Supplementary Information for

Different functional neural substrates for good and poor language outcome in autism

Michael V. Lombardo, Karen Pierce, Lisa Eyler, Cindy Carter Barnes, Clelia Ahrens-Barbeau, Stephanie Solso, Kathleen Campbell, Eric Courchesne

Supplementary Methods and Results

Experimental Procedures

Participants

This study was approved by the University of California, San Diego Institutional Review Board. Parents provided written informed consent according to the Declaration of Helsinki and were paid for their participation. A total of 103 toddlers participated in this study. These toddlers were part of a larger longitudinal study of toddlers at-risk for autism (age range 12-48 months) who were enrolled in neuroimaging studies at our center (www.autism-center.ucsd.edu). All toddlers were recruited in the same way via general community referral (e.g. website or outside agency) and a population-based screening method called the 1-Year Well-Baby Check-Up Approach (Pierce et al., 2011). Using this method, infants and toddlers at-risk for ASD, at-risk for a language (LD) or other developmental delay (DD), or not at risk for a developmental disorder were identified in pediatric offices with a broadband screening instrument, the Communication and Symbolic Behavior Scales-Developmental Profile Infant Toddler Checklist (Wetherby et al., 2002), and were recruited and tracked until 3 to 4 years of age. This method thus allowed for the prospective study of autism and LD/DD as compared with typical toddlers from ages as young as 1 and 2 years.

ASD diagnoses were made by multiple PhD-level licensed clinical psychologists with extensive experience in autism using the three initial modules of the Autism Diagnostic Observation Schedule (toddler, 1, or 2) (Lord et al., 2000; Luyster et al., 2009), the Mullen Scales for Early Learning (Mullen, 1995), and the Vineland Scales of Adaptive Behavior (Sparrow et al., 1984). Diagnoses were based on based on clinical judgment (i.e., a DSM-IV checklist as well as the clinician's overall diagnostic impression formed after observing the child during the 3-4 hour test visit) and ADOS scores. In all toddlers, behavioral exams were performed within 3 months of the fMRI scan (typically they were performed within the same week). The diagnosis of toddlers with ASD who were younger than 24 months at the time of the scan was also confirmed at later ages. For the purposes of this study, language ability and outcome was operationalized based on receptive and expressive language (EL or RL) T-scores from the Mullen Scales of Early Learning. Because the norm for each Mullen subtest is set to a T-score of 50, ±10 as 1 standard deviation, we used this as a criteria for classifying individuals into language outcome groups based on their EL and RL T-scores from the outcome testing time-point (generally about a year after assessment of brain function). Furthermore, in this study we assessed all individuals irrespective of diagnosis longitudinally with the Mullen, Vineland, and ADOS and attempted measurements approximately every 6 months since the intake assessment, although this was not always possible.

Of those 103 infants and toddlers in this study, 60 retained a diagnosis of ASD across longitudinal evaluations towards their outcome testing time-point. Further stratification of the ASD group was achieved by dividing individuals into 2 language outcome groups. Twenty-four individuals with ASD were classified as 'poor' language outcome (ASD Poor), based on the criteria of having both Mullen EL *and*

RL T-scores more than 1 standard deviation below the norm of 50 (i.e. T<40) at the outcome testing time-point (n = 24, 19 male, 5 female; mean age at fMRI scan = 27.13 months, SD at fMRI scan = 7.92, range = 12-47 months). Another 36 individuals with ASD were classified as 'good' language outcome (ASD Good), based on having either Mullen EL *or* RL T-scores greater than or equal to 40 (i.e. T ≥ 40) at outcome (n = 36, 28 male, 8 female; mean age at fMRI scan = 25.38 months, SD at fMRI scan = 8.69, range = 11-39 months). The usage of the term 'Good' here is not used to refer to ability level in absolute terms, but more reflects ability relative to the ASD Poor subgroup.

Two comparison groups were included in addition to the ASD subgroups. The first consisted of 24 typically-developing toddlers (TD) with no neurodevelopmental diagnosis and no family history of ASD or any other neurodevelopmental disorders. Typically-developing comparison participants were obtained from community and pediatrician referrals and were matched to the poor and good ASD language outcome groups on sex (19 male, 5 female) and age at fMRI scan (mean age at fMRI scan = 27.02 months, SD at fMRI scan = 9.69, range = 13-44 months). The second comparison group consisted of 19 toddlers with language and/or developmental delays (LD/DD; 17 male, 2 female) but did not meet criteria for diagnosis of an autism spectrum disorder (mean age at fMRI scan = 21.40 months, SD = 8.46, range = 12-39 months). Of the 19 LD/DD individuals, 13 were classified as 'language delay' (LD) and 6 were classified as 'developmental delay' (DD). At intake, within the domain of language measures (i.e. Mullen EL and RL T-scores), DD was similar to the LD on EL (DD mean = 32.5, SD = 15.06; LD mean = 37.15, SD = 9.69; $t(17) = 0.81$, $p = 0.42$) and was substantially lower on RL (DD mean = 30.66, SD = 7.89; LD mean = 42.07, SD = 8.22; $t(17) = 2.84$, $p = 0.011$). In addition to the fact that these scores are all well below T = 50 on the Mullen, which reflects norms of age-appropriate ability in typical development, the fact that the DD group also shows substantial delays within the domain of language demonstrates that all individuals within this combined LD/DD group are delayed in language development. Thus as a comparison group to ASD, this group offers another non-ASD comparison group, but where developmental delays particularly in the domain of language development are present.

Demographic and descriptive statistics for clinical measures for all groups at intake and outcome time-points are presented in Table S1. In depth analysis of group-differences in developmental trajectories and main effects of group are presented within the main text of the manuscript. However, here we present some initial analyses to assess similarities or differences between groups on main demographic characteristics such as age at fMRI. An ANOVA showed no significant effect of group on age at fMRI scan ($F(3,99) = 1.92$, $p = 0.13$, *partial* $\eta^2 = 0.055$), although the p-value and effect size suggests a trend for a difference that may be driven by the LD/DD being slightly younger than the other groups. Therefore, in all fMRI analyses, we included age at fMRI scan as a covariate.

Behavioral Data Analysis

All behavioral analyses reported in the main text of the manuscript employ mixed-effect analyses in order to model within-individual trajectories and group-level

trajectories. These analyses were implemented within R and used the *lme* function from the *nlme* package (http://cran.r-project.org/web/packages/nlme/index.html) for mixed-effect modeling. Main analyses consisted of mixed-effect ANOVAs (modeling random slopes and intercepts) and were followed up by post-hoc pairwise group comparisons. Given the 6 pairwise post-hoc comparisons, results were only deemed significant if they passed a Bonferroni-corrected alpha threshold of 0.0083.

fMRI Task Design and Behavioral Measures

The fMRI task was identical to that used in our previously published studies (Eyler et al., 2012; Redcay and Courchesne, 2008; Redcay et al., 2008) and consisted of three types of stimuli, presented in 20 second blocks: complex forward speech, simple forward speech and backward speech. All speech conditions were created using the same female speaker. During the complex forward speech condition, toddlers were exposed to segments of a children's story that was written at a comprehension level of over 48 months. During the simple forward speech condition, toddlers were exposed to excerpts from a children's story written at a comprehension level between 12 and 36 months. Finally, during the backward speech condition, toddlers were exposed to the simple story segments played backwards. There were also 20 second rest blocks (no presented stimuli) between each stimulus type. Each stimulus type was repeated three times in a pseudorandom order. The total task length was 6 min 25 seconds. The stimuli were presented using commercially available music presentation software with maximum volume set both for the software and the computer's speakers. Stimulus presentation was through pneumatic headphones (Confon, Inc.) set to a volume attenuation of -40 dB for all participants.

Stimuli

The comprehension levels for the simple and complex speech conditions were defined informally with the simple speech excerpts taken from a toddler's board book, "Its Time for Sleep" written by Mem Fox. Words in this book were simple, common words well known by toddlers (most listed in the McArthur Bates CDI) and were almost exclusively 1 and 2 word syllable words. For example, one line reads: "Its time for bed, little sheep, little sheep. The cows are out and fast asleep." The "complex" speech excerpts were taken from a book of poetic verses for older children called "A Child's Garden of Verses" written by Robert Luis Stevenson. The words were more complex and some of the words contained 3 or more syllables. For example, one line reads "I rose and found the dew on every buttercup."

fMRI Data Acquisition

Infants and toddlers were imaged in a 1.5 Tesla General Electric MRI scanner during natural sleep at night; no sedation was used. Parents were encouraged to forgo any usual naps by the child and engage the child in rigorous physical activity during the day. Families arrived at the scanning facility 1 h after the child's typical bedtime and most children had been asleep in the car for ~15 min prior to arrival. If not already asleep or if awakened by placement on the scanner bed, the child was allowed

to fall asleep in the waiting or scanning room and was placed on the scanner bed after ~15 min of sleep. The time that the child fell asleep was recorded for every participant. After placement of the headphones, padding of the head for comfort and motion reduction and covering the child with a weighted blanket for warmth and motion reduction, the scan was started. A research assistant was present in the room next to the child during the entire scan and stopped the scan if the child woke up or made a large movement.

The order of scans varied somewhat between individuals, but all participants first received a high-resolution, T1-weighted anatomical scan (repetition time=6.5, flip angle=12°, bandwidth=31.25, field of view = 24 cm, in-plane resolution = 1 x 1 mm, slice thickness = 1.2 mm, 170 slices, scan length = 7 min 24 s) for localization of functional signals and warping into standard atlas space. On average, the speech task was presented 18.3 ± 13.3 min after the onset of scanning. Other functional and anatomical scans were acquired before and after the speech task; data from these will not be presented in this article. If the child remained asleep for the entire procedure including these other scans, the total scan time was 1 h 15 min.

Blood oxygenation level-dependent signal was measured across the whole brain with echoplanar imaging during the language paradigm. Scan parameters were: echo time = 30 ms, repetition time = 2500 ms, flip angle = 90°, bandwidth = 70 kHz, field of view = 25.6 cm, in-plane resolution = 4 x 4 mm, slice thickness = 4 mm, 31 slices.

fMRI Data Analysis

Preprocessing of functional imaging data was implemented within the Analysis of Functional NeuroImages (AFNI) software package. The preprocessing pipeline was comprised of motion correction, normalization to Talairach space, and smoothing (8mm full-width at half-maximum (FWHM) Gaussian kernel). To examine head motion across the groups in more detail, we used the metric of framewise displacement (FD) quantified in millimeters (Power et al., 2012). Because of positively skewed distributions in mean FD, a Kruskal-Wallis one-way non-parametric ANOVA was used. We found significant between-group differences in mean FD ($\chi^2(3,99) = 11.65$, $p = 0.0087$), with this difference being driven primarily by outliers in LD/DD compared to ASD Poor (Fig S5A). Given this difference and the additional known issues with the influence of head motion on BOLD signal variation, we made attempts in the analysis to regress out motion parameters in 1st level analyses and regress out mean FD in 2nd level analyses (similar to the recommendations made by Yan and colleagues) (Yan et al., 2013).

In addition to covarying out mean FD, we also ran a further analysis on all individuals whom had mean FD less than or equal to 0.1mm (number excluded: n=5 TD; n=7 LD/DD, n=7 ASD Good; n=0 ASD Poor). This range restriction on head movement was sufficient for removing the group-difference in mean FD ($\chi^2(3,80) = 5.07$, $p = 0.16$) (Fig S5B). We then re-ran the primary ANCOVA on the NeuroSynth left-hemisphere temporal ROI and found that in this motion control analysis, that the group difference remained and that the effect size was larger than in the analysis on all individuals ($F(3,78) = 5.36$, $p = 0.002$, *partial* $\eta^2 = 0.171$). Thus, this additional

motion control analysis further demonstrates that the primary results are robust to contamination due to motion artifact.

First-level mass-univariate whole-brain activation analyses were conducted using the general linear model in SPM8 (http://www.fil.ion.ucl.ac.uk/spm/). Events were modeled using the canonical hemodynamic response function and its temporal derivative. All first-level GLMs included motion parameters as covariates of no interest. High-pass temporal filtering was applied with a cutoff of 0.0078 Hz (1/128 seconds) in order to remove low frequency drift in the time-series. Contrast images were computed for the general contrast of all speech conditions vs. rest as well as the contrast of forward vs backward speech for all first-level analyses. Within second-level analyses, we covaried for age at fMRI scan and mean FD. For ROI analyses, we used independent functionally-defined ROIs related to language processing by extracting bilateral frontal and temporal cortex clusters from a meta-analysis map downloaded from www.neurosynth.org (Yarkoni et al., 2011) of 725 studies associated with the feature 'language' (see top of Fig 2E). These ROIs were further constrained to only voxels labeled as frontal or temporal cortex in the MNI atlas packaged with the FSL software package (http://fsl.fmrib.ox.ac.uk/fsl/). All ROI analyses were conducted as ANCOVAs covarying for mean FD and age at scanning. For any ROIs where there was indication of violation of homogeneity of variance via Levene's test, we re-ran such analyses using Welch's ANOVA (i.e. using the *oneway.test* function in R). For all second-level whole-brain analyses we used a cluster-forming threshold of $p<0.025$ and corrected for multiple comparisons at the cluster-level to obtain an FDR $q<0.05$ (Chumbley et al., 2010).

Multi-Voxel Pattern Similarity Analyses

To quantitatively assess the similarity of multi-voxel activation patterns estimated at the group-level, we computed Pearson's r correlation between all group's second-level t-statistic map, masked for only voxels within the NeuroSynth language map (n=820 voxels). These second-level activation pattern similarity results were visualized as a correlation matrix, and it was visually clear that TD, ASD Good, and LD/DD groups were all highly similar in activation patterns, and ASD Poor was markedly most dissimilar to these other groups. In order to better visualize the separation of ASD Poor from all other groups, the resulting similiarity (correlation) matrix was first converted into a dissimilarity matrix (i.e. 1-r) and then entered into canonical multidimensional scaling to reduce dimensionality down to a 2-dimensions. To specifically test whether ASD Poor was highly dissimilar, we converted Pearson's r correlations to z-statistics using Fisher's z transformation (using the *paired.r* function within the *psych* package in R: http://bit.ly/1xtgf9i)), and then assessed statistical significance of the difference between all pairwise group-comparisons. Of the 12 total pairwise comparisons, the 6 comparisons involving a difference between a [ASD Poor vs non-ASD Poor] correlation vs a [non-ASD Poor vs non-ASD Poor] correlation, were the top 6 most highly enriched comparisons with all $p < 5.64 \times 10^{-14}$ (Table S4).

We also compared multi-voxel pattern similarity of individual subject activation maps (i.e. 1st-level t-statistic maps for all speech vs rest) to specific NeuroSynth feature maps constructed specifically for the purposes of isolating neural

systems for either general auditory processing outside the domain of language and speech or isolating neural systems that are specific to language and speech. To achieve these aims we used NeuroSynth core tools in Python (https://github.com/neurosynth/neurosynth), to run 2 meta-analyses not found on the general NeuroSynth website. The first includes only studies associated with the term 'auditory' but excludes any of those studies which also appear for the terms 'language' or 'speech' (n=601 total studies; 'Auditory AND NOT (Language OR Speech)'). This meta-analysis highlights neural systems for general auditory processing, excluding the influence of studies highlighted in the 'language' or 'speech' features. The second meta-analysis was conducted on to isolate neural systems from studies associated with both the terms 'language' and 'speech' (n=145 total studies; 'Language AND Speech'). This meta-analysis highlights neural systems found in studies that highly refer to both the terms 'language' and 'speech'. These meta-analyses produced whole-brain z-statistic maps that we could then compare against each individual subject's activation map. For each individual, we computed Pearson's r correlation for activation similarity within voxel masks defined by each meta-analysis (i.e. only voxels from the meta-analyses that survive at FDR $q<0.01$). We then tested for group-differences using ANCOVAs, covarying for mean FD and age at scanning.

Supplementary Analysis of Primary Auditory Cortex

Supplementary analysis of activation within primary auditory cortex was done by first defining a meta-analytic ROI of primary auditory cortex. This was achieved by using NeuroSynth core tools in Python to run a meta-analysis on the combination of all studies from the features 'auditory cortex', 'auditory cortices', and 'primary auditory'. Then we identified the peak voxel from this meta-analysis in left hemisphere superior temporal gyrus (MNI x = -48, y = -22, z = 6; z-stat = 22.91) and constructed a 10mm sphere around this centroid. This spherical peak ROI was then used for extracting percent signal change from the all speech vs rest contrast for all groups. Comparison of group-differences was done via an ANCOVA covarying for mean FD and age at scanning. Furthermore, we specifically tested whether each group showed activation greater than 0 via a one-sample t-test with a one-tailed p-value due to the directional nature of our hypothesis that primary auditory cortex activation should be greater than 0 (Fig S3).

Brain-Behavior Relationship Analysis

To assess brain-behavior relationships we used partial least squares correlation (PLSC) analysis (Krishnan et al., 2011; McIntosh and Lobaugh, 2004). PLSC is widely used in the neuroimaging literature, particularly when explaining multivariate neural responses in terms of multivariate behavioral patterns of variation or a design matrix. Given that the current dataset is massively multivariate both in terms of brain and behavioral datasets, and because there is known correlational structure amongst the measured behavioral variables for language variation (see Fig S2), we used PLSC to elucidate how variation in neural response to speech across large-scale neural systems covaries with behavioral variation across measures of language development. PLSC allows for identifying such relationships by finding latent brain-behavioral variable pairs (LV) that maximally explain covariation in the dataset and for which

are uncorrelated with other latent brain-behavior variable pairs. The strength of such covariation is denoted by the singular value (*d*) for each brain-behavior LV, and hypothesis tests can be made via using permutation tests on the singular values. Furthermore, identifying brain regions that most strongly contribute to each LV pair is made via bootstrapping, whereby a bootstrap ratio is created for each voxel, and represents the reliability of that voxel for contributing strongly to the LV pattern identified. The bootstrap ratio is roughly equivalent to a Z statistic and can be used to threshold data to find voxels that reliably contribute to an LV pair.

The PLSC analyses reported here were implemented within the plsgui Matlab toolbox (www.rotman-baycrest.on.ca/pls/). Here we input first-level all speech versus rest contrast images into the PLSC. For behavioral data, we input EL and RL subscales of the Mullen, as well as the Vineland communication subscale taken from individual's intake and outcome time-points. For statistical inference on identified brain-behavior LV pairs, a permutation test was run with 10,000 permutations. To identify reliably contributing voxels for brain-behavior LVs and to compute 95% confidence intervals (CIs) on brain-behavior correlations, bootstrapping was used with 10,000 resamples. To show voxels that most reliably contribute to significant brain-behavior LVs, we thresholded data for visualization at a bootstrap ratio (BSR) of 1.96 and -1.96. The strength of brain-behavior correlations for significant LVs was displayed as a bar graph with 95% bootstrap CIs as errorbars.

Classifier Analyses

Finally, for classifier analyses we used partial least squares linear discriminant analyses with 5-fold cross validation implemented with the Matlab toolbox libPLS (http://www.libpls.net). Features for the classifiers consisted of behavioral measures taken from the earliest clinical intake time-point or using fMRI speech-related activation (i.e. percent signal chance for all speech vs. rest) from all voxels extracted from the NeuroSynth defined left hemisphere superior temporal cortex ROI. Each feature was z-scored before being input into the classifier. Four classifiers were tested using the following features: 1) ADOS total scores, 2) all Mullen subscale T-scores, Vineland subscale standard scores, and ADOS total scores (i.e. 'clinical' measures), 3) fMRI activation, and 4) a combination of all clinical measures (Mullen, Vineland, and ADOS) plus fMRI activation features. The distinction being made in each classifier was the distinction between ASD Poor versus ASD Good. Receiver operating characteristic (ROC) curves and area under the curve (AUC) values were computed for each classifier in order to determine which of the four classifiers performed best, and accuracy, sensitivity, and specificity were also computed as measures of classifier performance.

Fig. S1. Developmental trajectories for Mullen Fine Motor and Vineland subscales.

This figure shows developmental trajectories for all groups (TD, red; ASD Good, blue; ASD Poor, purple, LD/DD green) on the Mullen fine motor subscale (FM) (A), Vineland communication (B), Vineland socialization (C), Vineland motor skills (D), Vineland daily living skills (E), and Vineland adaptive behavior (F) subscales. Plots show the group-level trajectory (solid line) along with 95% confidence bands, estimated from mixed-effect modeling after taking into account individual-level trajectories (dotted lines, unfilled circles). Related to Figure 1.

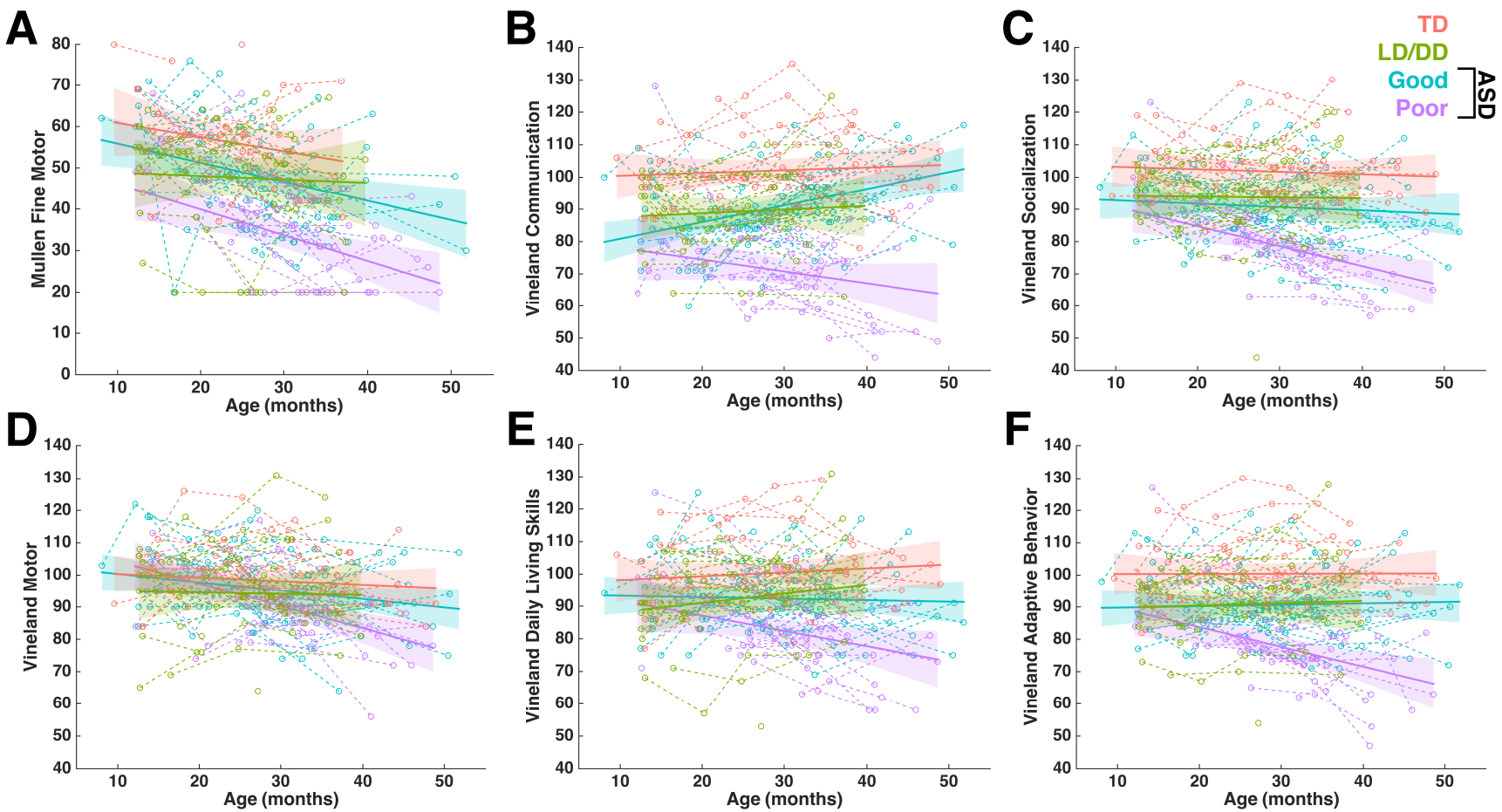

A
Mullen Fine Motor
Age (months)
B
Vineland Communication
Age (months)
C
TD
LD/DD
Good
Poor
ASD
Vineland Socialization
Age (months)
D
Vineland Motor
Age (months)
E
Vineland Daily Living Skills
Age (months)
F
Vineland Adaptive Behavior
Age (months)

Fig. S2. Correlations between all behavioral measures at intake and outcome assessments and correlations between all measure's developmental trajectories.

Panel A shows the correlation matrices for each group across all behavioral measures. Correlations for the intake assessment are shown in the left column of the figure, while correlations for the outcome assessment are shown in the right column. Panel B shows the correlation matrices for each group across all behavioral measures and is computed based on individual-level developmental trajectories estimated within mixed-effect models. Abbreviations: EL, Mullen expressive language; RL, Mullen receptive language; FM, Mullen fine motor; VR, Mullen visual reception; VineComm, Vineland communication; VineSoc, Vineland socialization; VineDly, Vineland daily living skills; VineMtr, Vineland motor skills; VineAdpBeh, Vineland adaptive behavior, ADOS, Autism Diagnostic Observation Schedule; TD, typical development; ASD, autism spectrum disorder; LD/DD, language/developmental delay. Related to Figure 1.

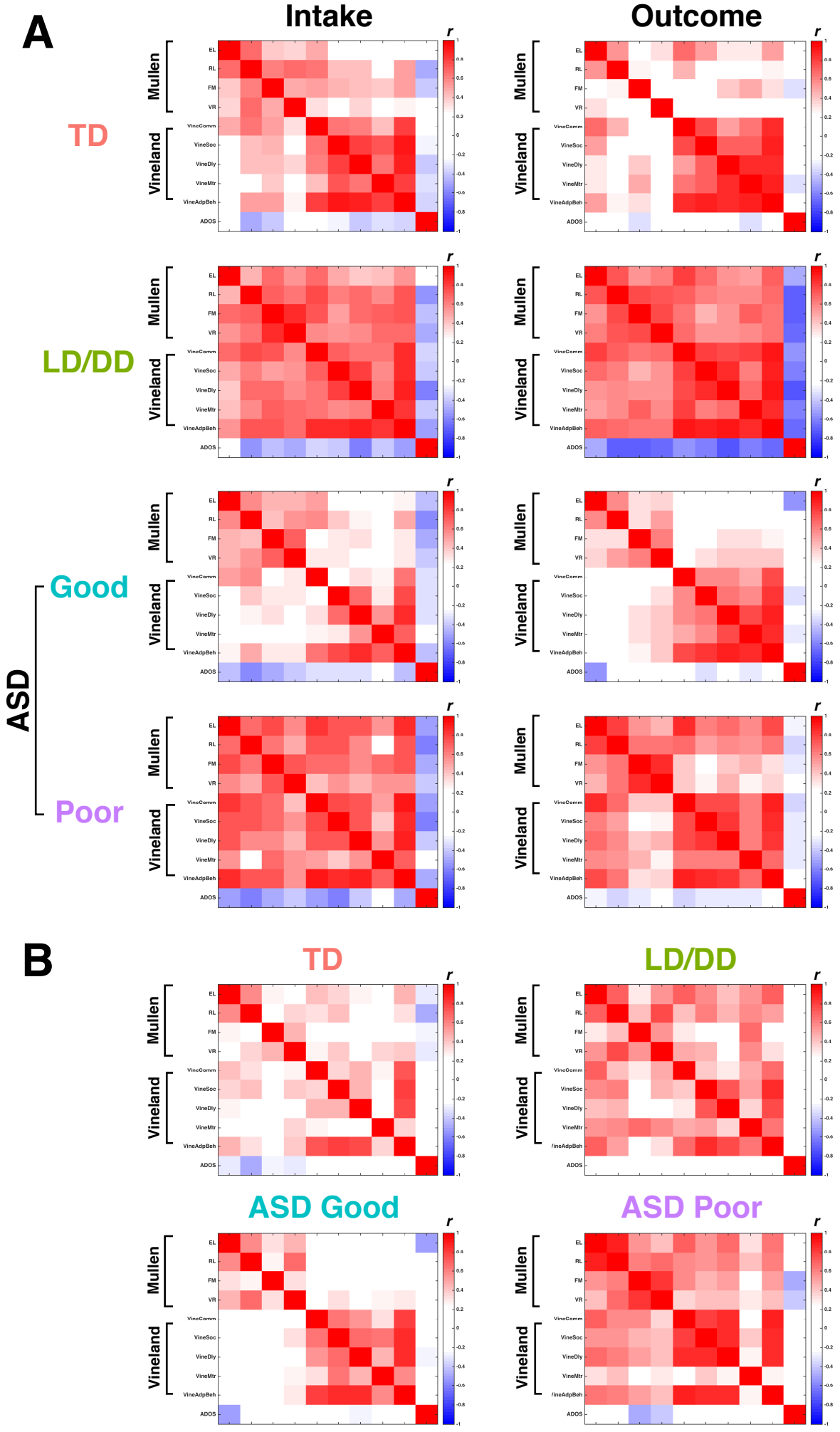

A
Intake
Outcome
TD
LD/DD
ASD
Good
Poor
Mullen
Vineland
EL
RL
FM
VR
VineComm
VineSoc
VineDly
VineMtr
VineAdpBeh
ADOS
r
B
TD
LD/DD
ASD Good
ASD Poor

Fig S3. ROI analysis for left hemisphere primary auditory cortex and montage of single-subject activations in ASD Poor.

This figure shows in panel A, Δ% signal change extracted from the all speech vs rest contrast specifically for the ROI of left hemisphere primary auditory cortex. Black dots indicate the mean and errorbars indicate 95% confidence intervals. This ROI was defined by taking a 10mm sphere from the peak voxel (MNI x = -48, y = -22, z = 6; z-stat = 22.91) of a NeuroSynth meta-analysis combining studies from the features 'auditory cortex', 'auditory cortices', and 'primary auditory'. Here we did not find any significant differences between-groups in an ANCOVA ($F(3,97) = 0.89$, $p = 0.44$, *partial* $\eta^2 = 0.027$). TD, LD/DD, and ASD Good all showed sizeable effect sizes for non-zero activation in this region (e.g., *Cohen's d* > 0.54, *one-tailed p* < 0.006). For ASD Poor, a one-sample t-test shows that there is weak group-level activation at trend-level significance for being different from 0 ($t(23) = 1.67$, *one-tailed p* = 0.054, *Cohen's d* = 0.33) Panel B shows a whole-brain analysis on ASD Poor, where activations are visualized at the very liberal threshold of $p<0.05$ uncorrected, in order to show that subtle group-level activation in left hemisphere primary auditory cortex is present. Panel C is a montage showing all speech vs rest activation in surface renderings of the lateral right and left hemispheres of each individual in the ASD Poor group (activations are thresholded at $p<0.025$ uncorrected). Related to Figure 2.

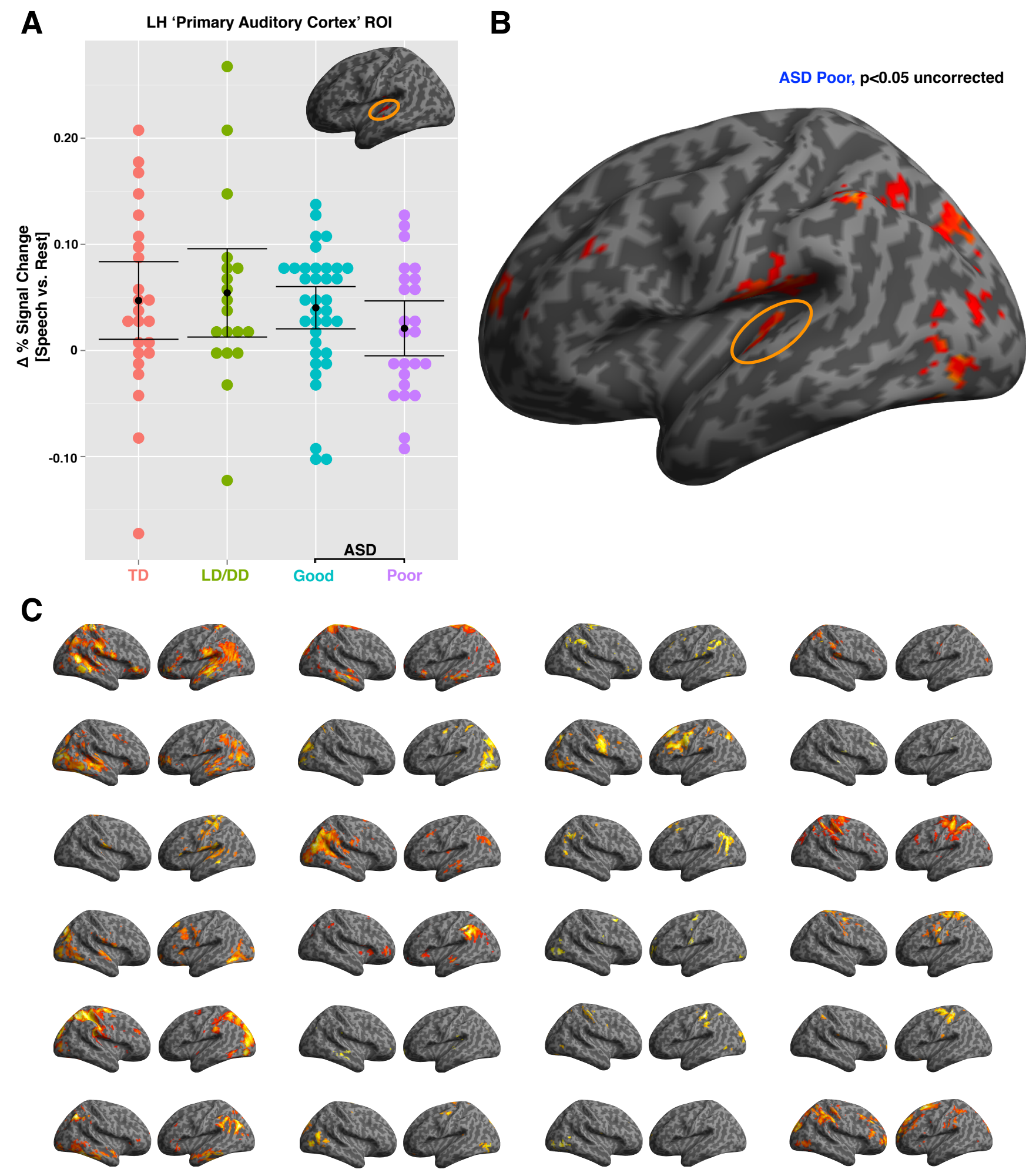

A
LH 'Primary Auditory Cortex' ROI
Δ % Signal Change [Speech vs. Rest]
0.20
0.10
0
-0.10
ASD
TD
LD/DD
Good
Poor
B
ASD Poor, p<0.05 uncorrected
C

Fig. S4. Dot plots showing individual data points for ROI activation and multi-voxel pattern similarity analyses

These plots show individual data points as colored dots (red = TD, green = LD/DD, blue = ASD Good, purple = ASD Poor) for the ROI activation analyses (panels A-D), and for the multi-voxel pattern similarity analyses (panels E-F). Black dots in each plot represent the mean and errorbars represent 95% confidence intervals. Related to Figure 2.

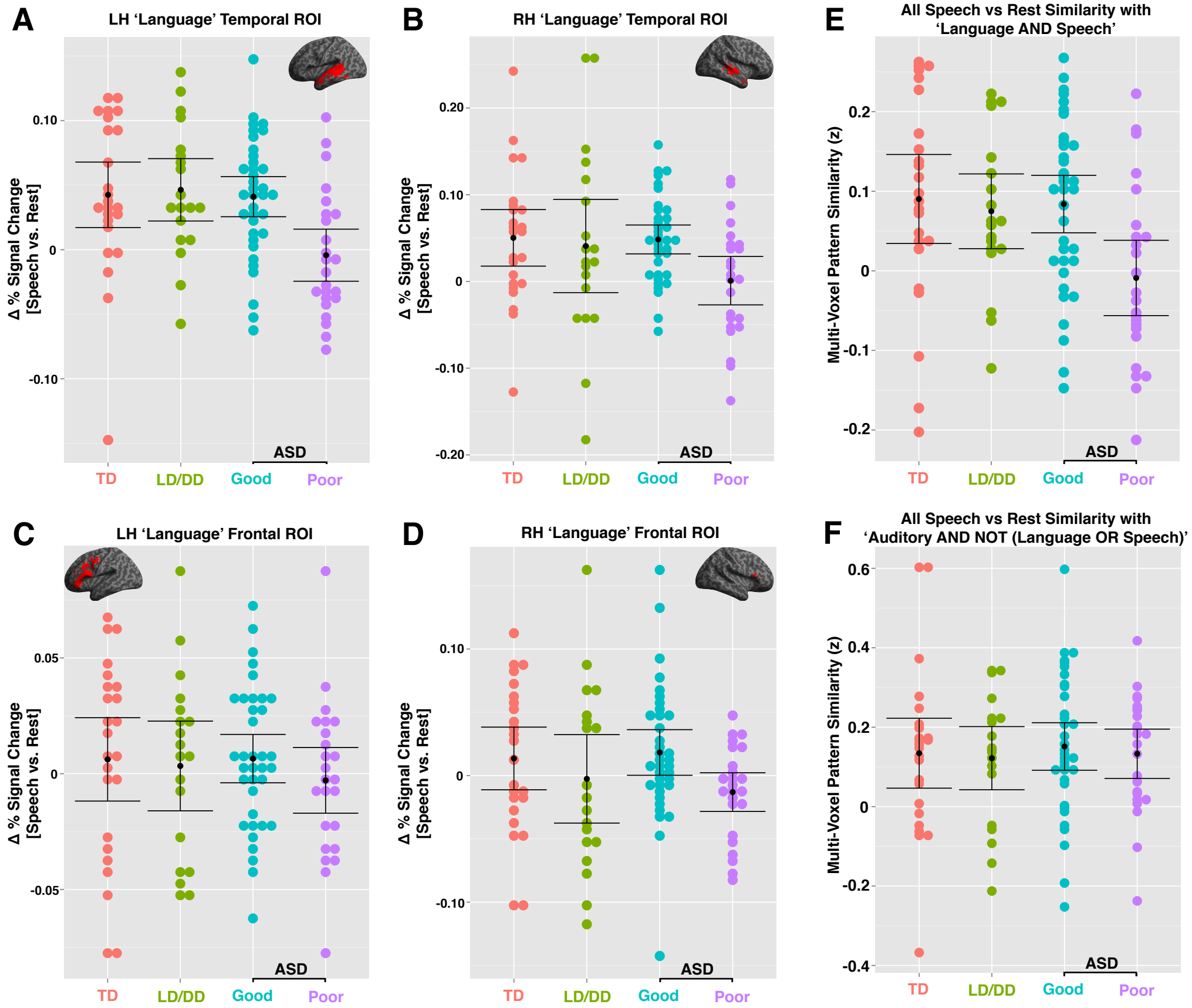

A
LH 'Language' Temporal ROI
Δ % Signal Change [Speech vs. Rest]
0.10
0
-0.10
ASD
TD
LD/DD
Good
Poor
B
RH 'Language' Temporal ROI
Δ % Signal Change [Speech vs. Rest]
0.20
0.10
0
-0.10
-0.20
ASD
TD
LD/DD
Good
Poor
E
All Speech vs Rest Similarity with 'Language AND Speech'
Multi-Voxel Pattern Similarity (z)
0.2
0.1
0
-0.1
-0.2
ASD
TD
LD/DD
Good
Poor
C
LH 'Language' Frontal ROI
Δ % Signal Change [Speech vs. Rest]
0.05
0
-0.05
ASD
TD
LD/DD
Good
Poor
D
RH 'Language' Frontal ROI
Δ % Signal Change [Speech vs. Rest]
0.10
0
-0.10
ASD
TD
LD/DD
Good
Poor
F
All Speech vs Rest Similarity with 'Auditory AND NOT (Language OR Speech)'
Multi-Voxel Pattern Similarity (z)
0.6
0.4
0.2
0
-0.2
-0.4
ASD
TD
LD/DD
Good
Poor

Fig. S5. Quantification of in-scanner head motion.

This figure plots the metric of mean framewise displacement (meanFD) for all individuals in all groups. Panel A shows mean FD for all individuals, while panel B shows mean FD after motion yoking through removal of any individuals with mean FD > 0.1mm (number excluded: n=5 TD; n=7 LD/DD, n=7 ASD Good; n=0 ASD Poor). This range restriction on head movement was sufficient for removing the group-difference in mean FD ($\chi^2(3,80) = 5.07$, $p = 0.16$). We then re-ran the primary ANCOVA on the NeuroSynth left-hemisphere temporal ROI and found that in this motion control analysis, that the group difference remained and that the effect size was larger than in the analysis on all individuals ($F(3,78) = 5.36$, $p = 0.002$, *partial* $\eta^2 = 0.171$). Related to Experimental Procedures.

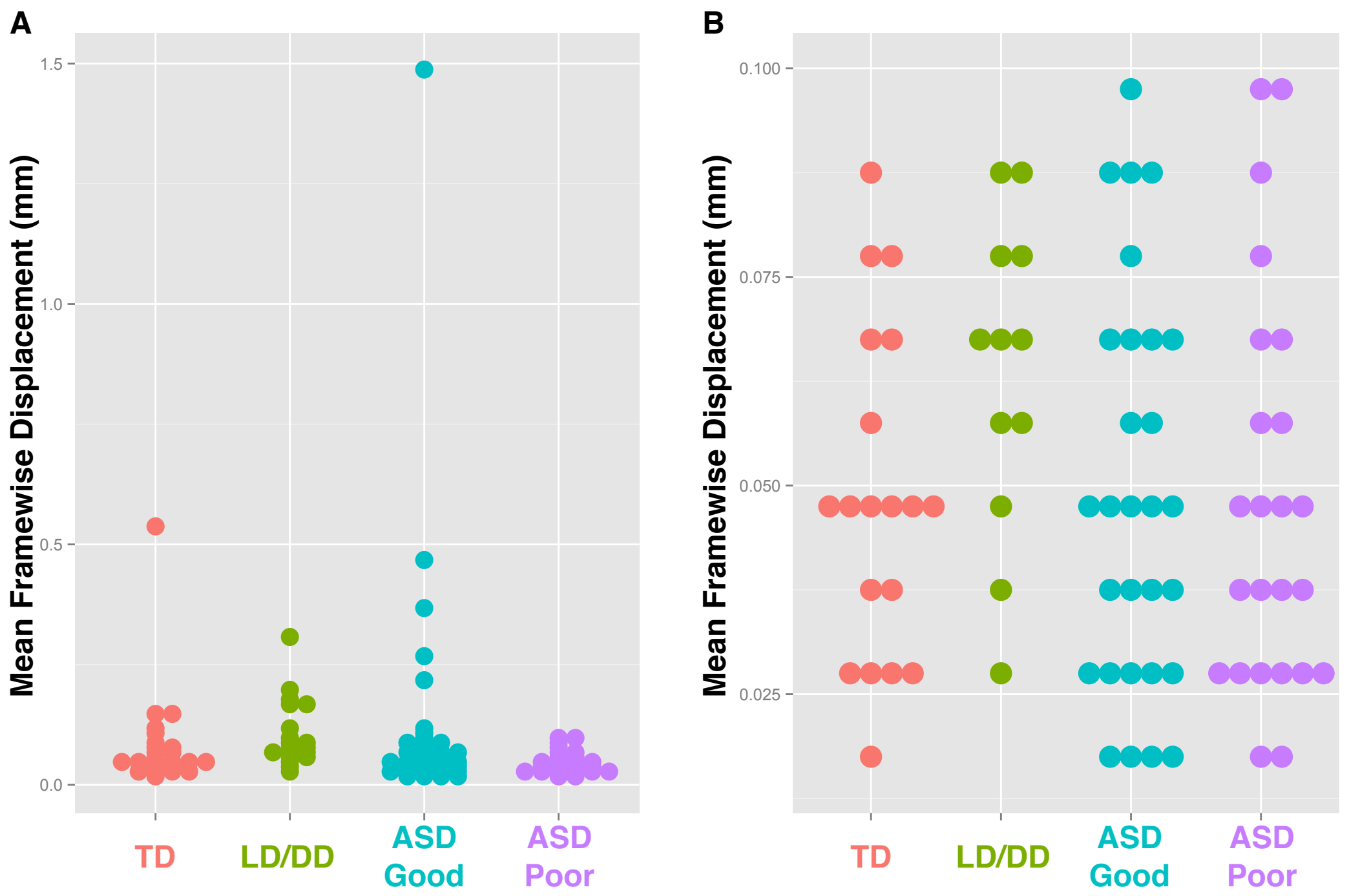
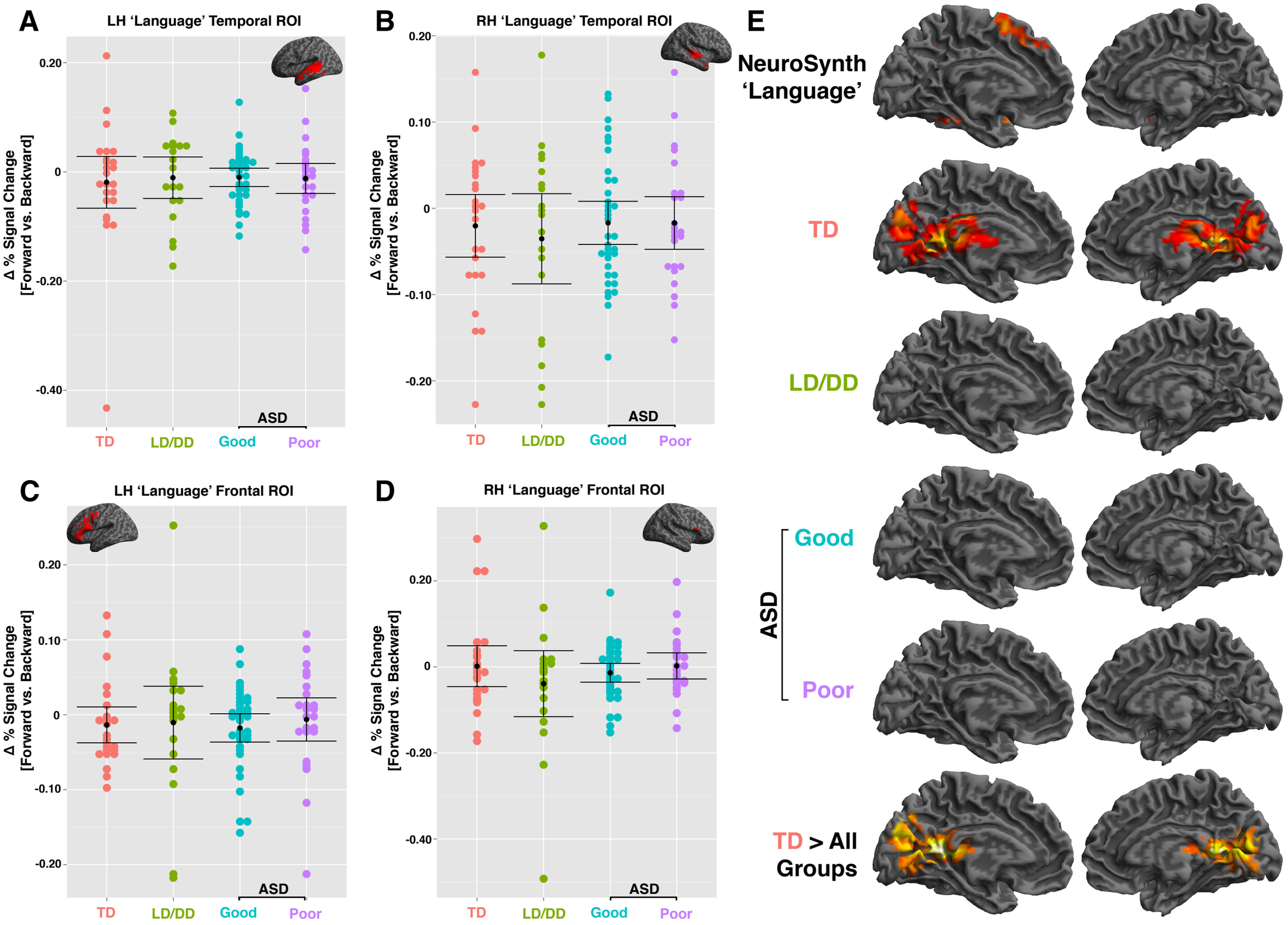

A
Mean Framewise Displacement (mm)
1.5
1.0
0.5
0.0
TD
LD/DD
ASD Good
ASD Poor
B
Mean Framewise Displacement (mm)
0.100
0.075
0.050
0.025
TD
LD/DD
ASD Good
ASD Poor

Fig. S6. Forward vs Backward speech ROI and whole-brain analysis results.

This figure shows neural response to the contrast of forward vs backwards speech across all groups. Panels A-D show the difference in BOLD percent signal change for this contrast from an ROI analysis using the frontal and temporal cortex regions derived from the NeuroSynth meta-analysis map for 'language'. Black dots represent the mean and errorbars represent 95% confidence intervals. The TD (red), LD/DD (green), ASD Good (blue), ASD Poor (purple) groups all show similar levels of Δ% signal change, indicating that these canonical language regions are not differentiated by group status. The top row of Panel E shows the full spatial extent of the NeuroSynth 'language' meta-analysis map along the medial wall of the brain. The subsequent rows within panel E show results from within-group activation analyses corrected at the whole-brain level at FDR $q<0.05$. The bottom row of panel E shows the whole-brain analysis for the specific contrast of TD versus all other groups. Related to Figure 2.


A
LH 'Language' Temporal ROI
Δ % Signal Change [Forward vs. Backward]
0.20
0
-0.20
-0.40
ASD
TD
LD/DD
Good
Poor
B
RH 'Language' Temporal ROI
Δ % Signal Change [Forward vs. Backward]
0.20
0.10
0
-0.10
-0.20
ASD
TD
LD/DD
Good
Poor
C
LH 'Language' Frontal ROI
Δ % Signal Change [Forward vs. Backward]
0.20
0.10
0
-0.10
-0.20
ASD
TD
LD/DD
Good
Poor
D
RH 'Language' Frontal ROI
Δ % Signal Change [Forward vs. Backward]
0.20
0
-0.20
-0.40
ASD
TD
LD/DD
Good
Poor
E
NeuroSynth 'Language'
TD
LD/DD
ASD
Good
Poor
TD > All Groups

Supplementary References

Chumbley, J., Worsley, K., Flandin, G., and Friston, K. (2010). Topological FDR for neuroimaging. Neuroimage 49, 3057-3064.

Eyler, L.T., Pierce, K., and Courchesne, E. (2012). A failure of left temporal cortex to specialize for language is an early emerging and fundamental property of autism. Brain : a journal of neurology 135, 949-960.

Krishnan, A., Williams, L.J., McIntosh, A.R., and Abdi, H. (2011). Partial Least Squares (PLS) methods for neuroimaging: a tutorial and review. Neuroimage 56, 455-475.

Lord, C., Risi, S., Lambrecht, L., Cook, E.H., Jr., Leventhal, B.L., DiLavore, P.C., Pickles, A., and Rutter, M. (2000). The autism diagnostic observation schedule-generic: a standard measure of social and communication deficits associated with the spectrum of autism. J Autism Dev Disord 30, 205-223.

Luyster, R., Gotham, K., Guthrie, W., Coffing, M., Petrak, R., Pierce, K., Bishop, S., Esler, A., Hus, V., Oti, R., *et al.* (2009). The Autism Diagnostic Observation Schedule-toddler module: a new module of a standardized diagnostic measure for autism spectrum disorders. J Autism Dev Disord 39, 1305-1320.

McIntosh, A.R., and Lobaugh, N.J. (2004). Partial least squares analysis of neuroimaging data: applications and advances. Neuroimage 23 Suppl 1, S250-263.

Mullen, E.M. (1995). Mullen scales of early learning. (Circle Pine, MN: American Guidance Service, Inc).

Pierce, K., Carter, C., Weinfeld, M., Desmond, J., Hazin, R., Bjork, R., and Gallagher, N. (2011). Detecting, studying, and treating autism early: the one-year well-baby check-up approach. J Pediatr 159, 458-465 e451-456.

Power, J.D., Barnes, K.A., Snyder, A.Z., Schlaggar, B.L., and Petersen, S.E. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. Neuroimage 59, 2142-2154.

Redcay, E., and Courchesne, E. (2008). Deviant functional magnetic resonance imaging patterns of brain activity to speech in 2-3-year-old children with autism spectrum disorder. Biol Psychiatry 64, 589-598.

Redcay, E., Haist, F., and Courchesne, E. (2008). Functional neuroimaging of speech perception during a pivotal period in language acquisition. Developmental science 11, 237-252.

Sparrow, S., Balla, D., and Cicchetti, D. (1984). Vineland Scales of Adaptive Behavior: Interview edition, survey form manual. (Circle Pines, MN: American Guidance Service.).

Wetherby, A.M., Allen, L., Cleary, J., Kublin, K., and Goldstein, H. (2002). Validity and reliability of the communication and symbolic behavior scales developmental profile with very young children. J Speech Lang Hear Res 45, 1202-1218.

Yan, C.G., Cheung, B., Kelly, C., Colcombe, S., Craddock, R.C., Di Martino, A., Li, Q., Zuo, X.N., Castellanos, F.X., and Milham, M.P. (2013). A comprehensive assessment of regional variation in the impact of head micromovements on functional connectomics. Neuroimage 76, 183-201.

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