Title: Improving effect size estimation and statistical 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], Rosemary J. Holt[3], Jack Waldman[3], Amber N. V. 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. fMRI signals also inherently possess various sources of non-BOLD noise that further hampers ability to detect subtle effects. Here we take a bottom-up approach to addressing these problems via implementing multi-echo fMRI data acquisition and denoising innovations that can substantially improve effect size estimation and statistical power. We show that effect sizes on two different tasks within the social cognitive domain of mentalizing/theory of mind were enhanced at a median rate of 27% in regions canonically associated with mentalizing, while much more substantial boosts (43-130%) were observed in non-canonical cerebellar areas. This effect size boosting is primarily a consequence of reduction of non-BOLD noise at the subject level, which then translates into consequent reductions in between-subject variance. Power simulations demonstrate that enhanced effect size enables highly-powered studies at traditional sample sizes. Moreover, the cerebellar effects observed after applying our multi-echo innovations may be unobservable with conventional imaging at traditional sample sizes. The 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 true and 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 is to increase sample size, increase within-subject scan time, or both. These recommendations are pragmatic mainly because these variables are within the control of the researcher during study design. 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. Poor and variable quality data 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]. 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 address problems related to statistical power through specific targeting of the problems related to non-BOLD artifact variability. 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. Therefore, ME-ICA is a biophysically and statistically principled bottom-up approach towards identifying and retaining BOLD-related variability while systematically removing non-BOLD variation.

We apply ME-ICA to task-related activation mapping with block-designs and as a proof-of-principle, we evaluate the method against other prominent denoising procedures (i.e. motion regression, GLMdenoise[11]). As a specific application of the method, we examine its impact on two separate tasks (i.e. the 'SelfOther' and 'Stories' tasks) tapping neural systems supporting the social cognitive domain or mentalizing and theory of mind and highlight its effects in terms of effect size estimation and power. We evaluate the impact of the method on two sets of brain regions; 'canonical' regions typically highlighted as important in the neural systems for mentalizing[12-18] and 'non-canonical' regions in the cerebellum[19].

Results

ME-ICA Denoising 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 of ME-ICA denoising directly on the time series. ME sequences capture the decay of EPI images and (time series) with increasing TE, shown in Fig 1A. For example, ME data show the signal evolution 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 data from ventromedial prefrontal cortex (vMPFC), posterior cingulate cortex/precuneus (PCC), and right cerebellum in order to demonstrate the effect of optimal combination on the time series, and then the effect of removing non-BOLD noise using ME-ICA relative to modeled task blocks. It is particularly apparent that ME-ICA, without prior information on task structure, 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.

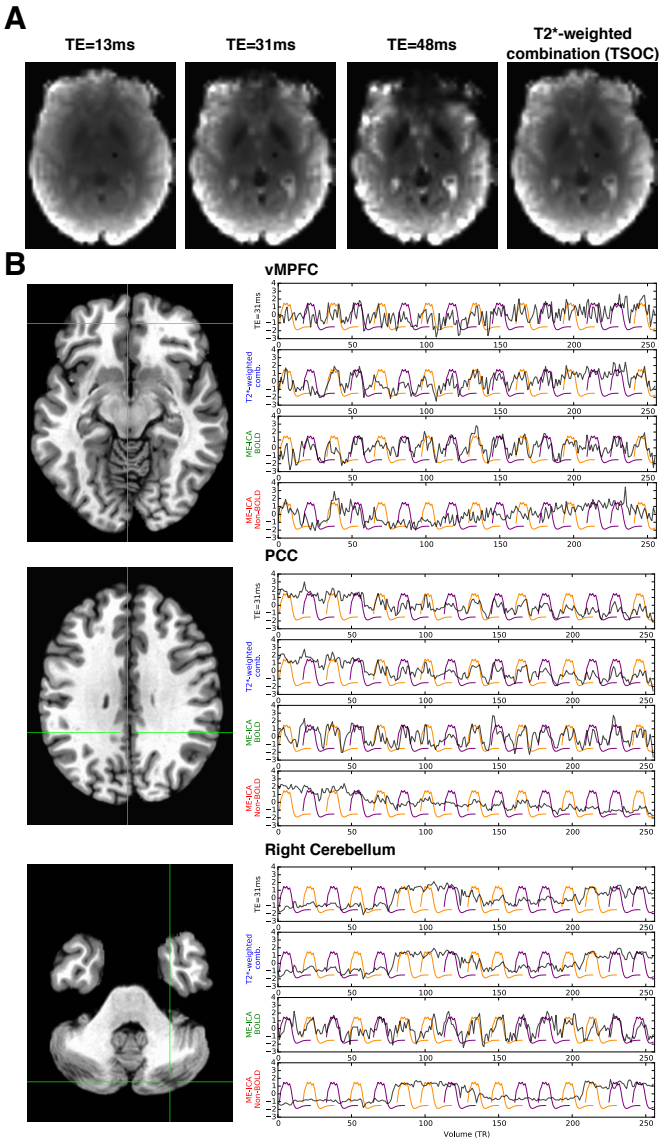

Fig 1: Multi-Echo Signal Characterization. Panel A shows the signal decay captured in multi-echo EPI images, for a single representative volume. With longer TE, gray/white contrast increases. Susceptibility artifact (e.g. dropout) also increases, as regions near in proximity to air-tissue boundaries have shorter T2*. The T2*-weighted optimal combination (TSOC) implements a matched-filter of TE images yielding a new image with optimized gray/white contrast (TE~T2*) and mitigation of susceptibility artifact. Panel B shows comparisons of time series data across three regions of interest; ventromedial prefrontal cortex (vMPFC), posterior cingulate cortex/precuneus (PCC), and right cerebellum. Each comparison shows the time series before model-based filtering of the middle TE image (black), TSOC image (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 evaluating ME-ICA-related effects on group-level inference, we examined the influence on non-BOLD denoising on effect size estimation. Effect size is operationalized here as a standardized measure of distance from 0 expressed in standard deviation units (i.e. mean/sd) and is analogous to Cohen's d. As illustrated in Fig 2, ME-ICA outperforms two other prominent methods for denoising (GLMdenoise[11] and regressing out motion parameters or TSOC+MotReg). This enhanced performance is evident across both mentalizing tasks and in nearly every single region investigated. Quantifying the magnitude of effect size boosting as the difference in effect size estimates, we find that the median ME-ICA induced boost for canonical mentalizing regions is 27%. Boosts were much larger (nearly always greater than 50%) in areas like vMPFC and left temporal pole (ITP) that characteristically suffer from signal dropout. Amongst cerebellar areas, right and left cerebellar Crus I/II areas showed evidence of even larger effect size boosts ranging from 55-130% increases when compared to GLMdenoise and

43-108% increases when compared to TSOC+MotReg. See Supplementary Table 1 for full characterization of effect size estimates and effect size boosts.

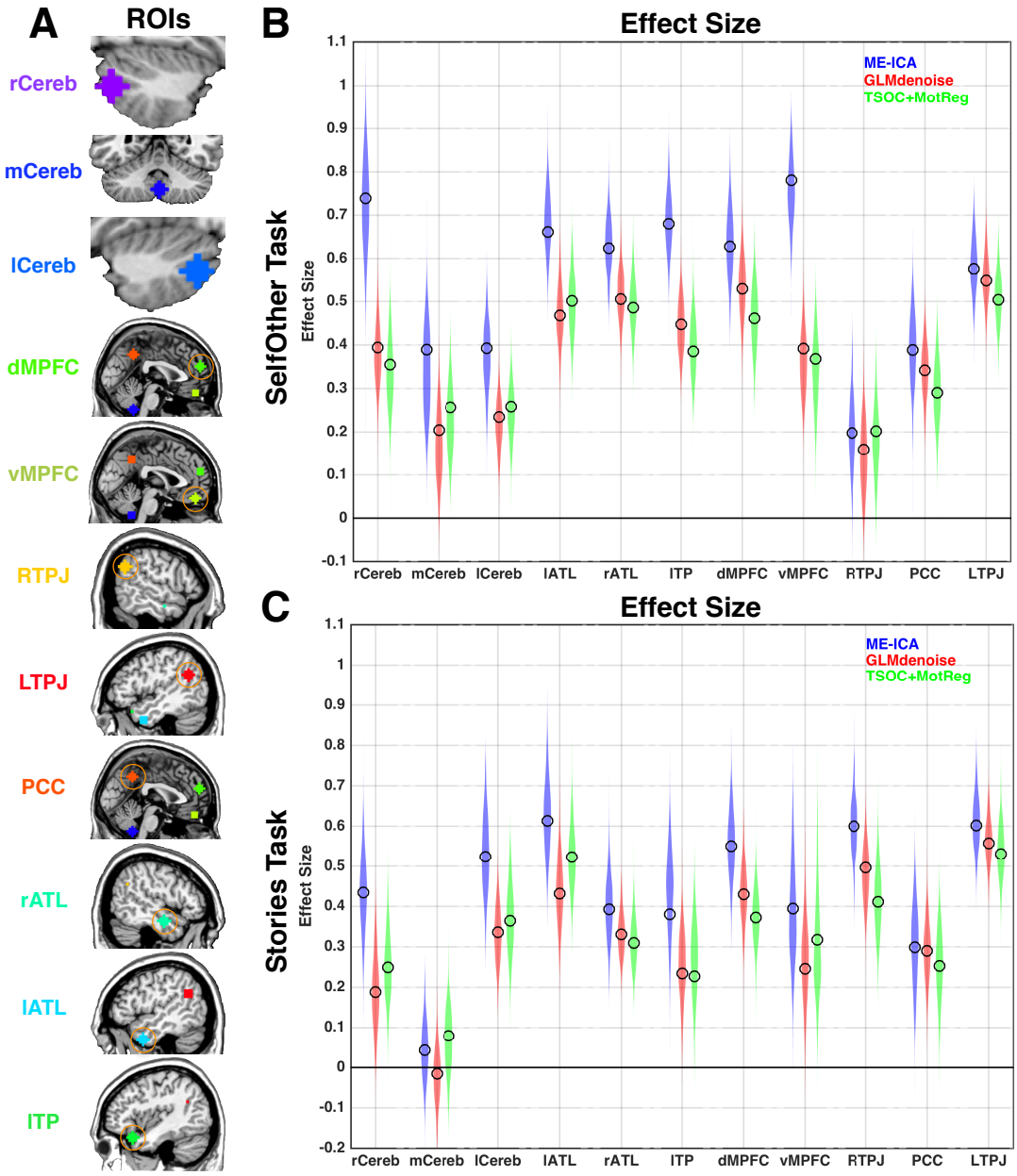

Fig 2: ME-ICA Effect Size Boosting. This figure shows effect size estimates (panels B-D) from all regions of interest (panel A). Effect sizes are expressed in standard deviation units and are analogous to Cohen's d. Colored clouds in each plot represent density of estimates obtained from 1000 bootstrap resamples, while unfilled circles represent estimates within the true dataset.

Because our operational definition of effect size is a standardized measure that incorporates both mean and variability measurements, we went further in decomposing how these boosts in effect size estimation manifested in terms of changes to either the mean and/or variability measurements. It is clear from Fig 3 that ME-ICA induces these boosts primarily by reducing estimates of variability at the 2nd level group analysis. Given that at a within-subject level ME-ICA is working to remove non-BOLD noise from the time series, it is clear that one consequence of this for group-level modeling is clear reduction of between-subject variance which works to enhance standardized effect size estimates.

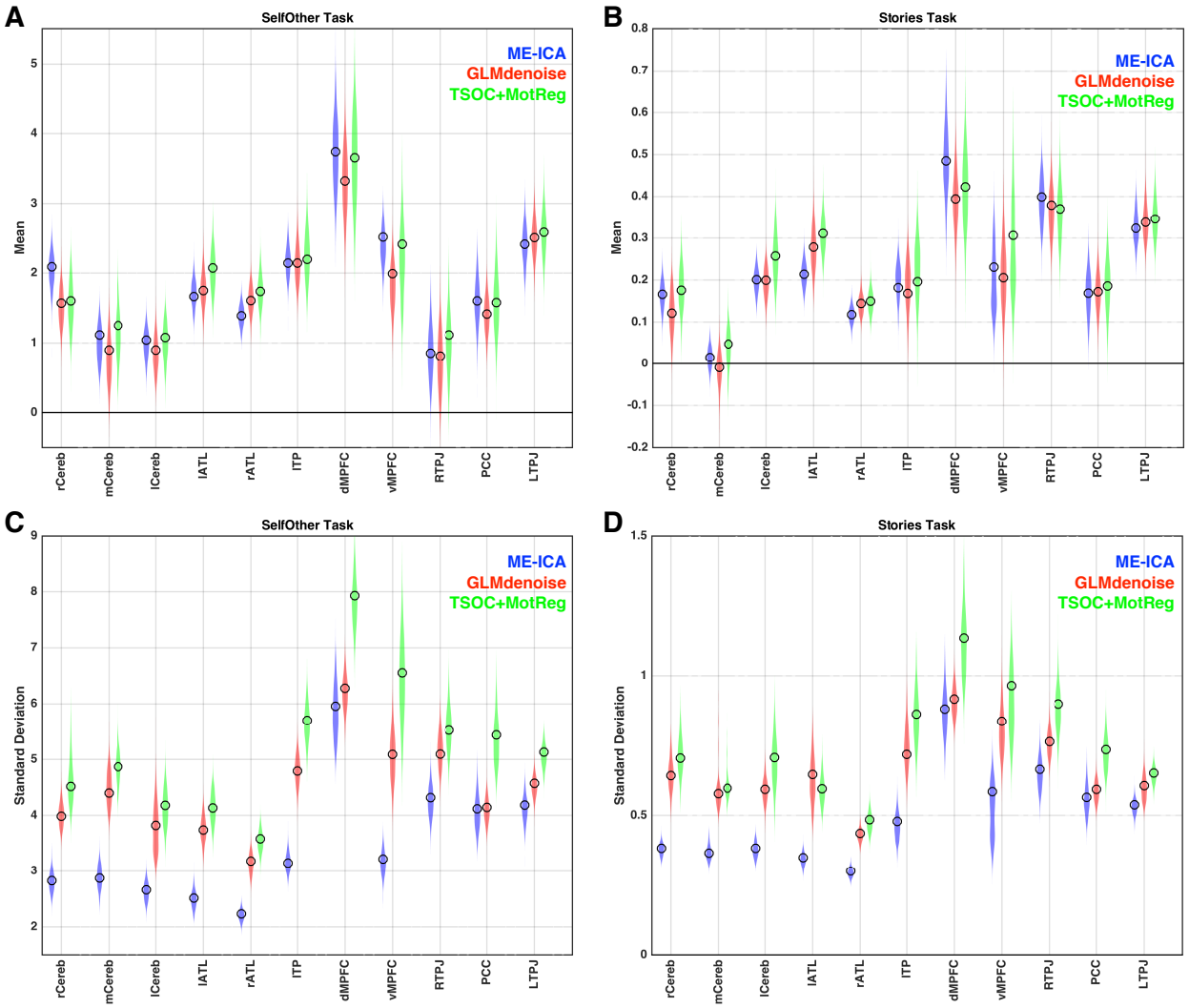

Fig 3: ME-ICA Reduction in Variance in Group-Level Analyses. This figure shows mean and standard deviation estimates from 2nd level group modeling that contribute to the standardized effect size calculations. Panels A and B show mean estimates for all regions in both tasks. Panels C and D show standard deviation estimates. Colored clouds in each plot represent density of estimates obtained from 1000 bootstrap resamples, while unfilled circles represent estimates within the true dataset.

Impact of ME-ICA on Statistical Power

Because ME-ICA improves standardized effect size estimation, it necessarily follows that statistical power will also be boosted, as such estimates are critical in such computations. However, for assessing the practical impact that ME-ICA may have, it is necessary to assess the impact such effect size boosting has on statistical power and sample size. Here we describe power simulations that mainly inform what we could expect in future work given effect size estimates similar to what we have observed in the current study under ME-ICA versus other analysis pipelines.

Power curves for each analysis pipeline across a range of sample sizes from n=5 to n=100 are illustrated in Fig 4A-B. Minimum sample size necessary for achieving 80% power at an alpha of 0.05 are shown in Fig 4C. Across all canonical regions and both tasks, the median minimum sample size to achieve 80% power at an alpha of 0.05 with ME-ICA is n=19. Minimum sample sizes across nearly all regions were well within reach of current standards for sample size (e.g., n<45). In contrast, for GLMdenoise and TSOC+MotReg the median minimum sample size for canonical regions is n=33 and n=41 respectively and there were several important regions whereby n>45 is necessary.

For cerebellar regions, the power benefits due to ME-ICA were even more pronounced. Aside from medial cerebellar region XI (mCereb) in the Stories task which did not result in a sizeable effect (e.g., effect size <0.1), the minimum sample size needed for the bilateral cerebellar Crus I/II areas (rCereb, lCereb) were always well within the a range of sample size that is typical for today's standards when using ME-ICA (e.g., n<45). This stands in contrast to the situation for GLMdenoise and TSOC+MotReg, where sample size always required n>40 and in many instances was not attained by n=100.

For further illustration of practical impact, these boosts in statistical power and reduction in sample size necessary for achieving 80% power can be quantified into monetary savings. Assuming a scan rate of $500 per individual, if one was only interested in canonical regions, using ME-ICA would amount to median savings of $6,000 compared to GLMdenoise and $10,500 compared to TSOC+MotReg. If one was interested in cerebellar regions, using ME-ICA would save $15,500 compared to GLMdenoise and $26,000 compared to TSOC+MotReg.

Visual examination of the power curves in Fig 4A-B highlights a point of diminishing returns when power is greater than 95%, as the improvements in power for adding more subjects diminishes substantially. We term this effect 'saturation'. When using ME-ICA, many regions quickly reach these saturation levels at sample sizes that are practically attainable (e.g., n<45). In contrast, other pipelines like GLMdenoise and TSOC+MotReg typically require considerably larger sizes to hit these saturation levels.

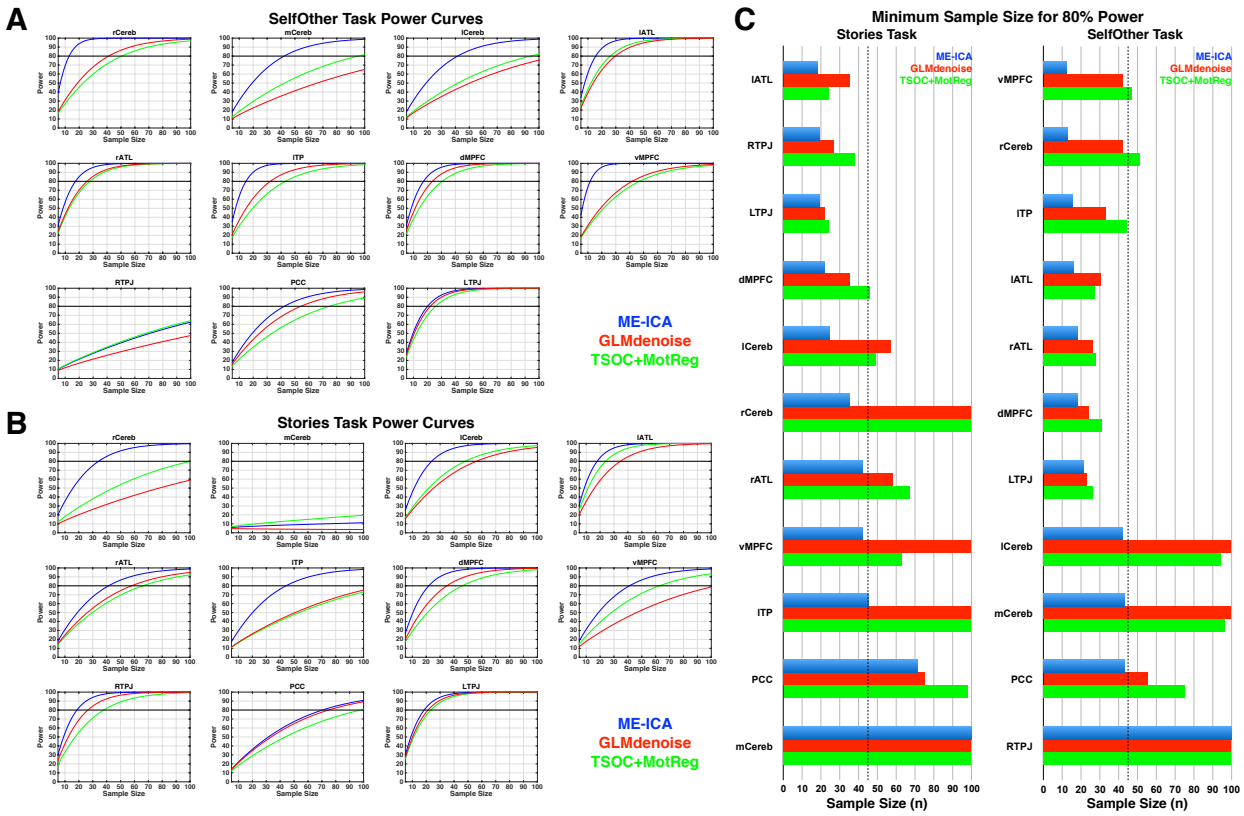

Fig 4: Power Simulations. This figure shows power curves constructed for each processing pipeline across a range of sample sizes from 5 to 100 (panels A-B). The minimum sample size necessary for achieving 80% power is shown in panel C for the Stories task (left) and SelfOther task (right). The dotted line indicates sample size of n=45.

Functional Connectivity Evidence for Cerebellar Involvement in Neural Systems Supporting Mentalizing

The improvements in effect size estimation particularly for cerebellar regions 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[20]. The default mode network incorporates many of the regions that are highly characteristic in task-evoked systems supporting mentalizing[21]. Meta-analytically defined cerebellar regions associated with mentalizing show some overlap with these cerebellar default mode areas[22]. 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 scanned on 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).

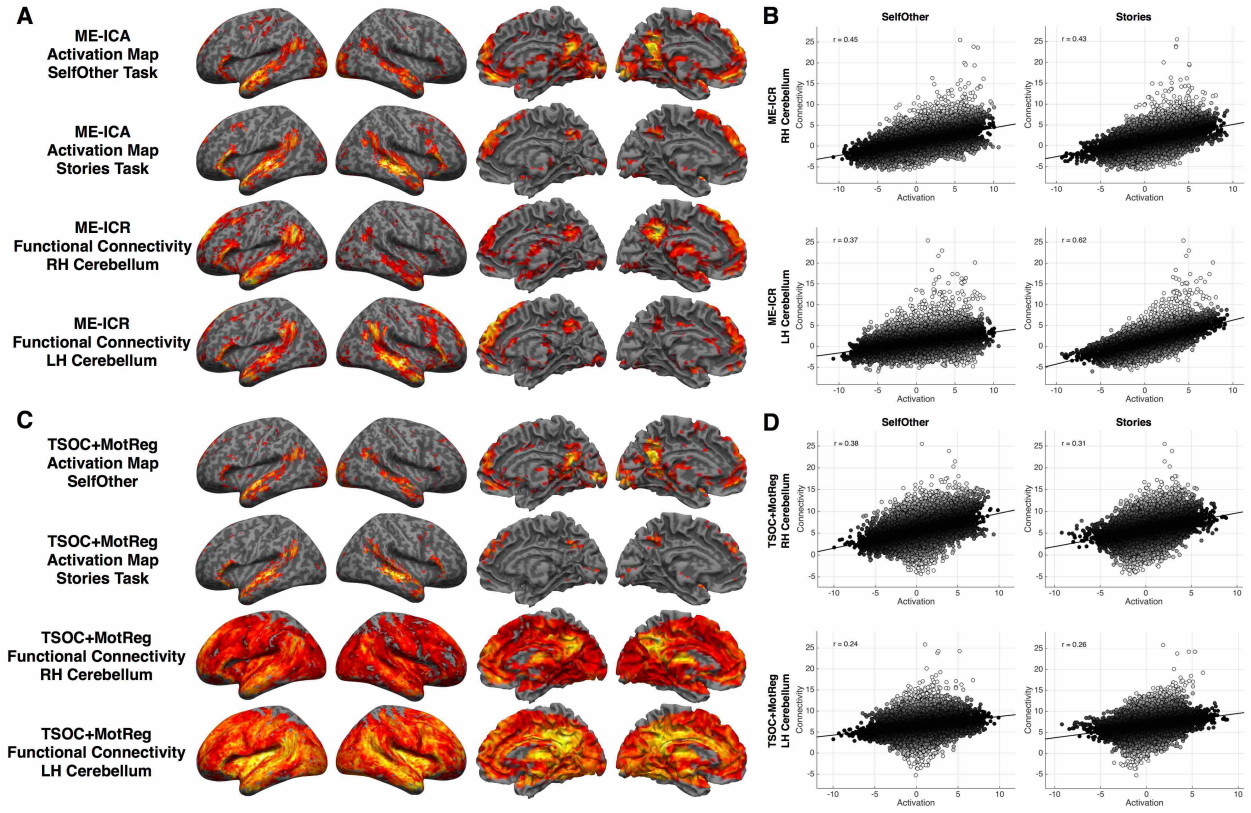

Fig 5: Resting state functional connectivity from cerebellar seed regions and pattern similarity with Mentalizing>Physical activation maps. This figure shows resting state connectivity from right and left cerebellar seed voxels (i.e. peak voxels from the NeuroSynth 'mentalizing' map) and their similarity 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[23]). 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.

Discussion

Task-based fMRI studies are characteristically of small sample size and thus underpowered for all but the largest and most robust effects. 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 show that our methodological innovation, ME-ICA, results in robust increases in effect size estimation and statistical power in block-design studies. These improvements are empirically demonstrated against other prominent denoising alternatives. As a consequence of these improvements in effect size and power, we also demonstrate application of this method towards identification of novel effects in the cerebellum involved in the neural systems supporting mentalizing. Assuming similar effect sizes in future work, power simulations suggest that discovery of these novel cerebellar effects will remain nonetheless hidden at characteristically small sample sizes and without the multi-echo denoising innovations we report here.

There are several practical points of impact that these results underscore. First, addressing the problem of statistical power in neuroscience, particularly fMRI studies[1,2], is a complicated matter as most recommendations for this problem rely on increasing the amount of data collected both at the within and between-subject levels. A practical barrier for most research labs however, is that increasing the scale of data collection (e.g. massive sample size studies) is typically cost prohibitive. Our innovations here take a different perspective on the problem of low statistical power, by addressing from the bottom up, the problem of non-BOLD noise, which directly has impact on the sensitivity of fMRI, and thus statistical power. In practical terms, we show that ME-ICA allows for such substantial boosts in effect size estimation and consequently statistical power whereby in most cases (i.e. canonical and cerebellar regions investigated here), requisite levels of statistical power are attainable at sample sizes that should not be out of reach for most research laboratories. Therefore, if in the future researchers were to take up our multi-echo innovations in combination with uptake of already prominent considerations to generally collect more data, we could envision that the situation for fMRI research could substantially improve.

It is particularly important to underscore here that we are not suggesting that ME-ICA is the panacea to the small sample size problem and that as a result, researchers could continue the tradition of small sample size studies. Rather, we advocate that there are always compelling reasons to collect more data and that if funds permit, researchers should go above and beyond data collection that will ensure that their studies are highly powered at traditional sample sizes. Such a situation will ensure that canonical large effects are robust and replicable. Moreover, boosts in the sensitivity of fMRI can open up a range of previously practically unattainable possibilities for new discoveries. Such new discoveries could take the form of much more enhanced sensitivity for detecting smaller and more subtle effects in brain regions that are currently not well understood or which are methodologically hampered by being continually veiled underneath blankets of non-BOLD noise. New discoveries could also be enabled with parsing apart further variability such as subgroups that may have important translational implications[24], parsing apart heterogeneity mapped onto individual differences, and/or more fine grained hypotheses/methods that result in much smaller effects than the typical and more basic activation mapping paradigm. All of these situations could be substantially improved with a methodological approach that dramatically improves statistical power, but at the same time promotes and motivates researchers to collect larger samples than what is typically characteristic.

As an empirical demonstration of ME-ICA's ability to enhance new discoveries for human brain functional organization, we have uncovered robust evidence that there are discrete cerebellar regions that should hold more prominence in discussions about the neural systems supporting mentalizing/theory of mind and the 'social brain'. The cerebellum is already a neglected and not well understood brain area, particularly in the context of its potential role in higher-level cognition[19,25-28]. Prior indications that these cerebellar regions may be plausible candidates for neural systems supporting mentalizing come from meta-analytic evidence[19]. However, while meta-analytic evidence alone might suggest plausibility for these regions, it was still unclear as to the exact reasons for why these cerebellar regions have not been the topic of more extensive focus.

In this study, one of the novel findings that may help explain why these cerebellar regions are missed, is that they are typically veiled in substantial amounts of non-BOLD noise that obscure researcher's ability to detect such effects with traditional types of methods and analysis pipelines. Effect sizes for these regions under more traditional analysis approaches are typically small and the sample size necessary for detecting those effects with high power are much greater than what is typical for fMRI research. However, after applying ME-ICA innovations, these effects are substantially boosted by 43-130%. As we show in this study, ME-ICA primarily boosts effect size estimation via noise reduction at the within-subject level and consequently has impact for reduction of variance at the group level. Therefore, it is clear that these regions are typically highly saturated in non-BOLD noise and this problem helps to obscure these effects from traditional research practices of small sample sizes, usage of single-echo EPI acquisition, and denoising procedures that do not fully identify and remove such noise variability.

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[29,30]. Translationally, the link between cerebellum and mentalizing is also particularly intriguing, given the longstanding, yet independent, literatures in autism regarding the cerebellum[31] and mentalizing[32]. Wang and colleagues[28] 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[33] and impairments in the development of mentalizing/theory of mind and self-referential cognition in autism[34,35] as well as atypical functioning of neural mechanisms that bolster such processes[36,37] 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.

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[38], 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[39].

The multi-echo innovations we provide here offer substantial improvements that can largely affect how the field conducts fMRI research. All of the tools for implementing these innovations are open source and most contemporary imaging facilities possess all the requisite requirements to enable actively taking up these innovations as standard practice. We hope that the community will actively take up these new innovations, as they are likely to have massive benefits for improving major issues that hamper the field and may further enable potential for new discoveries about human brain function.

References

1 Button, K. S. et al. Power failure: why small sample size undermines the reliability of neuroscience. Nature reviews. Neuroscience 14, 365-376, doi:10.1038/nrn3475 (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, doi:10.1371/journal.pmed.0020124 (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, doi:10.1177/0956797611417632 (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, doi:10.1016/j.neuroimage.2007.07.061 (2008).

7 Friston, K. Ten ironic rules for non-statistical reviewers. Neuroimage 61, 1300-1310, doi:10.1016/j.neuroimage.2012.04.018 (2012).

8 Lindquist, M. A., Caffo, B. & Crainiceanu, C. Ironing out the statistical wrinkles in "ten ironic rules". Neuroimage 81, 499-502, doi:10.1016/j.neuroimage.2013.02.056 (2013).

9 Murphy, K., Birn, R. M. & Bandettini, P. A. Resting-state fMRI confounds and cleanup. Neuroimage 80, 349-359, doi:10.1016/j.neuroimage.2013.04.001 (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, doi:10.1016/j.neuroimage.2011.12.028 (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, doi:10.3389/fnins.2013.00247 (2013).

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

13 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).

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

15 Lombardo, M. V. et al. Shared neural circuits for mentalizing about the self and others. J Cogn Neurosci 22, 1623-1635, doi:10.1162/jocn.2009.21287 (2010).

16 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, doi:10.1016/j.neubiorev.2014.01.009 (2014).

17 Spunt, R. P. & Adolphs, R. Validating the Why/How contrast for functional MRI studies of Theory of Mind. Neuroimage 99, 301-311, doi:10.1016/j.neuroimage.2014.05.023 (2014).

18 Schaafsma, S. M., Pfaff, D. W., Spunt, R. P. & Adolphs, R. Deconstructing and reconstructing theory of mind. Trends Cogn Sci 19, 65-72, doi:10.1016/j.tics.2014.11.007 (2015).

19 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, doi: 10.1016/j.neuroimage.2013.09.033 (2014).

20 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, doi:10.1152/jn.00339.2011 (2011).

21 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, doi:10.1016/j.neuroimage.2014.01.032 (2014).

22 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, doi:10.1080/17470919.2015.1005666 (2015).

23 Kundu, P. et al. Integrated strategy for improving functional connectivity mapping using multiecho fMRI. Proc Natl Acad Sci U S A 110, 16187-16192, doi:10.1073/pnas.1301725110 (2013).

24 Lombardo, M. V. et al. Different functional neural substrates for good and poor language outcome in autism. Neuron, doi:10.1016/j.neuron.2015.03.023 (2015).

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

26 Stoodley, C. J. & Schmahmann, J. D. Functional topography in the human cerebellum: a meta-analysis of neuroimaging studies. Neuroimage 44, 489-501, doi:10.1016/j.neuroimage.2008.08.039 (2009).

27 Buckner, R. L. The cerebellum and cognitive function: 25 years of insight from anatomy and neuroimaging. Neuron 80, 807-815, doi:10.1016/j.neuron.2013.10.044 (2013).

28 Wang, S. S., Kloth, A. D. & Badura, A. The cerebellum, sensitive periods, and autism. Neuron 83, 518-532, doi:10.1016/j.neuron.2014.07.016 (2014).

29 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).

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

31 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, doi:10.1056/NEJM198805263182102 (1988).

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

33 Lai, M. C., Lombardo, M. V. & Baron-Cohen, S. Autism. Lancet, doi:10.1016/S0140-6736(13)61539-1 (2013).

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

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

36 Lombardo, M. V. et al. Atypical neural self-representation in autism. Brain 133, 611-624, doi:10.1093/brain/awp306 (2010).

37 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, doi:10.1016/j.neuroimage.2011.02.067 (2011).

38 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, doi:10.1016/j.neuroimage.2014.10.051 (2015).

39 Schilbach, L. et al. Toward a second-person neuroscience. Behav Brain Sci 36, 393-414, doi:10.1017/S0140525X12000660 (2013).

Online Methods

Participants

Participants were 69 adolescents (34 males, 35 females, mean age = 15.45 years, sd age = 0.99 years, range = 13.22-17.18 years) sampled from a larger cohort of individuals whose mothers underwent amniocentesis during pregnancy for clinical reasons (i.e. screening for chromosomal abnormalities). The main focus for sampling from this cohort 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 individual 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[1-3]. 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[4,5]. 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 seconds 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 from the study by Gweon and colleagues[6] and we utilized the exact same stimuli and stimulus presentation scripts provided to us by 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[7,8] 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. In contrast to ME-ICA, we also ran denoising with two other prominent approaches; GLMdenoise9 and TSOC+MotReg. Each of these pipelines were fed TSOC data. For GLMdenoise, we extract global noise regressors it identified with cross validation and used these as regressors of no interest in first-level individual subject GLMs. For TSOC+MotReg we used motion parameters as regressors of no interest in first-level individual subject GLMs. When running first-level individual subject GLMs on ME-ICA denoised data, we did not include motion parameters as regressors of no interest because such artifact is already removed in principled manner at the prior denoising step. All first-level individual subject GLMs modeled the specific contrast of Mentalizing>Physical, and these contrast images were input into second-level random effects GLM (i.e. one sample t-test). Any whole-brain second-level group analyses we report are thresholded at a voxel-wise FDR $q<0.05$[10].

Resting State fMRI Connectivity Analysis

Resting state connectivity on ME-ICA processed data was estimated using the multi-echo independent components regression (ME-ICR) technique developed by Kundu and colleagues[7]. 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 the 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[11] 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)[12]. Effect size is operationalized here as a standardized measure of distance from 0 expressed in standard deviation units (i.e. mean/sd) and is 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)[13] 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[14].

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 ($ES_{ME\text{-}ICA}$ - $ES_{TSOC\ or\ GLMdenoise}$)/ abs($ES_{TSOC\ or\ GLMdenoise}$) * 100. We also 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 estimates. To further describe the effects of ME-ICA over and above GLMdenoise 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 $500. In cost savings computations, any regions that did not achieve requisite power before n=100 were excluded from such calculations.

References

1 Lombardo, M. V. et al. Atypical neural self-representation in autism. Brain 133, 611-624, doi:10.1093/brain/awp306 (2010).

2 Lombardo, M. V. et al. Shared neural circuits for mentalizing about the self and others. J Cogn Neurosci 22, 1623-1635, doi:10.1162/jocn.2009.21287 (2010).

3 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, doi:10.1016/j.neuroimage.2011.02.067 (2011).

4 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).

5 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).

6 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, doi:10.1111/j.1467-8624.2012.01829.x (2012).

7 Kundu, P. et al. Integrated strategy for improving functional connectivity mapping using multiecho fMRI. Proc Natl Acad Sci U S A 110, 16187-16192, doi:10.1073/pnas.1301725110 (2013).

8 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, doi:10.1016/j.neuroimage.2014.10.051 (2015).

9 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, doi:10.3389/fnins.2013.00247 (2013).

10 Genovese, C. R., Lazar, N. A. & Nichols, T. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. Neuroimage 15, 870-878, doi:10.1006/nimg.2001.1037 (2002).

11 Wager, T. D., Keller, M. C., Lacey, S. C. & Jonides, J. Increased sensitivity in neuroimaging analyses using robust regression. Neuroimage 26, 99-113, doi:S1053-8119(05)00036-4 [pii]10.1016/j.neuroimage.2005.01.011 (2005).

12 Mumford, J. A. & Nichols, T. E. Power calculation for group fMRI studies accounting for arbitrary design and temporal autocorrelation. Neuroimage 39, 261-268, doi:10.1016/j.neuroimage.2007.07.061 (2008).

13 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, doi:nmeth.1635 [pii] 10.1038/nmeth.1635 (2011).

14 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, doi: 10.1016/j.neuroimage.2013.09.033 (2014).