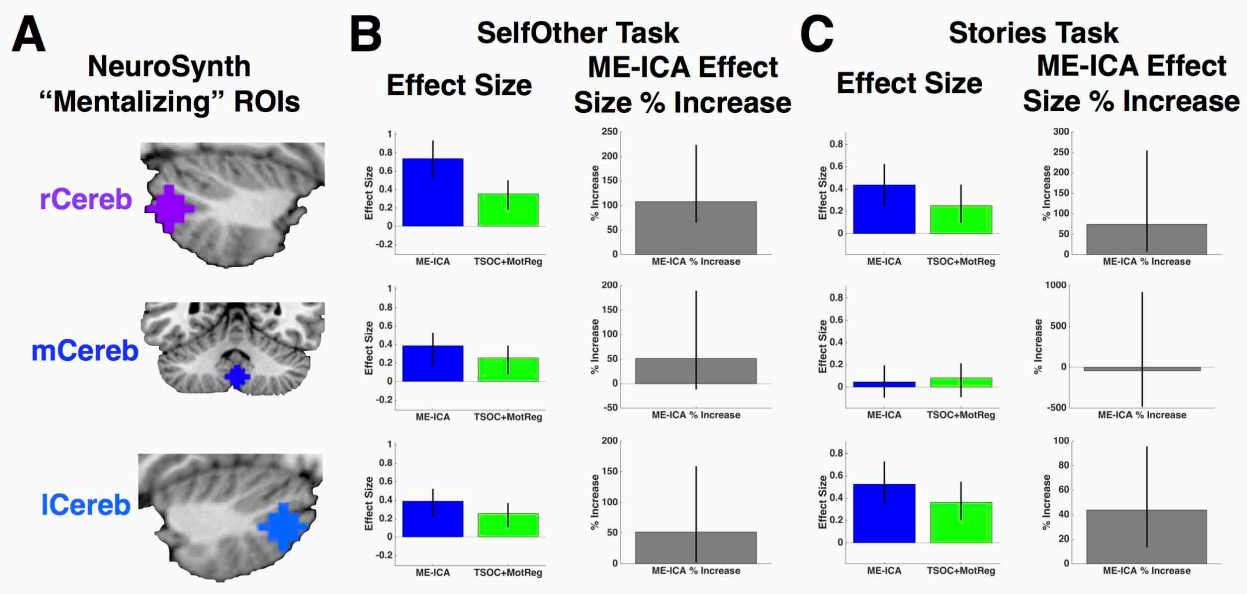

Fig 3: Effect size in cerebellar mentalizing regions. This figure shows effect size estimates and ME-ICA effect size percentage increases from each cerebellar mentalizing region extracted from 8mm spheres around the peaks in the NeuroSynth 'mentalizing' map (A) when using ME-ICA (blue) or TSOC+MotReg data (green). Panel B shows results from the SelfOther task and panel C shows results from the Stories task. All error bars are 95% bootstrap confidence intervals. The effect sizes are expressed in standard deviation units and are analogous to Cohen's d.

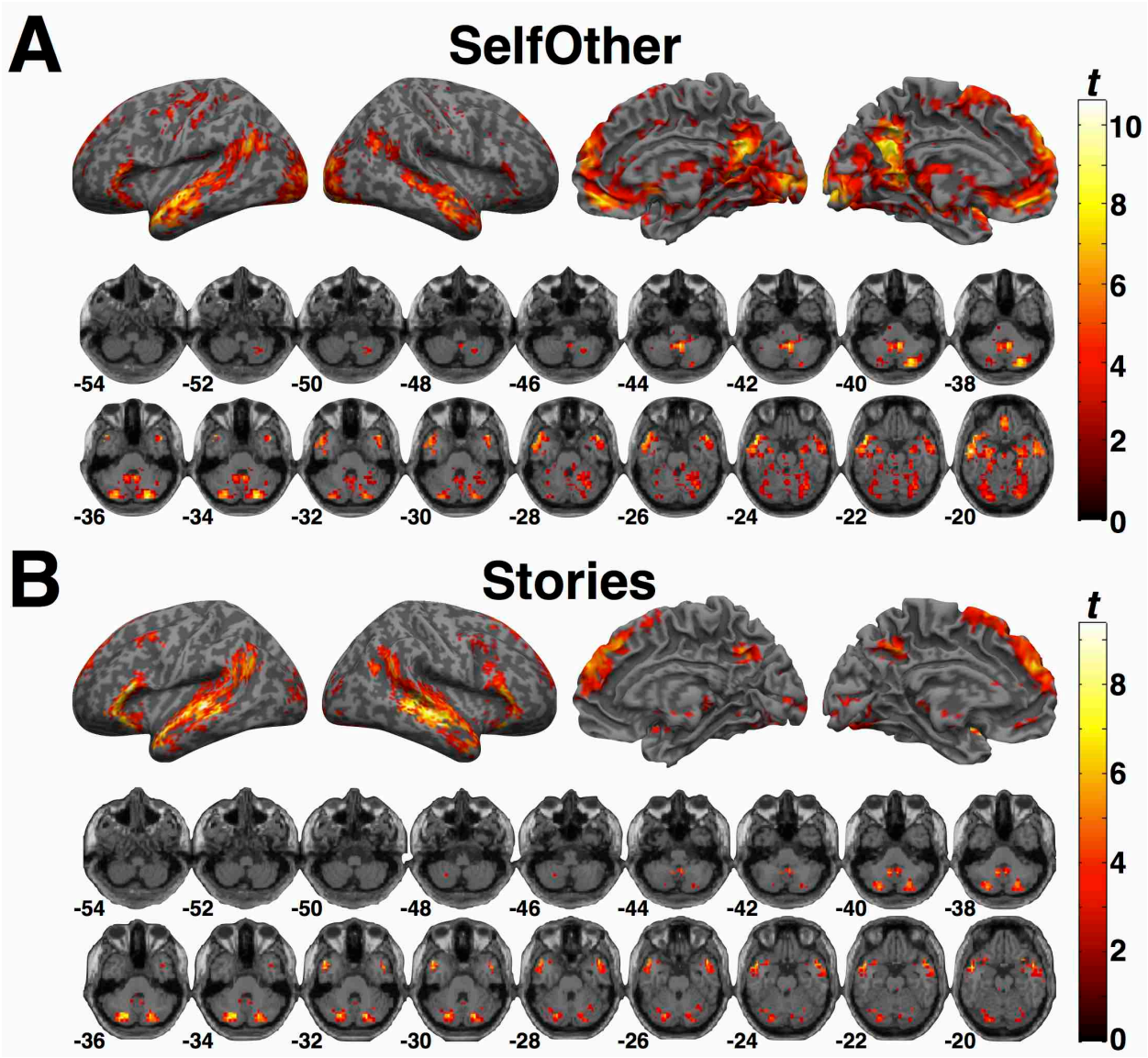

Fig 4: Cortical surface rendering and cerebellar montage of whole-brain activation for Mentalizing>Physical using ME-ICA. This figure shows a cortical surface rendering and cerebellar montage of whole-brain activations for the SelfOther task (A) and Stories task (B) using ME-ICA. All results are shown thresholded at a voxel-wise FDR q<0.05.

Connectivity Evidence for Cerebellar Involvement in Neural Systems Supporting Mentalizing

The improvements in effect size estimation particularly for cerebellar contributions is important as it potentially signals the ability of ME-ICA to uncover novel effects that may have been undetected in previous research. To further test the

importance of cerebellar contributions to mentalizing, we have examined resting state functional connectivity data and the relationship that cerebellar connectivity patterns may have with task-evoked mentalizing systems. Prior work suggests that specific cerebellar regions may be integral participants with the default mode network[27]. The default mode network incorporates many of the regions that are highly characteristic in task-evoked systems supporting mentalizing[28]. Meta-analytically defined cerebellar regions associated with mentalizing show some overlap with these cerebellar default mode areas[26]. Therefore, if cerebellar regions for which ME-ICA systematically produces boosts in effect size are integral participants in neural circuits associated with mentalizing, we hypothesized that resting state connectivity patterns with such cerebellar regions would be highly involved in the default mode network. Taking this hypothesis one step further, we also hypothesized that if these cerebellar nodes are truly important within the neural systems that support mentalizing, we should expect that cerebellar resting state functional connectivity patterns highlighted with multi-echo EPI methods would recapitulate the patterns observed for activational topology observed during mentalizing tasks across the whole-brain and within the same participants.

Confirming these hypotheses we find that bilateral cerebellar seeds involved in mentalizing show highly robust resting state functional connectivity patterns that resemble the default mode network within the same participants who participated in our task paradigms. Visually, the similarity between the ME-ICR connectivity maps and our Mentalizing>Physical activation maps are striking (Fig. 5A). Quantitatively we assessed this similarity through voxel-wise correlations (estimated with robust regression) across the whole-brain, and here we confirm that the resting state functional connectivity maps are strikingly similar in patterning to what we observe for task-evoked mentalizing activation patterns (all $r > 0.37$) (Fig. 5B). Relative to the activation-connectivity similarity observed in TSOC+MotReg data, the activation-connectivity similarity obtained with ME-ICA and ME-ICR is much larger (i.e. $z > 8.85$) (Fig 5B-5D).

It is worth noting that functional connectivity maps within conventional functional connectivity analysis on TSOC data were massively right-shifted compared to distributions in ME-ICR that centered approximately around 0 with some positive shift (Supp Fig 1). Because of the massive numbers of large positive correlations, thresholding group analyses at the same t-statistic threshold as when using ME-ICR with a FDR $q<0.05$, results in a map that shows nearly all of the brain positively functionally connected to cerebellar seeds (Fig 5C). This effect demonstrates a point already shown in prior work on ME-ICR[13], that conventional seed-based functional connectivity analyses are not properly statistically conditioned to control the effective degrees of freedom in the dataset and thus result in many more false positive associations. Here, we can see that this inflation in functional connectivity estimation in conventional seed-based analyses is a likely explanation for the reduced pattern similarity between mentalizing activation maps and functional connectivity from the cerebellum.

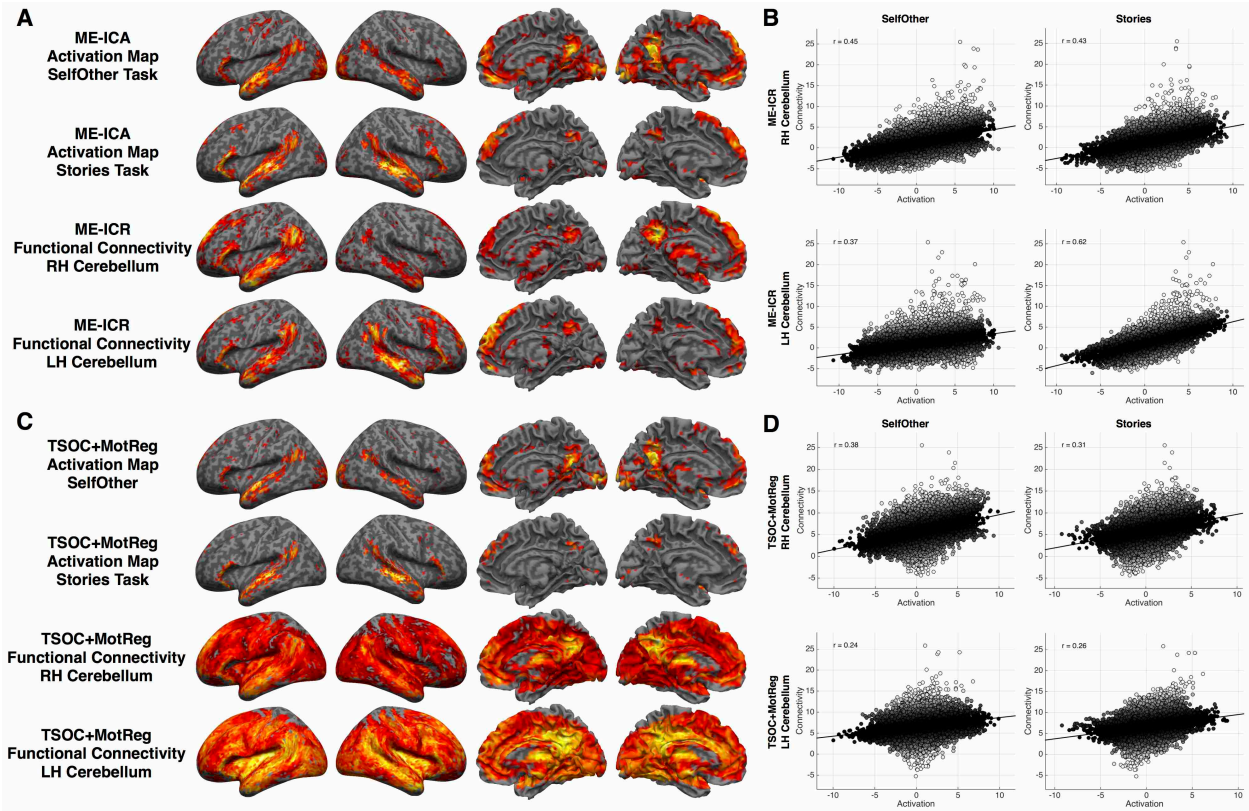

Fig 5: Resting state functional connectivity from cerebellar seed regions and pattern similarity with Mentalizing>Physical activation maps. This figure shows resting state connectivity from right and left cerebellar seed voxels (i.e. peak voxels from the NeuroSynth 'mentalizing' map) and their similarity to Mentalizing>Physical activation maps. Panel A shows activation and resting state functional connectivity maps when using ME-ICA and multi-echo independent components regression (ME-ICR[13]). All data are visualized at thresholded of voxelwise FDR q<0.05. Panel B shows scatterplots and robust regression correlations between whole-brain activation and connectivity patterns when using ME-ICA and ME-ICR. Robust regression was used to calculate the correlation in a way that is insensitive to the outliers in the connectivity map which are voxels that are proximally close to the seed region. Panel C shows activation and cerebellar functional connectivity maps for data when using conventional analysis approaches on TSOC data. Activation maps are thresholded at FDR q<0.05. Connectivity maps are thresholded at the same t-statistic threshold for defining FDR q<0.05 in ME-ICR analyses (which were already much higher than the FDR q<0.05 cutoff estimated from TSOC data), and were shown in this manner to show connectivity at the exact same t-threshold cutoff. Panel D shows activation and connectivity similarity estimated with robust regression in TSOC data.

Impact of ME-ICA on Statistical Power

Given that ME-ICA has a pronounced effect for boosting effect size estimation across canonical and cerebellar mentalizing regions, it follows that ME-ICA will have

favorable impact on statistical power. Here we have run power simulations for the purposes of future study planning in order to demonstrate the practical implications that ME-ICA has for researchers. These simulations mainly inform what we could expect in future work given the effect size estimates we have observed in the current study under ME-ICA and TSOC+MotReg analyses.

Using estimates of effect size from meta-analytically defined regions from the NeuroSynth mentalizing map (Fig 2A & 3A), we constructed power curves across a range of sample sizes (n=5 to n=100) and quantified the smallest sample size needed to obtain effects at an alpha level of 0.05 and power of 80%. Across all canonical regions, the median minimum sample size to achieve 80% power in ME-ICA analyses is n=18 and n=32 respectively across the SelfOther and Stories tasks. In contrast, the median minimum sample sizes for TSOC+MotReg is n=31 and n=46 respectively for the SelfOther and Stories task (see Fig 6 for power curves and minimum sample sizes to achieve 80% power) (Supp Table 1). In almost all cases, the required sample size for 80% power in ME-ICA were all well within ranges that are practical and characteristic of current practices (e.g., n<45). In contrast, with TSOC+MotReg, there were many cases where the minimum sample size needed to achieve 80% power exceeded ranges for a typical fMRI study (e.g., n>45). Computing the difference between minimum sample sizes needed to achieve 80% power for each canonical region shows that TSOC+MotReg requires an average of 18 to 19 more participants than ME-ICA across both tasks. Assuming a scanning cost of $300 per subject, this means that ME-ICA could amount to savings of $5,400 to $5,700.

If one was interested in cerebellar regions for mentalizing the situation is even more extreme, as the SelfOther task requires an average of 47 more participants for TSOC+MotReg compared to ME-ICA to achieve requisite 80% power levels; a savings of $14,100. The minimum sample sizes for 80% power in ME-ICA for cerebellar regions were within practically attainable ranges; at most n=43 (e.g., mCereb) and at best, only n=13 (e.g., rCereb). For the Stories task, in TSOC+MotReg data, 2 of the 3 cerebellar regions never reach 80% power over the maximum sample size of n=100 used in the simulations. Left cerebellum however, required 25 more participants in TSOC+MotReg compared to ME-ICA, amounting to a $7,500 savings (Fig 7). Left cerebellum is also indicative of a practical effect whereby using ME-ICA makes a much more feasible situation of needing to collect n=24, as compared to needing to collect n=49 without ME-ICA. ME-ICA would also allow 80% power to detect an effect in right cerebellum with n=35. The fact that these cerebellar power simulations show that statistical power in TSOC+MotReg is far below 80% within sample size ranges that are characteristic of typical fMRI studies, suggests that if the effect sizes we observe here are generalizable to the rest of the literature, the cerebellum may continue to go undetected in traditional studies due to low power for detecting subtle effects in the presence of severe artifacts from pulsation, bulk head motion, CSF flow, and 'dropout' when ME-ICA is not used to remove them.

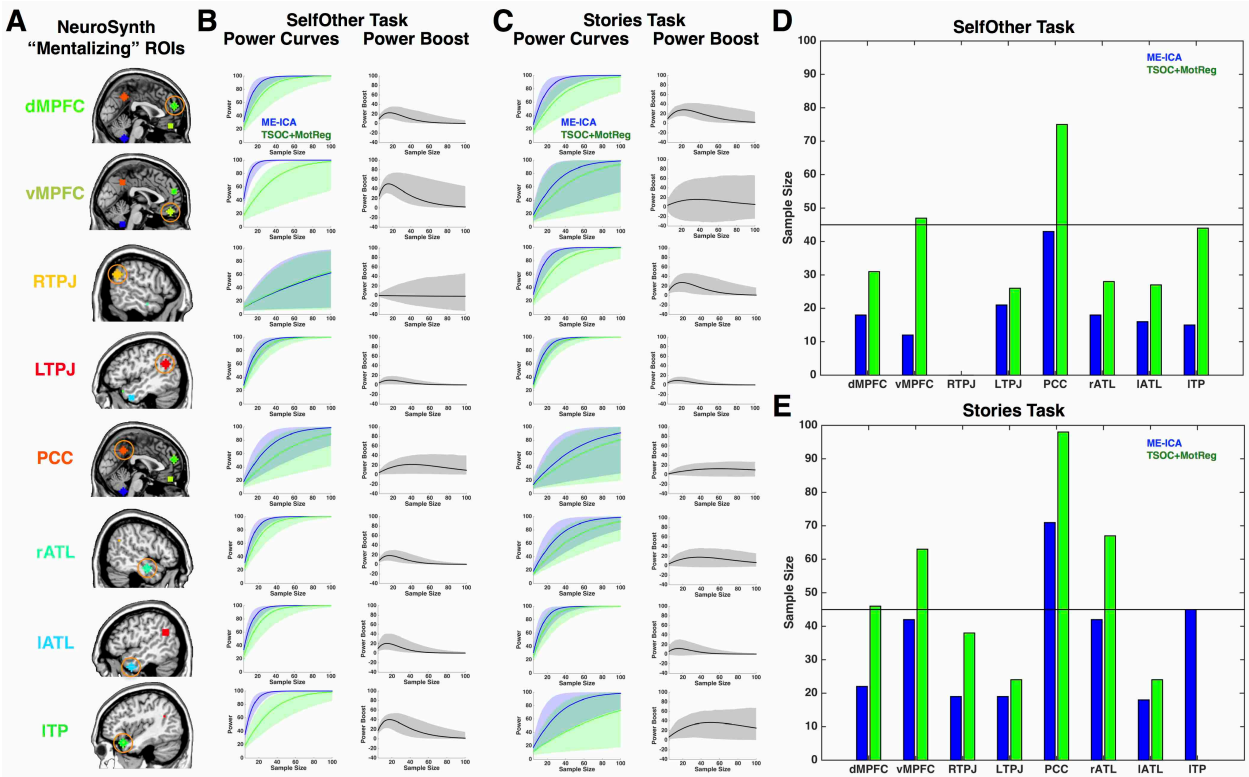

Fig 6: Statistical power simulations from canonical mentalizing regions. Panels B and C show power curves and ME-ICA-related power boosting over a range of sample sizes from n=5 to n=100, under effect sizes observed in ME-ICA (blue) and TSOC+MotReg (green). The shaded areas around curves in panels B and C indicate a range that encompasses the bootstrap 95% confidence interval. Panels D and E show the minimum sample sizes needed to detect effect at alpha = 0.05 and 80% power. The line at n=45 is intended to represent a cut-off, above which are sample sizes that are highly not characteristic of the typical task-based small sample size fMRI study. In Panels D and E, any instance where no bar is present indicates that power never reached 80% when going up to n=100. The requisite power for these specific instances is likely much greater than n=100, but these sample size ranges were not included in our simulations.

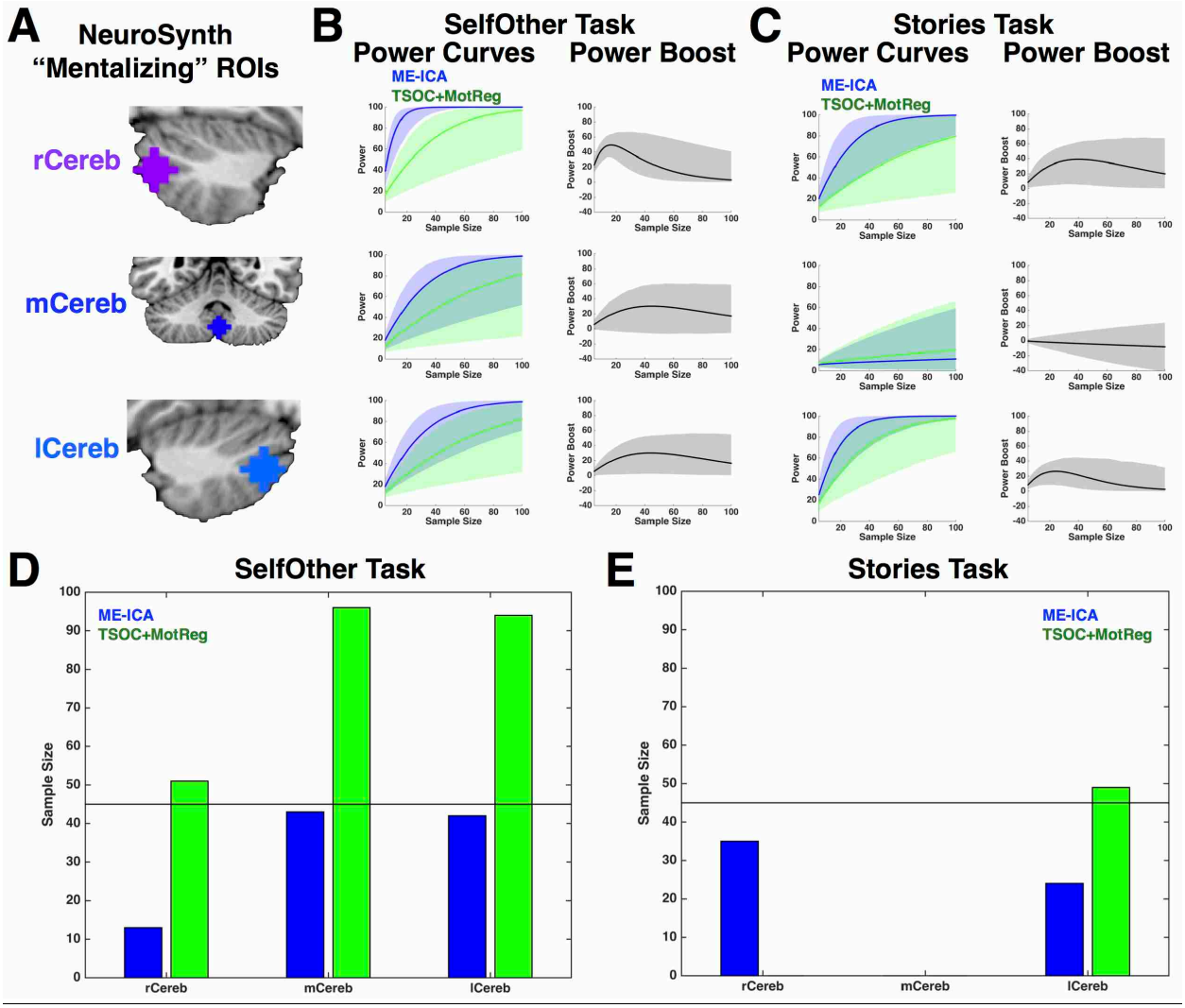

Fig 7: Statistical power simulations from cerebellar mentalizing regions. Panels B and C show power curves and ME-ICA-related power boosting over a range of sample sizes from n=5 to n=100, under effect sizes observed in ME-ICA (blue) and TSOC+MotReg (green). The shaded areas around curves in panels B and C indicate a range that encompasses the bootstrap 95% confidence interval. Panels D and E show the minimum sample sizes needed to detect effect at alpha = 0.05 and 80% power. The line at n=45 is intended to represent a cutoff, above which are sample sizes that are highly not characteristic of the typical task-based small sample size fMRI study. In Panels D and E, any instance where no bar is present indicates that power never reached 80% when going up to n=100. The requisite power for these specific instances is likely much greater than n=100, but these sample size ranges were not included in our simulations.

In examining power curves in Figs 6 and 7, another salient observation from the ME-ICA findings is a point of diminishing returns when power achieves levels of 95% or more, as the improvements in power for adding more subjects diminishes substantially.

We term this effect 'saturation'. It is important to underscore here that when using ME-ICA, most regions show power is saturated at sample sizes that are typical of fMRI studies and are thus practically attainable, while in TSOC+MotReg, sample size needed to hit saturation is mostly beyond ranges that are characteristic for a typical fMRI study (see Supp Table 1).

Discussion

Task-based fMRI studies are characteristically of small sample size and thus underpowered for all but the largest and most robust effects[1,2]. Furthermore, typical task-based fMRI studies do not apply advanced methods to mitigate substantial non-BOLD noise that is generally known to be inherent in such data. Combining small underpowered studies with little to no consideration of persistent non-BOLD noise that is present in the data even after typical pre-processing and statistical modeling creates a situation where most task-based studies are potentially missing key effects and makes for somewhat impractical conditions for most researchers where massive sample sizes are required to overcome such limitations. In this study we have taken a bottom-up approach to this issue by specifically addressing the problem of attenuating non-BOLD noise and the effects it may have on augmenting effect size and statistical power in task-based fMRI studies. We have employed methodological innovations in data acquisition using multi-echo fMRI, with integration of biophysically and statistically principled methods of isolating and removing non-BOLD sources of variation from fMRI data. This approach has proven extremely effective in applications to resting state fMRI data and connectivity estimation[10,13-16] as well as for mapping ultra-slow BOLD-related signal fluctuations during temporally extended tasks[29]. The goals of this study were to characterize the effect of removing non-BOLD noise from block-design task-based fMRI studies and we have shown the application of this method in the social cognitive domain of mentalizing.

ME-ICA-related Effect Size Boosting

Overall, we find that our methodological approach, ME-ICA, robustly enhances effect size estimation in second-level random-effects group analyses with one-sample t-tests. This holds true for both 'canonical' regions in social-cognitive domains such as mentalizing[18-24], but also in novel detection of 'non-canonical' regions located in the cerebellum[25,26]. Owing to its robustness, ME-ICA boosted effect size estimation across two different paradigms tapping the domain of mentalizing that vary greatly with regard to the stimuli, modality of stimulation, nature of the task, etc. With few exceptions, ME-ICA systematically increases effect size estimation around a median boost in effect size for canonical mentalizing regions of around 25-32%. We also found that in key regions thought to be important for social-cognitive function but which are characteristically affected by signal dropout such as ventromedial prefrontal cortex and temporal pole regions, effect size boosts are much greater and in the range of 50-100%. Similarly, in

non-canonical cerebellar regions effect sizes boosts were in the range of 43-108%. These data generally suggest as a proof-of-principle that ME-ICA allows for boosts in task-related effect sizes that are very substantial.

Practical Implications and Impact for Statistical Power and Future Study Planning

In addition to increasing ability to detect canonical and novel effects of interest, our analyses also suggest many practical implications that can have large impact on the way in which researchers typically conduct fMRI research. One obvious top-down approach to addressing issues of underpowered and substantially noisy studies is to make the experimental design decision to collect much more data within-subject and collect much larger sample sizes than are typically characteristic of task-based fMRI studies. While we would wholeheartedly agree and advocate that this kind of practice generally become more common in fMRI research, our simulations of statistical power and sample size also shed some additional insight on how the use of ME-ICA innovations can provide compelling and large steps forward towards ensuring studies are high-powered, while also balancing practical limitations such as scanning costs that typically cited as a main prohibitive reason for collecting more data. Here we find that in general terms, the use of ME-ICA induces boosts in effect size that are large enough to warrant studies with statistical power of 80% for detecting effects at an alpha level of 0.05 within sample size ranges that are typically more common and practical in fMRI research. For example, the median sample size for achieving 80% power across all canonical regions in both tasks is around n=18 to n=32 when using ME-ICA, and in cerebellar regions n=43 is sufficient. Without the use of ME-ICA, minimum sample sizes for 80% power are much larger and in many cases are outside of the realm that is typically characteristic of the average fMRI study. In our simulations we have also quantified these practical implications in terms of monetary savings to researchers and here it is also salient to see that ME-ICA can have true impact in monetary value. This practical point about monetary savings is one that is meant to address a pertinent point about the expensive nature of fMRI studies and how high costs are potentially prohibitive, especially for massive-N studies. If it is within a researcher's budget to collect more than is traditional for typical fMRI studies, we would most definitively advocate for this, and if researchers were to use ME-ICA, given the effect size boosts we characterize here, there is the potential for enabling very high powered studies within what is practically attainable for most research labs.

Finally, our power simulations also suggested another compelling point regarding costs versus benefits of acquiring very large samples after implementing an innovation like ME-ICA. Amongst the power curves for many of the regions investigated, we found that power quickly reached 'saturation levels' of greater than 95% power at sample sizes that are currently standard. These 'saturation levels' are of practical importance because they indicate where the returns in statistical power diminish substantially relative to investment in increasing sample size. Without ME-ICA, the sample sizes needed for the aforementioned power levels were summarily greater than n=40, and in many cases, not

attainable at the limit of n=100 for simulations. Thus, for many regions of interest in our study of neural systems involved in mentalizing, there is a theoretical ceiling to the power benefit of increasing sample size given current effect size exposed after ME-ICA denoising. Because most 'canonical' regions hit these saturation levels quickly when using ME-ICA, it may be more pertinent in future work to turn to examination of the more subtle effects and/or more fine-grained hypotheses, as the methodological innovation of ME-ICA would potentially allow for detection of more subtle effects with larger sample sizes. In addition there is further promise that with enhanced sensitivity due to ME-ICA, there are likely to be benefits on paradigms that are predicated on the detection of much more subtler effects such as activation differences between closely related stimuli (as in fMRI adaptation paradigms), multifactorial designs involving complex contrasts, etc.

It is important to stress that we are not advocating that researchers refrain from collecting large sample sizes when using ME-ICA. Besides the tacit understanding that statistical parameters we estimate will always be more precise with larger samples, there are also always other very compelling reasons to attain more data even when statistical power considerations such as these may have hit 'saturation levels' we speak of here. We would advocate that researchers still consider obtaining large samples even when such considerations are accounted for (funds permitting), as many new classes of questions or finer-grained hypotheses could be asked and evaluated. For example, rather than continuing to focus on hypotheses predicated on detection of effects that are present on-average in a large and potentially heterogeneous population, one could turn to a translationally more relevant approach towards making more individualized predictions that could start with defining discrete subgroups within a population where an effect systematically differs, and which might be linked back to specific genetic and/or environmental factors (e.g., [30]). In another example, we can turn away from a focus on the very large robust effects we currently consider as 'canonical' in many cognitive domains and begin to focus more on the much more subtle effects that may be of clinical significance, and which will still necessitate larger sample sizes. More towards the point we are trying emphasize here about arguments relevant to the 'saturation' effect on power, we want to mainly to address what many researchers would call a very big practical limitation on large costs for attaining large sample sizes. We would argue that when one uses ME-ICA, one can theoretically achieve high degrees of statistical power at levels that are not unattainable due to cost considerations. Given that we would still advocate for collecting as much data as possible, we view the use of ME-ICA as enabling studies to be highly powered for most canonical effects, and to also be well positioned for new discoveries of effects that are typically hidden in underpowered studies that do not address non-BOLD noise issues. Thus, the innovations described here with ME-ICA applied to the context of task-based fMRI research can plausibly have substantial impact on the common criticisms about fMRI research being underpowered, but may additionally have the benefit of progressing our knowledge of brain function even further by clearing up effects that are saturated in high-levels of non-BOLD noise and which are hidden at characteristically low sample sizes.

Enabling Discovery Science: Mentalizing and the Cerebellum

An especially compelling and important consequence of the ME-ICA induced effect size and power boosts is the potential for discovery of new effects that have been largely missed throughout the literature due to small underpowered studies that have not adequately addressed issues related to non-BOLD noise. As a case in point, here we find robust evidence for discrete cerebellar regions that should be more frequently considered as integral nodes of neural systems supporting mentalizing and theory of mind processes. Prior indications that these effects may be plausible come from meta-analytic evidence. In particular, van Overwalle and colleagues have suggested that specific cerebellar regions may be important in particular aspects of mentalizing[25,26]. For confirmation on this point, one need not look further than the NeuroSynth 'mentalizing' meta-analysis map for further support of cerebellar regions being important in mentalizing.

In this study, we found robust evidence apparent in whole-brain and targeted region of interest analyses across two different kinds of mentalizing tasks for the involvement of specific cerebellar regions in mentalizing. Upon inspection of the effect sizes derived from meta-analytically defined cerebellar regions, we find these cerebellar effects are quite subtle in analyses that do not implement ME-ICA innovations, and that the statistical power for such effects are expectedly low at standard sample sizes employed throughout the literature. This insight indicates that part of the explanation behind why such cerebellar regions potentially remain hidden from the literature is because the characteristically small sample size study is not sufficiently statistically powered to be able to detect such an effect. In contrast, after one employs ME-ICA innovations, these cerebellar effect sizes begin to approach the range of effects one would typically see in most 'canonical' regions and sample size needed to attain sufficient power for such effects are also within practically attainable ranges.

For further support regarding plausibility of cerebellar nodes being integral in neural circuits supporting mentalizing, we have also elucidated that resting state functional connectivity from respective cerebellar seed regions highlight networks that largely recapitulate the systems evoked during mentalizing tasks. These connectivity insights were discovered on the same set of subjects who participated in our activation paradigms. Furthermore, quantitative assessments of pattern similarity between resting state cerebellar connectivity patterns and mentalizing activation patterns provided striking evidence for a high degree of similarity across the entire brain for both maps. Interestingly, the ability to detect these high degrees of similarity are hindered without the use of ME-ICA and ME-ICR, as further analyses showed that activation-connectivity similarity is substantially attenuated under conditions for conventional functional connectivity analyses. This is a further methodological case in point for why the multi-echo innovations we utilize may be central for further enhancing neuroscientific discovery.

This evidence highlighting the similarity between neural systems evoked in mentalizing tasks with resting state cerebellar functional connectivity maps further illustrates the point that researchers studying social cognition, mentalizing and theory of mind, should focus more on the contributions that these cerebellar regions play in social cognitive processes. The cerebellum is historically a poorly understood and overlooked region in terms of its role in higher cognitive processes[25,31-34]. The ME-ICA innovations we present here should help researchers to gain a more stable foothold on cerebellar effects in the context of mentalizing and enable better circumstances for parsing apart how their role can further our understanding of such complex social cognitive processes. A promising avenue for future work on this topic would be to further understand the computational role the cerebellum plays in simulative processes that may be important in mentalizing[35]. Computationally, some theories about cerebellum posit that it may be integral for making internal models of representations relevant to higher-cognitive thought[36]. One proposed difference between cerebellar vs cortical learning mechanisms are that cerebellar learning takes more the form of supervised learning while unsupervised learning computations are more cortically-mediated[37]. Through the use of cerebello-thalamo-cortical loops, there could be a possibility for an important dual learning architecture to subserve early social learning[34], which may be important for lining up with dual process models of social cognition[38] and/or strategies such as anchoring and adjustment in social prediction[39]. In the context of simulation theories of mentalizing, this insight potentially opens up very important avenues of new research on how cerebellar computations might play some role in mental modeling of the social world.

Translationally, the link between cerebellum and mentalizing is also particularly intriguing, given the longstanding, yet independent, literatures in autism regarding the cerebellum[40-42] and mentalizing[43,44]. Wang and colleagues[34] have recently argued that developmental processes derailed within the cerebellum may be particularly important for understanding autism. Autism is well known for hallmark deficits in the domain of social-communication[45] and impairments in the development of mentalizing/theory of mind and self-referential cognition in autism[43,46-48] as well as atypical functioning of neural mechanisms that bolster such processes[49-51] are thought to be important as explanations behind social-communication deficits in autism. Thus, the intersection of developmental abnormalities in cerebellar development and their relationship to the development of mentalizing in autism will be an interesting new avenue of research enabled by these kinds of novel discoveries highlight in this study.

Methodological Issues, Caveats and Limitations

ME-ICA implements a sensitive and specific approach to identifying BOLD signals and removing non-BOLD signals from fMRI time series. While this method was established based on resting-state fMRI connectivity mapping, the method is shown here

to be highly effective in task-based activation mapping contexts. As ME-ICA is a first-level analysis and denoising approach, the specific gains from ME-ICA over TSOC+MotReg in second-level analysis are due to a combination of increased effect size and reduction in inter-subject variability of activation at the first level. The increase in effect size can be most associated to reduction of physiological and imaging artifacts that manifest as structured (non-thermal) signals. These sources of artifact are most prominent in the ventral brain regions, which manifest non-BOLD signals from cardiac and CSF pulsatility (i.e. periodic motion). These regions are also affected by susceptibility artifact due to proximity to air-tissue boundaries. Optimal combination of echoes was a key processing step because it utilized the signals of early TEs to circumvent signal dropout, making key regions of the ventral prefrontal cortex and anterior temporal cortex accessible to analysis. Based on these gains, it follows that the most significant gains due to ME-ICA over TSOC+MotReg were found in these brain regions. At the same time, these regions are amongst the least well understood in terms of normative function (particularly within social cognition) and roles in neuropsychiatric conditions.

An important caveat for this study is that our findings are based on block-design activation paradigms, utilizing relatively long-duration changes in susceptibility weighting. This differs from event-related paradigms, whereby activations may be associated with a significant inflow component that is S0-weighted. Future studies will involve assessing the suitability of ME-ICA for the analysis of event-related studies as well as other more novel task-designs. With regard to novel task-designs such as temporally extended tasks, we have previously shown that ME-ICA also has the ability to separate ultra-slow BOLD effects from slow non-BOLD effects[29], and this opens up a range of possibilities for new paradigms that may be particularly well-suited for temporally-extended and continuous tasks, such as more naturalistic paradigms for social cognition[52,53].

Another methodological point about the current study is that while the sample is otherwise a typically-developing sample (excluding one individual with an autism diagnosis), these individuals do represent a selected-sample from the general population, as all individuals were those who were born to mothers who underwent amniocentesis. Because amniocentesis is typically done for clinical reasons such as screening for chromosomal abnormalities, and is typically done on older parents, this is a potential caveat to keep in mind. The reason for this is that our sample here is originally designed to answer specific questions regarding fetal programming mechanisms of steroid hormones present in amniotic fluid samples for predicting variability in later brain development. It is possible that certain considerations regarding the association between increases in parental age and increases in the rate of de novo mutations are relevant for this type of population[54]. However, regarding specific points about boosts in effect size due to ME-ICA, our comparisons are not systematically biased by this issue of a selected sample, as all such head-to-head comparisons of ME-ICA vs TSOC are done within the same participants. Therefore, while we acknowledge the selected nature of this otherwise typically-developing sample, due to the within-

subject nature of the comparisons, it is unlikely that the main inferences we draw regarding ME-ICA related improvements in effect size and power are somehow biased by this fact. We think it would be alarming to see that such ME-ICA-related improvements disappear in a more randomly selected sample of the general population, and cannot see plausible explanations for why this would be expected, given that the nature of these innovations are centered in biophysically principled characteristics of BOLD signal and the acquisition of multi-echo data to leverage such characteristics in denoising.

Another caveat of the current study deals with the age range of participants in this sample. All participants were within adolescent age ranges, and given particular kinds of developmental effects that may exist for some regions it is possible that different results will manifest at different ages. For example, RTPJ exhibits a developmental effect for increasing specialization for mentalizing across this age range and has been previously shown specifically for the Stories task we have employed in the current study[55,56]. Regarding this point on the potentially less specialized nature of RTPJ for mentalizing at pre-adulthood ages and the insensitivity we observed in RTPJ on the SelfOther task, it may be that effect size improvements could be observed on this task in adulthood when mentalizing specialization in RTPJ is more prominent[21,50].

It is important to specifically point out that with this study we are not stating that our effect size boosts and power simulations are what one should expect across any type of experimental setting and any type of semi-related paradigm on mentalizing. As previously pointed out by Mumford and Nichols, for any given study, a specific power analysis is necessary to be applied to each specific type of experimental design and statistical model, as applying power insights directly from a paper such as this one into an experimental context that is quite different from ours (e.g., different paradigms, different number of runs of length of scanning per subject) will likely lead to over- or underestimation of power[6]. We advocate that careful piloting should still always be done before any experiment to make informed decisions about experimental design that are specific to the paradigm and statistical model one is implementing. However, our data does show a proof-of-principle for how in practice ME-ICA could generally translate into statistical power benefits in task-based activation mapping of neural systems supporting mentalizing, and therefore should be an impetus for researchers in future studies to consider the use of ME-ICA over traditional single echo EPI acquisition and analysis approaches.

Finally, our effect size estimates and power simulations comparing TSOC+MotReg to ME-ICA are likely to be conservative estimates of the gains from ME-ICA versus conventional single-echo fMRI acquisition and analysis. The reason for this, is that to be fair in making comparisons against ME-ICA, we used data that are optimally combined across multiple echo readouts from a multi-echo EPI sequence. Optimally combined time series data (TSOC) have already been shown to have substantial increases in temporal signal-to-noise ratio (tSNR; up to doubling) versus analogous

single-echo datasets[13], via homogenizing functional contrast across the brain while attenuating thermal noise (combination is a weighted average implementing a matched-filter). Because tSNR is reduced in traditional single-echo sequences, the expectation is that what would likely be found in traditional single-echo sequences is at best the same, but probably some degree worse than the TSOC+MotReg results we present here in this study. Therefore, a caveat to note is that results presented for TSOC+MotReg are likely to be overoptimistic about the situation likely in traditional single-echo acquisitions. This potentially means that comparisons of ME-ICA to traditional single-echo EPI data are likely to show even larger ME-ICA related improvements in effect size and power.

Conclusions

This study demonstrates that fMRI innovations based on multi-echo EPI data acquisition alongside ICA and principled removal of non-BOLD artifacts from BOLD-weighted fMRI datasets can substantially improve the task-based fMRI research by increasing effect size and statistical power at group-level (i.e. one-sample t-tests). The need or present technical capability to remove substantial amounts of non-BOLD noise has not been given adequate attention with respect to discussions on underpowered studies and the 'crisis of confidence', especially those centered expressly on sample size. Our study shows that attention should be drawn to this methodological aspect of the problem, alongside considering increases in scan time or sample size. We emphasize that the non-BOLD artifact removal by ME-ICA does not invoke any assumptions of task design or functional brain organization or localization and therefore is a truly unbiased technique. Going forward, we expect the sensitive removal of non-BOLD effects to be more readily considered a practically and financially accessible means of rigorously conducting fMRI research. In implementing these innovations, we may also be positioning ourselves for discovery of novel effects that may have been overlooked in the literature, which opens up possibilities for new discoveries in brain organization and its relationship to cognitive and behavioral function.

Methods

Participants

Participants in this study were 69 adolescents (34 males, 35 females, mean age = 15.45 years, sd age = 0.99 years, range = 13.22-17.18 years) from a cohort of individuals whose mothers underwent amniocentesis during pregnancy for clinical reasons (i.e. screening for chromosomal abnormalities), where the main focus was to study the fetal programming effects of steroid hormones on adolescent brain and behavioral development. At amniocentesis, none of the individuals screened positive for any chromosomal abnormalities and were thus considered typically developing. Upon recruitment for this particular study, we additionally checked for any self- or parent-reported neuropsychiatric conditions. One of the individuals had a diagnosis on the autism spectrum. The remaining participants did not have any other kind of neurological or psychiatric diagnosis. Analyses were done on the full sample of 69 individuals, as analyses leaving out the one patient with an autism diagnosis did not change any of the results.

Task Design

Participants were scanned using two block-design fMRI paradigms. The first paradigm, which we call the 'SelfOther' task, was a 2 x 2 within-subjects factorial design which contained two contrasts that tapped either self-referential cognition and mentalizing and was similar in nature to previously published studies[21,49,50]. Briefly, participants were asked to make reflective judgments about either themselves or the British Queen that varied as either a mentalistic (e.g., "How likely are [you/the Queen] to think that it is important to keep a journal?") or physical judgment (e.g., "How likely are [you/the Queen] to have bony elbows?"). Participants made their judgments on a 1-4 scale, where 1 indicated 'not at all likely' and 4 indicated 'very likely'. All stimuli were taken from Jason Mitchell's lab and have been used in prior studies on mentalizing and self-referential cognition[35,57]. The SelfOther task was presented in 2 scanning runs (8:42 duration per run; 261 volumes per run). Within each scanning run there were 4 blocks per condition, and within each block there were 4 trials of 4 second duration each. Task blocks were separated from each other by a 16 second fixation block. The first 5 volumes of each run were discarded to allow for T2 stabilization effects.

The second paradigm, which we call the 'Stories' task, was block-design which contained two contrasts that tapped mentalizing and language. The paradigm was taken directly from the study by Gweon and colleagues[56] and the stimuli and stimulus presentation scripts were acquired directly from Hyowon Gweon and Rebecca Saxe. Briefly, participants listened to a series of stories presented auditorily. The stories differed in content and could either be mentalistic, social, or physical. The social stories contained descriptions of people and characters but made no statements that referenced

mental states. Physical stories were segments of stories that described the physical setting and did not include people. Mental stories were segments that included references to people as main characters and made references to mental states that those characters held. The paradigm also included blocks for two other kinds of language control conditions that were not examined in this manuscript (i.e. stories read in a foreign language (e.g., Russian, Hebrew, and Korean) and blocks of music played by different instruments (e.g., guitar, piano, saxophone, and violin)). After participants heard each story segment they were given a choice of whether a specific auditory segment logically came next. This was introduced to verify that participants were paying close attention to the stories and the details inside each story segment. The Stories task was presented in 2 scanning runs (6:36 duration per run; 192 volumes per run) and within each scanning run there were 2 blocks per condition. The first 6 volumes were discarded to allow for T2 stabilization effects.

Resting state data was also collected on each participant with a 10 minute long 'eyes-open' run (i.e. 300 volumes), where participants were asked to stare at a central fixation cross and to not fall asleep. The multi-echo EPI sequence was identical to those used in the task paradigms.

fMRI Data Acquisition

All MRI scanning took place on a 3T Siemens Tim Trio MRI scanner at the Wolfson Brain Imaging Centre in Cambridge, UK. Functional imaging data during task and rest was acquired with a multi-echo EPI sequence with online reconstruction (repetition time (TR), 2000 ms; field of view (FOV), 240 mm; 28 oblique slices, alternating slice acquisition, slice thickness 3.8 mm; TE = 13, 31, and 48 ms, GRAPPA acceleration factor 2, BW=2368 Hz/pixel, flip angle, 90°). Anatomical images were acquired using a T1-weighted magnetization prepared rapid gradient echo (MPRAGE) sequence for warping purposes (TR, 2300 ms; TI, 900 ms; TE, 2.98 ms; flip angle, 9°, matrix 256 × 256 × 256, field-of-view 25.6 cm).

Multi-Echo ICA (ME-ICA) Pipeline

Data were processed by ME-ICA using the tool *meica.py* as distributed in the AFNI neuroimaging suite (v2.5 beta10), which implemented both basic fMRI image preprocessing and decomposition-based denoising. *meica.py* implemented AFNI tools for preprocessing. For the processing of each subject, first the anatomical image was skull-stripped and then warped nonlinearly to the MNI anatomical template using AFNI *3dQWarp*. The warp field was saved for later application to functional data. For each functional dataset, the first TE dataset was used to compute parameters of motion correction and anatomical-functional coregistration, and the first volume after equilibration was used as the base EPI image. Matrices for de-obliquing and six-parameter rigid body motion correction were computed. Then, 12-parameter affine anatomical-functional coregistration was computed using the local Pearson correlation

(LPC) cost functional, using the gray matter segment of the EPI base image computed with AFNI *3dSeg* as the LPC weight mask. Matrices for de-obliquing, motion correction, and anatomical-functional coregistration were combined with the standard space non-linear warp field to create a single warp for functional data. The dataset of each TE was then slice-time corrected and spatially aligned through application of the alignment matrix, and the total nonlinear warp was applied to the dataset of each TE. Critically, data were not spatially smoothed using a full-width-half-max (FWHM) spatial filter. The effective smoothness of the data after preprocessing (which inadvertently adds smoothing due to interpolation and re-gridding) was found to be ~5mm, compared to isotropic voxel size of 3.8mm. Note that the application of FWHM smoothing *adds* to the image smoothness, such that 6mm FWHM smoothing yields an 11m FWHM effective smoothing. No time series filtering was applied in the preprocessing phase.

Time series denoising occurred over several steps, based on fitting multi-echo data and its statistical components to signal models reflecting the T2* and S0 signal decay processes. This has been detailed in our prior work[13,29] and is summarized here. T2* and S0 images were computed from means of time series of different TEs. The separate TE time series datasets were "optimally combined" as a weighted average, with weights being a function of TE and local T2*. This procedure implemented a matched-filter that produced a contrast-optimized or "high dynamic range" time series dataset where the functional contrast-to-noise at each voxel was maximized and thermal noise is attenuated. The optimally combined data was then decomposed to further remove [approximately Gaussian distributed] thermal noise and concurrently reduce dimensionality by a known number of degrees of freedom. This was done by principal components analysis (PCA) decomposition, followed by TE-dependence analysis of each principal component. PCA components that exhibited neither TE-dependence nor TE-independence and explained less than a data-driven threshold for variance explained were counted as thermal noise and projected out. This procedure is referred to as ME-PCA. Next, ICA (FastICA with *tanh* contrast function) was applied to the dimensionally reduced dataset to yield non-Gaussian spatial components indicating distinct signal processes that were orthogonal and statistically independent, alongside a time course mixing matrix. The mixing matrix was fit to the time series of each separate TE, producing coefficient maps for each component and TE. The signal scaling of each component across TEs was then use to compute Kappa (κ) and Rho (ρ), which were pseudo-F statistics indicating component-level TE-dependence and TE-independence, respectively. A component classification algorithm was then applied that differentiated components into BOLD and non-BOLD categories. Lastly, the linear combination of BOLD component maps and their time series (both derived from the optimally combined time series) produced the ME-ICA BOLD dataset.

Task-fMRI Data Analysis

All first and second level statistical modeling was performed in SPM8 (http://www.fil.ion.ucl.ac.uk/spm/), using the general linear model (GLM). First level analyses modeled the hemodynamic response function (HRF) with the canonical HRF,

and used a high-pass filter of 1/128 Hz. For analyses on TSOC data, we further modeled out motion parameters as regressors of no interest from all first-level individual subject analyses. However, we did not include such motion regressors in data preprocessed via ME-ICA, as such artifact is already removed in principled manner. Each subject's first-level analysis modeled the specific contrast of Mentalizing>Physical, and these contrast images were input into second-level random effects group analyses. All whole-brain second-level group analyses are thresholded at a voxel-wise FDR $q<0.05$[58].

Resting State fMRI Connectivity Analysis

Resting state connectivity was estimated using the multi-echo independent components regression (ME-ICR) technique developed by Kundu and colleagues[13]. This analysis technique effectively controls for false positives in connectivity estimation by using the number of independent components estimated by ME-ICA as the effective degrees of freedom in single-subject connectivity estimation. Once ME-ICA has estimated number of components, these component maps are concatenated, and connectivity is estimated by computing the correlation of ICA coefficients between the seed and other brain voxels. The seed regions we have chosen are the peak voxels from the NeuroSynth 'mentalizing' map in right and left hemisphere cerebellum (RH MNI x = 29, y = -82, z = -39; LH MNI x = -25, y = -78, z = -39). Connectivity GLM analyses were implemented within SPM and the second-level group connectivity maps are thresholded with a voxel-wise FDR threshold of $q<0.05$.

To assess the similarity between whole-brain resting state connectivity and Mentalizing>Physical activation maps, we used robust regression[59] to compute the correlation between the whole-brain connectivity and activation maps. Robust regression allows for protection against the effects of outliers that are particularly pronounced in the connectivity maps, since voxels that contain or are proximally close to the seed voxel exhibit very large connectivity values.

Conventional functional connectivity analyses were also implemented on the TSOC data. Here we used AFNI *3dBandpass* to bandpass filter the data between 0.01 and 0.1 Hz, and specifically used the *-ort* argument to additionally remove motion-related variability all in one step. No other steps were taken to denoise the data (e.g., global signal regression, white matter regression, etc). The bandpass filtered and motion-regressed data were then inserted into GLMs in SPM8.

To compare the difference between activation-connectivity correlations for ME-ICR vs TSOC+MotReg, we use the *paired.r* function within the *psych* R library (http://cran.r-project.org/web/packages/psych/) to obtain z statistics to describe the difference between correlations. However, no hypothesis tests (i.e. p-values) are computed for these analyses as they are not needed since the comparisons are on

correlations estimated from the entire population of interest (i.e. all voxels within whole-brain volume).

Effect Size Estimation and Power Simulations

All effect size and power estimates were computed with the *fmripower* Matlab toolbox (http://fmripower.org)[6]. The effect sizes are expressed in standard deviation units and are analogous to Cohen's d. The Type I error was set to 0.05 and we computed power across a sample size range from n=5 to n=100. All effect size and power estimates were estimated from independently defined meta-analytic ROIs identified by NeuroSynth (http://neurosynth.org) for the feature 'mentalizing'. This feature contained 98 studies and 4526 activations. The NeuroSynth 'mentalizing' mask was first resampled to the same voxel sizes as the current fMRI datasets. Because regions surviving the NeuroSynth analysis at FDR q<0.01 were large and contained multiple peaks (e.g., medial prefrontal cortex comprised both dorsal and ventral subregions), we constrained ROIs further by finding peak voxels within each region, and constructing a 8mm sphere around each peak. This resulted in 11 separate ROIs. Eight of the 11 have been reported and heavily emphasized in the literature (dorsomedial prefrontal cortex (dMPFC): x = -2, y = 60, z = 22; ventromedial prefrontal cortex (vMPFC): x = -2, y = 48, z = -20; right temporo-parietal junction (RTPJ): x = 59, y = -55, z = 27; left temporo-parietal junction (LTPJ): x = -48, y = -55, z = 26; posterior cingulate cortex/precuneus (PCC): x = 2, y = -52, z = 42; right anterior temporal lobe (rATL): x = 48, y = -6, z = -20; left anterior temporal lobe (lATL): x = -52, y = 6, z = -35; left temporal pole (lTP): x = -40, y = 21, z = -24). The remaining 3 regions are located in the cerebellum (right hemisphere cerebellar region Crus II (rCereb): x = 29, y = 82, z = -39; medial cerebellar region IX (mCereb): x = 2, y = -52, z = -47; left hemisphere cerebellar region Crus II (lCereb): x = -25, y = -78, z = -39) and have been relatively overlooked in the literature, with some exceptions that also rely on meta-analytic inference[25].

To get an indication of how big the effect size boost due to ME-ICA was, we computed a measure of effect size percentage increase operationalized as ($ES_{ME\text{-}ICA}$ - ES_{TSOC})/abs(ES_{TSOC}). In order to get an indication of how big the boost due to ME-ICA was on power curves, we computed the difference score between power curves for ME-ICA and TSOC (e.g., $PowerCurve_{ME\text{-}ICA}$ - $PowerCurve_{TSOC}$). We used bootstrapping (1000 resamples) to re-run SPM second-level group analysis and fmripower computations in order to construct 95% confidence intervals around effect size and power estimates. To further describe the effects of ME-ICA over and above TSOC+MotReg pipelines, we have computed the minimum sample size to achieve 80% power, minimum sample size to achieve 95% or more power (what we call 'power saturation' levels), and the sample size and cost reduction due to using ME-ICA to achieve a study with 80% power, assuming a per subject scanning cost of $300.

References

1 Button, K. S. et al. Power failure: why small sample size undermines the reliability of neuroscience. Nature reviews. Neuroscience 14, 365-376, (2013).

2 Yarkoni, T. Big correlations in little studies: Inflated fMRI correlations reflect low statistical power - Commentary on Vul et al.(2009). Perspectives on Psychological Science 4, 294-298, (2009).

3 Ioannidis, J. P. Why most published research findings are false. PLoS medicine 2, e124, (2005).

4 Simmons, J. P., Nelson, L. D. & Simonsohn, U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol Sci 22, 1359-1366, (2011).

5 Desmond, J. E. & Glover, G. H. Estimating sample size in functional MRI (fMRI) neuroimaging studies: statistical power analyses. J Neurosci Methods 118, 115-128, (2002).

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

7 Friston, K. Ten ironic rules for non-statistical reviewers. Neuroimage 61, 1300-1310, (2012).

8 Lindquist, M. A., Caffo, B. & Crainiceanu, C. Ironing out the statistical wrinkles in "ten ironic rules". Neuroimage 81, 499-502, (2013).

9 Murphy, K., Birn, R. M. & Bandettini, P. A. Resting-state fMRI confounds and cleanup. Neuroimage 80, 349-359, (2013).

10 Kundu, P., Inati, S. J., Evans, J. W., Luh, W. M. & Bandettini, P. A. Differentiating BOLD and non-BOLD signals in fMRI time series using multi-echo EPI. Neuroimage 60, 1759-1770, (2012).

11 Kay, K. N., Rokem, A., Winawer, J., Dougherty, R. F. & Wandell, B. A. GLMdenoise: a fast, automated technique for denoising task-based fMRI data. Frontiers in neuroscience 7, 247, (2013).

12 Salimi-Khorshidi, G. et al. Automatic denoising of functional MRI data: combining independent component analysis and hierarchical fusion of classifiers. Neuroimage 90, 449-468, (2014).

13 Kundu, P. et al. Integrated strategy for improving functional connectivity mapping using multiecho fMRI. Proceedings of the National Academy of Sciences of the United States of America 110, 16187-16192, (2013).

14 Kundu, P., Santin, M. D., Bandettini, P. A., Bullmore, E. T. & Petiet, A. Differentiating BOLD and non-BOLD signals in fMRI time series from anesthetized rats using multi-echo EPI at 11.7 T. Neuroimage 102 Pt 2, 861-874, (2014).

15 Kundu, P. et al. Robust resting state fMRI processing for studies on typical brain development based on multi-echo EPI acquisition. Brain imaging and behavior, (2015).

16 Olafsson, V., Kundu, P., Wong, E. C., Bandettini, P. A. & Liu, T. T. Enhanced identification of BOLD-like components with multi-echo simultaneous multi-slice (MESMS) fMRI and multi-echo ICA. Neuroimage, (2015).

17 Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C. & Wager, T. D. Large-scale automated synthesis of human functional neuroimaging data. Nat Methods 8, 665-670, (2011).

18 Frith, U. & Frith, C. D. Development and neurophysiology of mentalizing. Philosophical transactions of the Royal Society of London 358, 459-473, (2003).

19 Saxe, R. & Powell, L. J. It's the thought that counts: Specific brain regions for one component of theory of mind. Psychol Sci 17, 692-699, (2006).

20 van Overwalle, F. Social cognition and the brain: A meta-analysis. Hum Brain Mapp 30, 829-858, (2009).

21 Lombardo, M. V. et al. Shared neural circuits for mentalizing about the self and others. J Cogn Neurosci 22, 1623-1635, (2010).

22 Schurz, M., Radua, J., Aichhorn, M., Richlan, F. & Perner, J. Fractionating theory of mind: a meta-analysis of functional brain imaging studies. Neurosci Biobehav Rev 42, 9-34, (2014).

23 Spunt, R. P. & Adolphs, R. Validating the Why/How contrast for functional MRI studies of Theory of Mind. Neuroimage 99, 301-311, (2014).

24 Schaafsma, S. M., Pfaff, D. W., Spunt, R. P. & Adolphs, R. Deconstructing and reconstructing theory of mind. Trends Cogn Sci 19, 65-72, (2015).

25 Van Overwalle, F., Baetens, K., Marien, P. & Vandekerckhove, M. Social cognition and the cerebellum: a meta-analysis of over 350 fMRI studies. Neuroimage 86, 554-572, (2014).

26 Van Overwalle, F., Baetens, K., Marien, P. & Vandekerckhove, M. Cerebellar areas dedicated to social cognition? A comparison of meta-analytic and connectivity results. Soc Neurosci, 1-8, (2015).

27 Buckner, R. L., Krienen, F. M., Castellanos, A., Diaz, J. C. & Yeo, B. T. The organization of the human cerebellum estimated by intrinsic functional connectivity. Journal of neurophysiology 106, 2322-2345, (2011).

28 Andrews-Hanna, J. R., Saxe, R. & Yarkoni, T. Contributions of episodic retrieval and mentalizing to autobiographical thought: evidence from functional neuroimaging, resting-state connectivity, and fMRI meta-analyses. Neuroimage 91, 324-335, (2014).

29 Evans, J. W., Kundu, P., Horovitz, S. G. & Bandettini, P. A. Separating slow BOLD from non-BOLD baseline drifts using multi-echo fMRI. Neuroimage 105, 189-197, (2015).

30 Lombardo, M. V. et al. Different functional neural substrates for good and poor language outcome in autism. Neuron, (2015).

31 Schmahmann, J. D. The cerebellum and cognition. (Academic Press, 1997).

32 Stoodley, C. J. & Schmahmann, J. D. Functional topography in the human cerebellum: a meta-analysis of neuroimaging studies. Neuroimage 44, 489-501, (2009).

33 Buckner, R. L. The cerebellum and cognitive function: 25 years of insight from anatomy and neuroimaging. Neuron 80, 807-815, (2013).

34 Wang, S. S., Kloth, A. D. & Badura, A. The cerebellum, sensitive periods, and autism. Neuron 83, 518-532, (2014).

35 Mitchell, J. P., Macrae, C. N. & Banaji, M. R. Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. Neuron 50, 655-663, (2006).

36 Ito, M. Control of mental activities by internal models in the cerebellum. Nature reviews 9, 304-313, (2008).

37 Doya, K. What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? Neural Networks 12, 961-974, (1999).

38 Lieberman, M. D. Social cognitive neuroscience: a review of core processes. Annual review of psychology 58, 259-289, (2007).

39 Epley, N., Keysar, B., Van Boven, L. & Gilovich, T. Perspective taking as egocentric anchoring and adjustment. J Pers Soc Psychol 87, 327-339, (2004).

40 Courchesne, E., Yeung-Courchesne, R., Press, G. A., Hesselink, J. R. & Jernigan, T. L. Hypoplasia of cerebellar vermal lobules VI and VII in autism. The New England journal of medicine 318, 1349-1354, (1988).

41 Murakami, J. W., Courchesne, E., Press, G. A., Yeung-Courchesne, R. & Hesselink, J. R. Reduced cerebellar hemisphere size and its relationship to vermal hypoplasia in autism. Archives of neurology 46, 689-694, (1989).

42 Courchesne, E. et al. Cerebellar hypoplasia and hyperplasia in infantile autism. Lancet 343, 63-64, (1994).

43 Baron-Cohen, S., Leslie, A. M. & Frith, U. Does the autistic child have a "theory of mind"? Cognition 21, 37-46, (1985).

44 Baron-Cohen, S. Mindblindness: An essay on autism and theory of mind., (MIT Press, 1995).

45 Lai, M. C., Lombardo, M. V. & Baron-Cohen, S. Autism. Lancet, (2013).

46 Lombardo, M. V., Barnes, J. L., Wheelwright, S. J. & Baron-Cohen, S. Self-referential cognition and empathy in autism. PLoS One 2, e883, (2007).

47 Lombardo, M. V. & Baron-Cohen, S. Unraveling the paradox of the autistic self. WIREs Cognitive Science 1, 393-403, (2010).

48 Lombardo, M. V. & Baron-Cohen, S. The role of the self in mindblindness in autism. Consciousness and cognition, (2010).

49 Lombardo, M. V. et al. Atypical neural self-representation in autism. Brain 133, 611-624, (2010).

50 Lombardo, M. V., Chakrabarti, B., Bullmore, E. T., Consortium, M. A. & Baron-Cohen, S. Specialization of right temporo-parietal junction for mentalizing and its association with social impairments in autism. Neuroimage 56, 1832-1838, (2011).

51 Yang, D. Y., Rosenblau, G., Keifer, C. & Pelphrey, K. A. An integrative neural model of social perception, action observation, and theory of mind. Neurosci Biobehav Rev 51C, 263-275, (2015).

52 Zaki, J., Weber, J., Bolger, N. & Ochsner, K. The neural bases of empathic accuracy. Proceedings of the National Academy of Sciences of the United States of America 106, 11382-11387, (2009).

53 Schilbach, L. et al. Toward a second-person neuroscience. Behav Brain Sci 36, 393-414, (2013).

54 Kong, A. et al. Rate of de novo mutations and the importance of father's age to disease risk. Nature 488, 471-475, (2012).

55 Saxe, R. R., Whitfield-Gabrieli, S., Scholz, J. & Pelphrey, K. A. Brain regions for perceiving and reasoning about other people in school-aged children. Child Dev 80, 1197-1209, (2009).

56 Gweon, H., Dodell-Feder, D., Bedny, M. & Saxe, R. Theory of mind performance in children correlates with functional specialization of a brain region for thinking about thoughts. Child development 83, 1853-1868, (2012).

57 Jenkins, A. C., Macrae, C. N. & Mitchell, J. P. Repetition suppression of ventromedial prefrontal activity during judgments of self and others. Proceedings of the National Academy of Sciences of the United States of America 105, 4507-4512, (2008).

58 Genovese, C. R., Lazar, N. A. & Nichols, T. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. Neuroimage 15, 870-878, (2002).

59 Wager, T. D., Keller, M. C., Lacey, S. C. & Jonides, J. Increased sensitivity in neuroimaging analyses using robust regression. Neuroimage 26, 99-113, (2005).