Published in final edited form as:
Neuroimage. 2009 April 15; 45(3): 1020–1032. doi:10.1016/j.neuroimage.2008.12.034.

Lack of generalizability of sex differences in the fMRI BOLD activity associated with language processing in adults

S. K. Z. Ihnen[a,*], **Jessica A. Church**[a], **Steven E. Petersen**[a,b,c,e], and **Bradley L. Schlaggar**[a,b,c,d]

[a] *Department of Neurology, Washington University School of Medicine, St. Louis, MO 63110, USA*

[b] *Department of Radiology, WUSM, St. Louis, MO 63110, USA*

[c] *Department of Anatomy and Neurobiology, WUSM, St. Louis, MO 63110, USA*

[d] *Department of Pediatrics, WUSM, St. Louis, MO 63110, USA*

[e] *Department of Psychology, WUSM, St. Louis, MO 63130, USA*

Abstract

A lack of consensus exists as to whether there are sex differences in the fMRI BOLD signal correlates of language processing in the human brain. Here, whole-brain fMRI was used to examine the neural activity of 46 adults performing one of two sets of language tasks. Conservative quantitative and qualitative criteria identified a handful of statistically significant regions of “sex difference” within each task separately. When each of the two sets of regions was investigated in the group of subjects performing the other task set, however, most of the identified “sex differences” failed to generalize. Identical analyses of the same subjects divided into sex-matched pseudorandom control groups for each task set separately revealed that it is possible to observe a similar number of statistically significant regions of “group difference” in the task-associated BOLD signal, even when the groups do not differ on any of the measured behavioral parameters, or any obvious demographic characteristic. Together, these results suggest that one should be cautious when interpreting studies that purport to have identified regions of difference between groups, whether those groups are divided by sex or by any other criterion. In particular, generalization or replication of a result in independent data sets is necessary for establishing conclusive support for any hypothesis about differences in brain function between groups.

Introduction

Neuroscientists and men, women, and children the world over seem to agree that, despite being largely similar, the average human male brain is different from the average human female brain in important and predictable ways. It is the consistent characterization of these differences that has eluded researchers interested in the neurobiology of sex differences, particularly when it comes to differences in language function (e.g. Sommer et al., 2004). A number of behavioral studies have demonstrated *average* performance discrepancies between the sexes for certain categories of cognitive tasks, despite an overwhelming overlap in the male and female

*Corresponding author. WUSM, Department of Neurology-Campus Box 8111, 660 S. Euclid, St. Louis, MO 63110, USA. Fax: +1 314 362 6110. E-mail address: ihnenk@wusm.wustl.edu.

distributions of the scores on these tasks. Specifically, males tend to outperform females at tasks involving spatial reasoning, the perception of vertical and horizontal, mathematical reasoning and spatio-motor targeting, while females tend to outperform males at tasks involving verbal fluency, perceptual speed, verbal and item memory and some fine motor skills (Kimura, 1996).

In addition to performance discrepancies in some tasks, there are reported sex differences in the macrostructure, connectivity and chemistry of the brain that have prompted neuroimaging investigations of how males and females might recruit functional areas somewhat differently (Cahill, 2006). The present report focuses on the identification and the validation of sex differences in the neural activity associated with language processing as measured using fMRI. Because we wanted to emphasize validation of results, we chose to examine sex differences using groups of subjects that some suggest are smaller in number than ideal (e.g. Thirion et al., 2007), since many previous fMRI investigations of sex differences have used groups of comparable size (e.g. Chen et al., 2007; Clements et al., 2006; Weiss et al., 2003; Rossell et al., 2002; Philips et al., 2001). Importantly, the emphasis on *validation* widens the scope of this report, since our conclusions are applicable to small group studies of any kind. Briefly, the major findings motivating the present investigation are reviewed.

A few gross structural dimorphisms between the sexes have been documented; on average, men have larger total brain volumes (by about 10%), larger white matter (WM) fractions and larger cerebrospinal fluid (CSF) fractions; women have larger gray matter (GM) fractions (Lemaitre et al., 2005). Particular brain regions have been compared individually between males and females, and relative size differences (corrected for overall brain volumes) have been reported in a long list of cortical regions, including some that are important for language (Goldstein et al., 2001). The largest effects have been noted in the regions known to have the greatest density of sex steroid receptors during development in animal models, suggesting a hormonal influence on their differentiation (Goldstein et al., 2001). In an MRI study, Schlaepfer et al. discovered a relatively larger GM fraction in the superior temporal gyrus in women compared to men, and Harasty et al. replicated this finding in a postmortem study, adding that another language-associated region, Broca's area, was also proportionally larger in females (Harasty et al., 1997; Schlaepfer et al., 1995). Some investigators have suggested that the corpus callosum is relatively larger in women than in men (Witelson, 1989), potentially facilitating faster interhemispheric transfer, though others point out that this finding may be a byproduct of the fact that smaller brains tend to have larger corpus callosa, and females tend to have smaller brains (Jancke et al., 1997).

McGlone reported a divergence in the outcomes of male and female stroke patients: among those who had experienced left-hemisphere strokes, men were three times more likely than women to become aphasic, and those men that did not become aphasic were still significantly more likely to experience some degree of verbal impairment (McGlone, 1977). Despite the fact that subsequent lesion studies have not consistently replicated McGlone's findings (De Renzi et al., 1980; Kertesz and Benke, 1989; Miceli et al., 1981), the model of greater bilaterality of cortical language representation in females than in males has influenced many neuroimaging investigations of sex differences. Sommer et al. conducted a meta-analysis of fourteen imaging studies of sex differences to determine whether a consensus would emerge from those independent inquiries as to whether or not women really do have a more bilateral representation of language than men (Sommer et al., 2004). Combining across tasks including semantic decision-making, verbal fluency, story listening and rhyme judgment, they concluded that there is insufficient evidence for a task-general, population-level sex difference in the cerebral lateralization of language, pointing out that affirmative reports of greater bilaterality in women than men tended to derive from the studies with the smallest number of subjects. These authors were unable to identify a single task characteristic that seemed to consistently predict a lateralization difference between the sexes, ruling out the suggestions of others that the distinction might be one of phonological vs. semantic or word vs. non-word. In a more recent meta-analysis, Sommer et al reaffirmed their conclusion that sex differences in language lateralization do not exist, this time pooling across data from 26 fMRI tasks, 12 dichotic listening studies and 12 investigations of the asymmetry of the planum temporale (Sommer et al., 2008).

One of the larger single fMRI studies to examine sex differences in language processing compared 50 men and 50 women, and employed both voxelwise and ROI analyses to conclude that the sexes showed very similar, strongly left lateralized activation patterns during an auditory comprehension task (Frost et al., 1999). Numerous other studies attempting to address the question of sex differences in hemispheric lateralization for language have relied heavily on lateralization indices (LIs) to compare males and females. Often, the assumption of this approach is that when a region is recruited, activity in the homotopic region in the contralateral hemisphere provides some meaningful comparison measure (in these cases, homotopic region pairs are created by duplicating one region of some specific interest from one hemisphere into an equally sized mirror image of that region in the opposite hemisphere; in other cases, LIs are instead computed at the level of the entire hemisphere). For an LI, regions are created, a threshold is established, voxels are counted, indices are computed and groups are assessed for an average difference between those LIs. Many studies utilizing some explicit or implicit permutation of an LI analysis have suggested that male and female brains are organized differently for language (Chen et al., 2007; Clements et al., 2006; Rossell et al., 2002; Phillips et al., 2001; Kansaku et al., 2000; Pugh et al., 1996; Shaywitz et al., 1995), while others have concluded that male and female brain activity for language is similar (Weiss et al., 2003; Frost et al., 1999).

The use of lateralization indices is problematic, however, partly because of the inherent assumption that differences in brain organization are best investigated by comparing homotopic regions, and partly because of the unreliability of voxel counting as a metric of brain activation. For an empirical comparison of voxel counts vs. activation magnitudes in terms of inter-trial and inter-subject stability, scatter, statistical power, signal-to-noise sensitivity and associated type II error rates, the reader is encouraged to consult (Cohen and DuBois, 1999). Even if one accepts LIs as useful measures of brain activity, it is complicated to compare functional imaging results couched as voxel counts to results expressed as activation magnitudes (see Jansen et al., 2006 for an interesting exploration of the effects of analysis parameters on LI robustness and reproducibility).

Beyond the question of lateralization differences, the functional neuroimaging literature on sex differences is, in general, decidedly mixed. Perhaps this is partly because “language” tasks take many forms, each engaging its own distinct set of processing demands. Some authors have explicitly postulated that task characteristics determine whether sex differences emerge; Kitazawa and Kansaku have repeatedly argued that tasks involving the processing of global structure (whole sentence or sentences) invoke sex differences, while tasks involving the semantic processing of individual words do not (Kansaku and Kitazawa, 2001; Kansaku et al., 2000; Kitazawa and Kansaku, 2005). Shaywitz and colleagues used fMRI to identify sex differences in the lateralization of activity associated with a phonological task but not with an orthographic or semantic task, and concluded that studies examining sex differences should pay attention to the specific component of language being assessed (Shaywitz et al., 1995). One recent study attributed the inconsistency in the literature to the dependence of sex differences on both task *and* analysis method (Harrington and Farias, 2008). The authors of another study that actually suggested that males are more bilateral than females for language likewise specified that the effect was analysis-dependent; differences were noted at the level of the group, but not on an individual subject basis (Kaiser et al., 2007).

With regard to studies of sex difference in language processing that have utilized tasks similar to the ones we used in the present investigation, a few reports are suggestive of some differences, but together the results do not converge on a single model. Buckner et al. used positron emission tomography (PET) to characterize prefrontal activations across tasks and sexes in subjects performing Verb Generation and Stem Completion (Buckner et al., 1995). While they noted that there were no qualitative differences between males and females (i.e. neither group significantly activated regions that the other group did not also activate), they did report larger activations in men than in women for Verb Generation for ten out of the eleven prefrontal ROIs analyzed. Another PET study examined sex differences associated with five language tasks of varying difficulty, including two reading conditions and three past tense verb generate conditions (Jaeger et al., 1998). These authors found that women activated a greater number of voxels in the right hemisphere than men for verb generate but not for word reading; these discrepancies were seen in both anterior and posterior brain areas. In a less similar investigation of five related PET studies, subjects were asked to name visually presented concrete entities including tools, animals and utensils. Despite similar task performance between the sexes, there were several regions in which men and women showed differential activity, including a region that males activated more than females and a region that females activated more than males (Grabowski et al., 2003).

PET and fMRI studies have likewise yielded a grab bag of results regarding sex differences in cerebral physiology. Using PET, Gur et al. (Gur et al., 1995) identified some regions in which men had relatively higher cerebral metabolism than women, and other regions in which the inverse was true. These authors also noted that hemispheric asymmetries of metabolism could be identified in many regions for both males and females. Levin et al. (Levin et al., 1998) showed that, compared to men, women had substantially lower BOLD signal responses in primary visual cortex (V1), especially in the right hemisphere, and that women additionally activated V1 more symmetrically than men. In contrast to Levin et al., Kastrup et al. (Kastrup et al., 1999) reported that compared to males, females had greater baseline regional cerebral blood flow (rCBF) in visual cortex, larger rCBF increases and increased BOLD signal changes. Marcar et al. (Marcar et al., 2004) noted that the time-dependent attenuation of the BOLD signal amplitude in V1 was larger in women than in men, though the peak value of the signal was not significantly different.

The specific objective of the present study was to investigate BOLD signal sex differences in two similar though distinct sets of language tasks in order to reveal and explore some of the issues that have prevented the literature from converging on a unitary model of how male and female brains might handle language differently. Each of the two sets of tasks was analyzed in the absence of any prediction about how male and female brains might accomplish the task differently, and sex differences that were detected were assessed using the same rigorous criteria that our laboratory uses for the analysis of any group difference. The use of two task sets allowed us to address the critical matter of the generalizability (which is distinct from replicability) of group differences. Because of the prominent place of the lateralization hypothesis (that females are more bilateral and males more left-lateralized for language) in the existing literature, one step of the analysis was additionally dedicated to an evaluation of the data for any evidence for a sex difference in the hemispheric specialization for language.

Materials and General Methods

Subjects

Forty-six healthy adults (23 males, 23 females) aged 18–32 years old each participated in one of two separate fMRI studies of language tasks. Males and females within each study were not significantly different from one another in terms of age, so potentially confounding developmental effects were avoided (see Table 1 for task performance and demographic measures). All subjects were right-handed native English speakers from the Washington University campus and surrounding community. All subjects gave their informed consent. Potential subjects were screened by interview and questionnaire and were excluded from participating for any of the following conditions: metal implants, heart arrhythmias, claustrophobia; history of head trauma, neurological or psychiatric illness; or use of psychotropic medications. The Washington University Human Studies Committee approved both studies, and subjects were compensated for their participation.

Word Generate Task Conditions

The Word Generate task conditions, along with many of the general methods described in this paper, have been detailed at length by Brown et al. (Brown et al., 2005). Subjects (n =13 males and 13 females) in the Word Generate study were scanned while performing three types of lexical association: 1.) verb generation in response to a noun (e.g. stimulus: ‘BALL’, a correct response: ‘bounce’); 2.) opposite generation (e.g. stimulus: ‘LESS’, a correct response: ‘more’); and 3.) rhyme generation (e.g. stimulus: ‘BEACH’, a correct response ‘peach’). Stimuli from all three tasks were presented either visually (via a back projection screen) or aurally (via headphones), with each run comprising only one task condition and one stimulus modality, for a total of six task runs. In response to each stimulus, subjects were asked to generate aloud a single word, per the condition of that run. Subjects were encouraged to respond quickly and accurately and to minimize head movement. Verbal responses were recorded during scans so that accuracy and reaction times could be determined post hoc. For each stimulus, there were usually multiple possible correct responses (for example, correct responses for BALL would include “throw,” “bounce” and “catch”), and accuracy was evaluated consistently between subjects. Stimuli were white on a black background, with each letter subtending about a half of a degree of visual angle. At the beginning of each run, a fixation crosshair appeared at the center of the screen, where it remained for the duration of the run, except when it was replaced by the visual stimulus. During auditory runs, the crosshair remained on the screen throughout. Visual stimulus duration was 1.37 s, while auditory stimulus duration varied by word length. Each run lasted 3 min 39 s and consisted of 21 stimulus trials, followed by ~90 s of rest between runs. Stimuli were presented every second, third or fourth MR frame (T_R = 3.08 s; average interstimulus interval = 9.24 s) in pseudo-random order. Jittering the stimulus presentation in this way allowed the extraction of the timecourses associated with the trial-by-trial events of interest.

Word/Nonword Read Task Conditions

Subjects (n=10 males and 10 females) in the Word/Nonword Read study (Church et al., 2006) were scanned while reading words of one of five types: 1.) low-frequency one-syllable words (e.g. ‘oaf’, ‘gape’, ‘dirge’); 2.) one-syllable nonwords (e.g. ‘nax’, ‘blep’, ‘stril’); 3.) low-frequency three-syllable words (e.g. ‘conifer’, ‘pastrami’, ‘ombudsman’); 4.) three-syllable nonwords (e.g. ‘hapical’, ‘bertiset’, ‘adrotteng’); and 5.) one-syllable high-frequency words (e.g. ‘got’, ‘thank’, ‘strange’). Each run featured 21 stimuli of one condition only, and each subject performed three runs of each condition, for a total of fifteen task runs. As in the Word Generate study, verbal responses were recorded during scans so that accuracy and reaction times could be determined post hoc. For each stimulus, multiple attempts, omissions and incorrect pronunciations were scored as incorrect; responses that followed standard English pronunciation rules were accepted as correct. A single rater assessed the data to ensure that accuracy was fairly calculated between subjects. The remaining parameters for the Word/Nonword Read study (physical description of visual stimuli, stimulus presentation duration, jittering) were identical to those described above for the Word Generate study.

Performance matching

Previously, we have demonstrated that interpretation of between-group effects in neuroimaging studies requires careful attention to between-group task performance discrepancies, since such discrepancies alone can produce differences in neural activity (Brown et al., 2005; Schlaggar et al., 2002). One way of addressing this issue is to ensure that the groups being compared are matched, on average, along a behavioral parameter that is measured during task performance, such as reaction time (RT) or accuracy. In these analyses, the males and females from each task set were not significantly different from one another in terms of RT. In the Word Generate study, percent correct scores were recorded for each subject, and although only correct trials were included in the analyses, males and females were not significantly different from one another on this measure. Because accuracy for both males and females averaged over 99% for the Word/Nonword Read study, all of the responses from all of the tasks in this study were included in the analyses (see Table 1 for RT and accuracy data for both task sets).

Imaging Data Acquisition and Preprocessing

Structural and functional neuroimaging data were collected using a Siemens 1.5 Tesla MAGNETOM Vision System (Erlangen, Germany). A sagittal magnetization-prepared rapid gradient echo (MP-RAGE) sequence having the following parameters was utilized to acquire rapid three-dimensional high-resolution structural images: slice T_E= 4 ms, T_R= 9.7 ms, T_I= 300 ms, flip angle = 12°, 128 slices, 1.25 × 1 × 1 mm voxels. An asymmetric spin-echo echo-planar pulse sequence sensitive to blood oxygenation level-dependent (BOLD) contrast having the following parameters was used to acquire functional data parallel to the anterior commissure – posterior commissure plane: T_R= 2.18 s with a 904 ms delay (total TR = 3.08 s), T_2* evolution time = 50 ms, flip angle = 90°. Each scan was comprised of 73 frames of 16 contiguous interleaved 8 mm axial slices (3.75 × 3.75 mm in-plane resolution), permitting coverage of the entire brain. Steady state was assumed after three frames (~9 s), so acquisition of functional data began with the fourth frame of each run.

Automated preprocessing procedures included the following: removal of a single pixel spike caused by signal offset; whole-brain normalization of signal intensity across MR frames; correction for subject movements both within and between runs; and slice-by-slice normalization to correct for changes in signal intensity introduced by the acquisition of interleaved slices.

Functional BOLD data were registered to structural MP-RAGE data on a subject-by-subject basis. All data were then transformed into a common stereotactic space (based on Talairach and Tournoux, 1988), facilitating direct comparisons across groups.

To help discourage movement, subjects were positioned in the scanner using a thermoplastic mask individually fitted to the face and attached to the head coil. Subject motion was corrected and quantified so that adjustments could be made to realign head movement on a frame-by-frame, post hoc basis. Such adjustments were based on root mean square (RMS) variance values for translation and rotation in each of the *x, y* and *z* planes (in millimeters), then total RMS values were calculated on a run-by-run basis for each subject. In both studies, males and females were similar in terms of average movement across the entire scanning session (see Table 1).

Imaging Data Analyses

Statistical analysis of event-related BOLD fMRI data was based on a general linear model (GLM) and computed using in-house software programmed in the Interactive Data Language (IDL) (Research Systems, Inc., Boulder, CO). The goal in the use of these two task sets was to identify brain regions exhibiting statistically significant sex x time interactions in language processing tasks. Characterization of potentially orthogonal effects of stimulus condition or stimulus modality was not an objective. Therefore we collapsed across all six task conditions in the Word Generate study (two modalities, three stimulus types) and likewise across all five task conditions in the Word/Nonword Read study.

The GLM design for both studies incorporated time as a seven-level factor, with the seven levels corresponding to successive MR frames following presentation of the stimulus. No assumptions were made about the shape of the hemodynamic response function (HRF), which allowed for the detection of a BOLD response with any shape over the period of ~22 seconds during which it was modeled (3.08 seconds per frame). Timecourses for all analyses were entered into ANOVAs using random effects models. For whole-brain analyses, a correction based on a Monte Carlo simulation was undertaken to guard against false positives that potentially emerge when a large number of statistical comparisons are made across images.

All the data shown have been screened for highly aberrant values; outlier timecourses were considered to be those in which the percent signal change for a single subject exceeded a magnitude of two at any one of its seven timepoints (based on Brown et al). Regions having outlying timecourses in 10% or more of subjects were completely excluded, while regions having outlying timecourses in fewer than 10% of subjects were treated by removing the aberrant subjects' timecourses and re-computing the statistics.

Methods: Data Analysis

For each whole-brain (voxel-wise) analysis, a stringent two-step approach was utilized to establish regions of significant difference between groups (both for males vs. females and for pseudorandom group A vs. group B, see below). First, a region was defined as significant in the group (2 levels, e.g. male and female) x time (7 levels; MRI acquisition volumes) ANOVA as one in which at least 24 contiguous voxels demonstrated an interaction at $Z > 3.5$ ($p < 0.05$, corrected). The second, post hoc check required that each significant region meet the following criteria: 1) retains a significant ($p < 0.05$) group x time effect when the analysis is restricted to the heart of the timecourses, timepoints two through four (TP2-TP4); 2) has no outliers (subjects with greater than two percent signal change for any timepoint, see above) or has been purged of outliers and still continues to meet all criteria; 3) exhibits biologically plausible timecourses (timecourses that resemble either activating, deactivating or flat hemodynamic response functions) in both groups; 4) is located in cerebral gray matter, as best as can be determined by consulting both the Talairach atlas and the IDL atlas created by averaging the anatomies of the subjects from each study separately; and 5) has a percent signal change difference (between males and females, e.g.) at TP3 that is at least of magnitude 0.1. (We have previously used 0.1% as a minimum biologically meaningful group difference, believing that it allows us to exclude from consideration timecourses in which there is a *statistical* difference between groups that is too small to be important for information processing (Brown et al., 2005).) Regions resulting from the whole-brain ANOVAs that also met the above five post hoc criteria were subsequently entered into regionwise (ROI) ANOVAs incorporating factor levels and statistical parameters identical to those used in the whole-brain analyses.

For each ROI analysis, a group (2 levels) x time (7 levels) ANOVA was computed in particular sets of pre-defined voxels (ROIs). A group x time effect was significant in these ROIs if $p < 0.05$ and the group difference in percent signal change at TP3 was ≥ 0.1. Box's sphericity correction was used, adjusting for temporal autocorrelation and possible inhomogeneity of variance over the repeated measure, time (Box, 1954; McAvoy et al., 2001).

The interpretation of ROI ANOVAs depends on how the regions are applied. When regions are reapplied to the data from which they are derived, the objective is not to recompute the statistics, which would be biased, but to extract the timecourses of activity from those regions. As explained above, each whole-brain analysis in the present study was therefore followed by a regionwise ANOVA for the purpose of extracting timecourses. Regions can alternatively be applied to completely independent sets of data, in which case the resulting statistics are unbiased, and the power to detect effects is greater than it is for the original whole-brain analysis. These statistics can therefore be correctly interpreted as indicating the significance of the effect of interest in those regions within that independent data set, an analysis technique that was also utilized in the present study.

Are there statistically significant BOLD signal sex differences in adults performing a language task?

Analysis step 1: Identifying sex differences in adults performing a set of Word Generate tasks: First a whole-brain sex x time ANOVA was computed of the BOLD activation data from the 26 adult subjects performing the Word Generate tasks. Here the result of interest was the set of regions in which there were statistically significant sex x time interactions that additionally met all five of the post hoc criteria described above.

Timecourses for individual subjects were also computed and plotted for each of the regions identified in this analysis.

Do the sex differences generalize across sets of language tasks?

Analysis Step 2.1: Applying sex x time regions from Word Generate to a Word/Nonword Read task set: The next objective was to determine whether the regions of sex x time interaction identified in the Word Generate task set would *generalize* when applied to a separate but comparable language task, Word/Nonword Read. Though Word Generate and Word/Nonword Read certainly share some features and computational processes, including the nature of the visual stimuli and the need for an articulated response, the two sets of tasks also make distinct cognitive demands. Furthermore, the two data sets are independent from one another, since they were acquired as two distinct investigations using completely non-overlapping subjects. Assessing Word/Nonword Read for effects identified in Word Generate is therefore an attempt at generalization, not replication. An ROI sex x time ANOVA was computed of the BOLD activation data from the adult subjects engaged in a Word/Nonword Read task set, using the regions identified in the previous whole-brain analysis of the Word Generate task set. As described above, this ROI analysis asks whether the regions identified as showing sex x time interactions in one study retain these interaction effects in a second study in terms of both statistical significance ($p < 0.05$) and magnitude (a group difference at TP3 of at least magnitude 0.1% signal change).

Analysis Step 2.2: Identifying sex x time interaction regions in adults performing Word/Nonword Read: The ROI analysis just described establishes whether or not the Word/Nonword Read set of tasks involves the *same* set of regions of “sex difference” as the Word Generate set of tasks -- in other words, whether or not one set of observed “sex differences” generalizes to an independent data set. Regardless of the outcome of that analysis, there may be a distinct set of regions of statistical difference between males and females in the Word/Nonword Read task set when it is analyzed independently. To separately test for this possibility, a second whole-brain sex x time ANOVA was computed of the BOLD activation data from only the subjects performing Word/Nonword Read. Here the result of interest was the set of regions in which there were sex x time interactions that additionally met all five of the post hoc criteria described above.

Timecourses for individual subjects were also computed and plotted for each of the regions identified in this analysis.

Analysis Step 2.3: Applying sex x time regions from Word/Nonword Read to Word Generate: In this analysis, the objective was again to determine the generalizability of the results, this time of the regions of sex x time interaction identified in the Word/Nonword Read set of tasks. An ROI sex x time ANOVA was computed of the BOLD activation data from the subjects engaged in the Word Generate set of tasks, using the regions identified in the previous whole-brain analysis of the Word/Nonword Read set of tasks. This analysis is essentially the inverse of analysis step 2.1, and it asks whether the regions identified as showing sex x time interactions in one study retain these interaction effects in a second, independent data set in terms of both statistical significance ($p < 0.05$) and magnitude (a group difference at TP3 of at least magnitude 0.1% signal change).

Do the results support any single lateralization hypothesis?

Analysis Step 3: Evaluating the laterality of the observed sex differences: To explicitly evaluate whether the results from both sets of tasks support any single lateralization hypothesis, each region of sex difference was characterized post hoc according to a 2 × 2 classification scheme. A region was classified according to the direction of its effect (M>F or F>M for the peak of the activation) and the cerebral hemisphere in which the region was located (right or left), resulting in four categories of observations. Each category of observation supports one or two specific hypotheses about sex-specific brain lateralization, as described in the results section.

For each set of tasks separately, regions were classified into each of the four categories, and the number of regions falling into each category was counted.

Are the differences observed decidedly attributable to sex?

Analysis Step 4: Reliability check: In order to determine the likelihood that “group differences” unrelated to any known between-subject variable can be obtained in an analysis of these two particular BOLD activation data sets, the subjects were divided, from each task separately, into two pseudorandom “control” groups. The groups were composed of equal numbers of males and females, and were well-matched for head size, age and performance, but were otherwise chosen randomly (see Table 2 for the demographic and task performance data for the control groups for Word Generate, Table 3 for Word/Nonword Read). Analysis streams exactly identical to those used to query for “sex differences” were applied to look for “group differences.” Thus a whole-brain group x time ANOVA was computed for each set of tasks separately, and in each case the result of interest was the set of regions in which there were statistically significant group x time interactions that additionally met all five of the post hoc criteria described above. As in analysis step 1 and step 2.2, timecourses for individual subjects were also computed and plotted for each of the regions identified in these two analyses.

Results

Are there statistically significant BOLD signal sex differences in adults performing a language task?

Analysis step 1: Identifying sex differences in adults performing a set of Word Generate tasks—The whole-brain sex x time ANOVA that was computed on the BOLD activation data from the 26 adult subjects performing the Word Generate tasks identified 17 regions of “sex difference.” These regions each showed a statistically significant sex x time interaction that additionally met all five of the post hoc criteria described in the methods. In each of the 17 regions, the average timecourse of activation for the males was of larger magnitude than that of the females. Figure 1 depicts a whole-brain surface rendering of all 17 regions, along with group timecourses from 8 representative regions. Table 4 lists all of the regions' Talairach coordinates, volumes, approximate Brodmann areas (BA), and TP3 effect sizes (percent signal change difference between males and females at time point three).

Supplemental figure 1 shows the individual timecourses in addition to the group timecourses for three of the regions depicted in Figure 1.

Do the sex differences generalize across sets of language tasks?

Analysis Step 2.1: Applying sex x time regions from Word Generate to a Word/Nonword Read task set—Only one out of the 17 regions that showed a statistically significant sex x time interaction in Word Generate also showed a significant sex x time interaction in Word/Nonword Read. As shown in Figure 2, the direction of the "sex difference" switched between the two task sets, so that for Word Generate, males activated more than females, and for Word/Nonword Read, females activated more than males. Illustrated also in Figure 2 are the timecourses from three example regions in which the sex x time interaction identified in Word Generate did *not* generalize to the Word/Nonword Read.

Analysis Step 2.2: Identifying sex x time interaction regions in adults performing Word/Nonword Read—The second whole-brain sex x time ANOVA, computed of the BOLD activation data from only the subjects performing Word/Nonword Read, identified 13 regions of "sex difference." These regions each showed a statistically significant sex x time interaction that additionally met all five of the post hoc criteria described in the methods. In 8 of the regions, the average timecourse of activation for the females was larger in magnitude than that of the males, while in 5 regions, the opposite was true. Figure 3 depicts a whole-brain surface rendering of all 13 regions, along with group timecourses from 6 representative regions. Table 5 lists all of the regions' Talairach coordinates, volumes, approximate Brodmann areas, and TP3 effect sizes (percent signal change difference between males and females at time point three).

Analysis Step 2.3: Applying sex x time regions from Word/Nonword Read to Word Generate—Four of the 13 regions that showed a statistically significant sex x time interaction in Word/Nonword Read also showed a significant sex x time interaction in Word Generate. As shown in Figure 4, the "sex difference" was in the same direction in two of the regions and in different directions for the other two regions. Illustrated also in Figure 4 are the timecourses from three example regions in which the sex x time interaction identified in Word/Nonword Read did not generalize to the Word Generate task set.

Do the results support any single lateralization hypothesis?

Analysis Step 3: Evaluating the laterality of the observed sex differences—The 2 × 2 post hoc classification scheme assigned regions to one of four categories based on the observed sex x time interaction: M>F (for the average peak of activation) in the left hemisphere; F>M in the right hemisphere; M>F in the right hemisphere and F>M in the left hemisphere. Table 6 indicates the number of regions from each task set that fall into each category, along with the hypotheses regarding the lateralization of language processing that are suggested by each one.

Are the differences observed decidedly attributable to sex?

Analysis Step 4: Reliability check

Word Generate Set of Tasks: In the Word Generate set of tasks, the pseudorandom "control group" x time whole-brain ANOVA identified 1 region that met both the statistical criteria and the five post hoc criteria. This region therefore exhibited a "group difference" of an unknown origin within the context of a single set of language tasks, Word Generate. The region was located in the lingual gyrus (TC 19, −80, 05), and Supplemental figure 2 depicts both the average timecourses and the individual timecourses for the region. The p value for the “group difference” in this region was < 0.00005.

Word/Nonword Read Set of Tasks: In the Word/Nonword Read set of tasks, the pseudorandom “control group” x time whole-brain ANOVA identified 14 regions that met both the statistical criteria and the five post hoc criteria. These regions therefore exhibited “group differences” of an unknown origin within the context of a single set of language tasks, Word/Nonword Read. For each of the fourteen regions, the effect was in the same direction: the average timecourse of activation of group two was of larger magnitude than that of group one. Table 7 lists the 14 regions’ Talairach coordinates, volumes, approximate Brodmann areas, and TP3 effect sizes (percent signal change difference between group one and group two at time point three). Figure 5 depicts a whole-brain surface rendering of these 14 regions, along with example timecourses from 8 of them.

Supplemental figure 3 shows the individual timecourses in addition to the group timecourses for three of the regions depicted in Figure 5.

Discussion

The purpose of this investigation was to determine whether there are *statistically significant* sex x time interactions in the neural correlates of two separate types of language tasks as examined with fMRI, and to evaluate objectively whether or not those interactions represent *functionally meaningful* differences. Whole-brain analyses that included rigorous, post hoc qualitative and quantitative criteria were used to identify “sex differences” in the BOLD signal associated with two sets of tasks, Word Generate and Word/Nonword Read. To determine how generalizable the differences are, each set of observed “sex differences” was reciprocally queried in a regionwise analysis of the other set of tasks, and the differences were found to be largely non-overlapping. Furthermore, each set of results was evaluated for the extent to which it might suggest a sex discrepancy in the left/right lateralization of neural activity during these tasks, and no support was found for any single model of hemispheric specialization. Finally, subjects from each study separately were shuffled into sex-matched, pseudorandom “control groups” to determine whether “group differences” comparable in number and magnitude to the suggested “sex differences” might emerge in studies of this size and power. Such inexplicable “group differences” were indeed discovered, throwing into question the reliability of the observed “sex differences.”

It is possible to identify regions of presumed “sex difference” in a single data set

The first question asked was whether there are sex differences in adults in the neural activity associated with a set of Word Generate tasks, involving task conditions that have been well-characterized by our group (Brown et al., 2005; Buckner et al., 1995; Petersen et al., 1988). Seventeen regions of sex difference, spread throughout both cerebral hemispheres, were identified, and these regions emerged despite similar age and similar task performance between males and females. In all 17 regions, the average timecourses of activations of the males were of larger magnitude than those of the females. Despite the potentially suggestive uniformity of the direction of the sex effect (M>F), it is highly unlikely that the results could be attributable to the population average head size difference between males and females, given that the actual difference between these two samples is so small (see Table 1), and given that the effects generally held up in a re-analysis of smaller, head-size matched groups (data not shown). None of the regions overlaps with classic Broca’s or Wernicke’s areas, and in general the spatial distribution of the regions defies a tidy mechanistic description of how men and women could possibly be performing the tasks differently.

Functionally meaningful regions of “sex difference” should generalize across task sets

Results supporting or refuting the existence of a sex difference in language-associated brain activity should hold up to independent *replication* of the same task or to independent *generalization* to a related but distinct task. Thus our finding of regions of *statistically significant* “sex difference” in the neural correlates of one language task in one set of adults prompted us to ask whether those regional differences could also be identified in a second set of adults performing a related but distinct language task (i.e. to ask whether those results generalize). Only one of the 17 ROIs from Word Generate did retain a sex x time effect in Word/Nonword Read that met our criteria, and this region was located in the right middle temporal gyrus (BA 37; TC 54, -53, 06). The p value for the sex x time interaction in this region in Word/Nonword Read was 0.02, close to our threshold of 0.05. The direction of the effect was reversed between the two task types, such that in Word Generate males were more active than females, while in Word/Nonword Read, females were more active than males. Because of this switch in effect direction between the task types, an interpretation of the reliability and of the importance of this region as a region of “sex difference” in language processing cannot be established in the present study. The importance of the 16 other “sex difference” regions that failed to generalize to the second task set remains indeterminable.

The applied ROI analysis just described showed that the Word/Nonword Read task set largely did not involve the *same* set of regions of sex difference as the Word Generate set of tasks. However, we were able to identify13 regions of “sex difference” in the Word/Nonword Read task set when this task set was analyzed separately. These regions, like the Word Generate regions, were distributed throughout the left and right hemispheres (see Figure 3 and Table 5). Again, these differences in neural activity emerged despite similar age and similar task performance between the males and females. Also, none of the regions overlaps with classic Broca’s or Wernicke’s areas, and the spatial distribution of the regions defies a succinct functional description. Four of the 13 regions that showed a statistically significant sex x time interaction in Word/Nonword Read task set also showed a significant sex x time interaction in Word Generate task set, two in the same direction in both tasks and two in opposite directions. As concluded with regard to Word Generate, the failure of the majority of the regions of “sex difference” identified in Word/Nonword Read to generalize to a similar, independent data set in a consistent way renders a definitive functional description of these regions unobtainable with the present data.

The “sex differences” we identified do not support a single lateralization hypothesis

Given the curiosity in the literature on sex differences regarding hemispheric specialization, we did feel that it was necessary to objectively address the issue of lateralization in our own data. Analysis step 3 therefore involved a straightforward quantification of the right/left distribution of “sex difference” regions in each set of tasks separately. Regions were classified, post hoc, into categories defined according to the direction of the effect (M>F vs. F>M) and the hemisphere in which the region was located (right vs. left). Particularly because all regions reported fell within a narrow distribution of volumes (from 0.22 to 1.01 cm^3), we felt that this was an appropriate method of asking to what extent the observed sex differences suggested a particular sex discrepancy in left vs. right hemisphere recruitment of functional areas. This scheme required no assumptions about hemispheric specialization, and any model regarding lateralization could potentially be supported by the results. This approach furthermore avoided the limitations both of voxel counting and of the homotopic region assumption (discussed in the Introduction).

In Word Generate, 75% of the regions showed an effect of M>F in the right hemisphere (one region was located along the midline and thus not included in the counts). As indicated in Table 6, the lateralization scenarios supported by this majority finding are that males are more right-lateralized for language and females more bilateral, or females are more left-lateralized and males more bilateral. Both of these interpretations are contrary to the popular theory that males are more left-lateralized and females more bilateral for language. The remaining 25% of the regions in Word Generate showed an effect of M>F in the left hemisphere, suggesting that males are more left-lateralized for language and females more bilateral, or females are more right-lateralized and males more bilateral. This minority finding does therefore potentially (though not necessarily) support the widespread notion that males are more left-lateralized and females more bilateral for language. Taken together, the 17 regions of sex difference in Word Generate fail to support a single straightforward model of a hemispheric sex difference in language processing, in fact substantiating, by a majority count, the inverse of the usual supposition.

In the Word/Nonword Read task set, the regions were more evenly distributed between the categories of classification, as shown in Table 6. Since there are comparable numbers of regions showing each effect, a single conclusion regarding sex differences in lateralization and language processing is again not warranted by the data. There are two reasonable explanations for the fact that this conclusion is inconsistent with previous reports of greater left hemispheric language lateralization in males than females. The first is methodological, and pertains in part to the problem of lateralization indices alluded to in the introduction. The second is biological, since it is plausible that some other types of language tasks engage distinct language processes that *do* result in sex differences in the lateralization of brain activity. Our conclusion here is therefore fairly specific: as measured with fMRI, males and females are not dissimilarly lateralized for the neural activity associated with either single word generation or the reading of individual words and nonwords. As Sommer et al. (Sommer et al., 2004) concluded, there may truly be a small lateralization difference at the population level that is difficult to detect; lateralization differences may be task-dependent (and not important for either of our task sets); or, simply, there may not be a substantial sex difference in the cerebral lateralization for language processing.

It is possible to identify regions of inexplicable “group difference” in a single data set using the same analyses that identified “sex differences”

After having identified “sex differences” in the BOLD activity associated with two separate sets of language tasks and characterizing the laterality of those effects, we wanted to check the reliability of our findings by determining the likelihood that “group differences” unrelated to any known between-subject variable could be obtained in an analysis of these two particular BOLD activation data sets derived from relatively small numbers of subjects. For each set of tasks separately, subjects were divided into two pseudorandom “control” groups, composed of equal numbers of males and females and well matched for every factor for which data had been collected (education level, RT, accuracy, movement, head size and age). Any differences identified in the analyses of these control groups must derive either from some unknown parameter along which the groups have unintentionally been segregated or from sufficient between-subject variability in the data so as to promote the detection of artifactual statistical distinctions (false positives) or from some combination of these factors. Regardless of their source, the extent to which unascribable group differences can be discovered in the data affects the degree of confidence with which one can interpret the result of interest.

The whole-brain analysis of Word Generate identified only one region that met all of the statistical and post hoc criteria for “group differences.” This single region compares with 17 regions identified in the sex effect analysis of Word Generate, and in the context of a statistical hypothesis test with a p value set to 0.05, is not a particularly remarkable finding. In the Word/Nonword Reading set of tasks, the whole-brain ANOVA identified 14 regions that met the statistical and post hoc criteria for “group differences,” compared to 13 regions in the sex effect analysis. This result *is* surprising, especially since we believed that our conservative criteria for defining regions would adequately exclude from consideration any effects not attributable to biologically driven differences in brain activity. Of note also is the fact that for all 14 regions, the average timecourses of group 2 were of larger magnitude than those of group 1, suggestive, though not confirmatory, of some systematic difference between these particular combinations of subjects. The results of this analysis suggest that it is possible that the regions identified as exhibiting "sex differences" in Word/Nonword Read are confounded to some degree *either* by a legitimate though overlooked between-group factor or by artifactual statistical effects due to high between-subject variability (i.e. false positives). These two possibilities are neither mutually exclusive nor distinguishable with our current methods. As a general caution, one should be aware of the potential for group comparisons of fMRI BOLD signal data to be similarly confounded, particularly when those groups are comparable in number to the groups examined here. This caveat applies to any between-group analysis, whether those groups are divided on the basis of sex, disease status or any other level of comparison.

Aside from increasing subject numbers to decrease noise, how might one avoid the contamination of a group effect of interest by some unnoticed factor? The best safeguards are probably the old standbys, a prudent experimental design and explicit hypothesis testing. For example, the current studies might be extended and improved not only by substantially increasing the numbers of participants, but also by a more thorough collection of demographic and neuropsychological data for the purpose of identifying lurking variables, and the inclusion of hypothesis-driven ROI-based analyses for the purpose of increasing the power to detect subtle effects. Furthermore, any single result should be considered provisional until it has been replicated in, or generalized to, an independent investigation. These are exactly the challenges that permeate the literature on sex differences in language. Many studies seem to arise as orthogonal queries of experimental paradigms designed to ask some other question, which means that the task conditions are not optimized to probe specific hypotheses about male/ female differences. Partly as a result of incidental experimental design, many studies (like this one) either lack a clear hypothesis on the one hand or (like many other studies) pursue a single theory of lateralization on the other hand, both of which are vulnerable to misconstruing the data or missing out on potentially interesting effects.

A final issue worth reiterating is the fact that "a language task" is not a monolith, and it should ideally be evaluated as a series of component processes, especially in the context of group differences, since male and female neural activity may differ for some processes but not others. Unfortunately, the independent characterization of some task components is outside of the *temporal* resolution of fMRI, though such an evaluation is amenable to other techniques, as illustrated by a couple of clever, recent event-related potentials (ERP) studies (see Hill et al., 2006; Wirth et al., 2007).

Conclusion

In the present study, conservative quantitative and qualitative criteria were applied to whole-brain fMRI data from two distinct language tasks to identify a scattered handful of statistically significant regions of "sex difference" within each task separately. This initial step in our analysis mimics what many others have done in an attempt to investigate sex differences in language, and often these studies have arrived at confusingly disparate conclusions. We hoped to have shown that additional analytic steps are critical for appropriately interpreting the finding of "sex differences." First, our regions of "sex difference" largely failed to generalize across our two task sets, so that if they do truly represent areas that are recruited differentially by males and females, it must be specified that this conclusion has been validated only *within the context of particular task sets* and perhaps also within particular small groups of adult subjects. Also, in neither task set were the regions distributed throughout the brain in a way that suggests a particular sex-based lateralization difference, nor did most of the regions localize to functional brain areas the significance of which for language processing has been established. Thus our results do not support a particular model of sex differences in language processing, and they overwhelmingly do not specifically corroborate previous fMRI studies of sex differences in language. Perhaps most significantly, our discovery of comparable regions of "group difference" in both task sets undermines the overall reliability of the finding of sex differences, because it demonstrates clearly that the analysis of small samples for group differences may be confounded by spurious positive findings unrelated to the comparison of interest. Thus from both a cognitive neuroscience perspective and from the standpoint of experimental design, it is important to recognize the possibility that sex differences may exist in functional neuroimaging studies of language and other cognitive domains. It is equally important, however, to recognize that conclusions about between-group differences in fMRI studies should be made conservatively, acknowledging the possibility that unaccounted variability may contaminate the data. Specifically, generalization and replication of results is essential for avoiding misleading interpretations of results from single analyses.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Glossary

fMRI

functional magnetic resonance imaging

BOLD

blood oxygenation level-dependent

References

Box GEP. Some theorems on quadratic forms applied in the study of analysis of variance problems, II. Effects of inequality of variance and of correlation between errors in the two-way classification. Annals of Mathematical Statistics 1954;25:484–498.

Brown TT, Lugar HM, Coalson RS, Miezin FM, Petersen SE, Schlaggar BL. Developmental changes in human cerebral functional organization for word generation. Cerebral Cortex 2005;15:275–290. [PubMed: 15297366]

Buckner RL, Raichle ME, Petersen SE. Dissociation of human prefrontal cortical areas across different speech production tasks and gender groups. Journal of Neurophysiology 1995;74:2163–2173. [PubMed: 8592204]

Cahill L. Why sex matters for neuroscience. Nat Rev Neurosci 2006;7:477–484. [PubMed: 16688123]

Chen C, Xue G, Dong Q, Jin Z, Li T, Xue F, Zhao L, Guo Y. Sex determines the neurofunctional predictors of visual word learning. Neuropsychologia 2007;45:741–747. [PubMed: 16999980]

Church, JA.; Petersen, SE.; Schlaggar, BL. Regions showing developmental effects in reading studies show length and lexicality effects in adults. Society for Neuroscience; Atlanta, GA: 2006.

Clements AM, Rimrodt SL, Abel JR, Blankner JG, Mostofsky SH, Pekar JJ, Denckla MB, Cutting LE. Sex differences in cerebral laterality of language and visuospatial processing. Brain Lang 2006;98:150–158. [PubMed: 16716389]

Cohen MS, DuBois RM. Stability, repeatability, and the expression of signal magnitude in functional magnetic resonance imaging. J Magn Reson Imaging 1999;10:33–40. [PubMed: 10398975]

De Renzi E, Faglioni P, Ferrari P. The influence of sex and age on the incidence and type of aphasia. Cortex 1980;16:627–630. [PubMed: 7226860]

Frost JA, Binder JR, Springer JA, Hammeke TA, Bellgowan PS, Rao SM, Cox RW. Language processing is strongly left lateralized in both sexes. Evidence from functional MRI. Brain 1999;122(Pt 2):199–208. [PubMed: 10071049]

Goldstein JM, Seidman LJ, Horton NJ, Makris N, Kennedy DN, Caviness VS Jr, Faraone SV, Tsuang MT. Normal sexual dimorphism of the adult human brain assessed by in vivo magnetic resonance imaging. Cereb Cortex 2001;11:490–497. [PubMed: 11375910]

Grabowski TJ, Damasio H, Eichhorn GR, Tranel D. Effects of gender on blood flow correlates of naming concrete entities. NeuroImage 2003;20:940–954. [PubMed: 14568464]

Gur RC, Mozley LH, Mozley PD, Resnick SM, Karp JS, Alavi A, Arnold SE, Gur RE. Sex differences in regional cerebral glucose metabolism during a resting state. Science 1995;267:528–531. [PubMed: 7824953]

Harasty J, Double KL, Halliday GM, Kril JJ, McRitchie DA. Language-associated cortical regions are proportionally larger in the female brain. Arch Neurol 1997;54:171–176. [PubMed: 9041858]

Harrington GS, Farias ST. Sex differences in language processing: functional MRI methodological considerations. Magn Reson Imaging 2008;27:1221–1228.

Hill H, Ott F, Herbert C, Weisbrod M. Response execution in lexical decision tasks obscures sex-specific lateralization effects in language processing: evidence from event-related potential measures during word reading. Cereb Cortex 2006;16:978–989. [PubMed: 16177269]

Jaeger JJ, Lockwood AH, Van Valin RD Jr, Kemmerer DL, Murphy BW, Wack DS. Sex differences in brain regions activated by grammatical and reading tasks. Neuroreport 1998;9:2803–2807. [PubMed: 9760124]

Jancke L, Staiger JF, Schlaug G, Huang Y, Steinmetz H. The relationship between corpus callosum size and forebrain volume. Cereb Cortex 1997;7:48–56. [PubMed: 9023431]

Jansen A, Menke R, Sommer J, Forster AF, Bruchmann S, Hempleman J, Weber B, Knecht S. The assessment of hemispheric lateralization in functional MRI - Robustness and reproducibility. Neuroimage 2006;33:204–217. [PubMed: 16904913]

Kaiser A, Kuenzli E, Zappatore D, Nitsch C. On females' lateral and males' bilateral activation during language production: A fMRI study. Int J Psychophysiology 2007;63:192–198.

Kansaku K, Kitazawa S. Imaging studies on sex differences in the lateralization of language. Neurosci Res 2001;41:333–337. [PubMed: 11755219]

Kansaku K, Yamaura A, Kitazawa S. Sex differences in lateralization revealed in the posterior language areas. Cereb Cortex 2000;10:866–872. [PubMed: 10982747]

Kastrup A, Li TQ, Glover GH, Kruger G, Moseley ME. Gender differences in cerebral blood flow and oxygenation response during focal physiologic neural activity. J Cereb Blood Flow Metab 1999;19:1066–1071. [PubMed: 10532630]

Kertesz A, Benke T. Sex equality in intrahemispheric language organization. Brain Lang 1989;37:401–408. [PubMed: 2478252]

Kimura D. Sex, sexual orientation and sex hormones influence human cognitive function. Curr Opin Neurobiol 1996;6:259–263. [PubMed: 8725969]

Kitazawa S, Kansaku K. Sex difference in language lateralization may be task-dependent. Brain 2005;128:E30. [PubMed: 15845628]author reply E31

Lemaitre H, Crivello F, Grassiot B, Alperovitch A, Tzourio C, Mazoyer B. Age- and sex-related effects on the neuroanatomy of healthy elderly. Neuroimage 2005;26:900–911. [PubMed: 15955500]

Levin JM, Ross MH, Mendelson JH, Mello NK, Cohen BM, Renshaw PF. Sex differences in blood-oxygenation-level-dependent functional MRI with primary visual stimulation. Am J Psychiatry 1998;155:434–436. [PubMed: 9501761]

Marcar VL, Loenneker T, Straessle A, Girard F, Martin E. What the little differences between men and women tells us about the BOLD response. Magn Reson Imaging 2004;22:913–919. [PubMed: 15288131]

McAvoy MP, Ollinger JM, Buckner RL. Cluster size thresholds for assessment of significant activation in fMRI. NeuroImage 2001;13:S198.

McGlone J. Sex differences in the cerebral organization of verbal functions in patients with unilateral brain lesions. Brain 1977;100:775–793. [PubMed: 608120]

Miceli G, Caltagirone C, Gainotti G, Masullo C, Silveri MC, Villa G. Influence of age, sex, literacy and pathologic lesion on incidence, severity and type of aphasia. Acta Neurol Scand 1981;64:370–382. [PubMed: 7347996]

Petersen SE, Fox PT, Posner MI, Mintun M, Raichle ME. Positron emission tomographic studies of the cortical anatomy of single-word processing. Nature 1988;331:585–589. [PubMed: 3277066]

Phillips MD, Lowe MJ, Lurito JT, Dzemidzic M, Mathews VP. Temporal lobe activation demonstrates sex-based differences during passive listening. Radiology 2001;220:202–207. [PubMed: 11425998]

Pugh KR, Shaywitz BA, Shaywitz SE, Constable RT, Skudlarski P, Fulbright RK, Bronen RA, Shankweiler DP, Katz L, Fletcher JM, Gore JC. Cerebral organization of component processes in reading. Brain 1996;119(Pt 4):1221–1238. [PubMed: 8813285]

Rossell SL, Bullmore ET, Williams SC, David AS. Sex differences in functional brain activation during a lexical visual field task. Brain Lang 2002;80:97–105. [PubMed: 11817892]

Schlaefer TE, Harris GJ, Tien AY, Peng L, Lee S, Pearlson GD. Structural differences in the cerebral cortex of healthy female and male subjects: a magnetic resonance imaging study. Psychiatry Res 1995;61:129–135. [PubMed: 8545497]

Schlaggar, BL.; Brown, TT.; Lugar, HM.; Coalson, RS.; Petersen, SE. fMRI in performance-matched children and adults: Modality dependent and independent age-related differences in lexical processing. Society for Neuroscience; Orlando, FL: 2002.

Shaywitz B, Shaywitz S, Pugh K, Constable T, Skudlarski P, Fulbright R, Bronen R, Fletcher J, Shankweiler D, Katz L, Gore J. Sex differences in the functional organization of the brain for language. Nature 1995;373:607–609. [PubMed: 7854416]

Sommer IE, Aleman A, Bouma A, Kahn RS. Do women really have more bilateral language representation than men? A meta-analysis of functional imaging studies. Brain 2004;127:1845–1852. [PubMed: 15240433]

Sommer IE, Aleman A, Somers M, Boks MP, Kahn RS. Sex differences in handedness, asymmetry of the Planum temporale and functional language lateralization. Brain Res 2008;1206:76–88. [PubMed: 18359009]

Talairach, J.; Tournoux, P. Co-Planar Stereotaxic Atlas of the Human Brain. Thieme Medical Publishers, Inc.; New York: 1988.

Thirion B, Pinel P, Meriaux S, Roche A, Dehaene S, Poline J. Analysis of a large fMRI cohort: Statistical ad methodological issuess. NeuroImage 2007;35:105–120. [PubMed: 17239619]

Van Essen DC. A Population-Average, Landmark- and Surface-based (PALS) Atlas of Human Cerebral Cortex. Neuroimage 2005;28:635–662. [PubMed: 16172003]

Weiss EM, Siedentopf C, Hofer A, Deisenhammer EA, Hoptman MJ, Kremser C, Golaszewski S, Felber S, Fleischhacker WW, Delazer M. Brain activation pattern during a verbal fluency test in healthy male and female volunteers: a functional magnetic resonance imaging study. Neurosci Lett 2003;352:191–194. [PubMed: 14625017]

Wirth M, Horn H, Koenig T, Stein M, Federspiel A, Meier B, Michel CM, Strik W. Sex differences in semantic processing: event-related brain potentials distinguish between lower and higher order semantic analysis during word reading. Cereb Cortex 2007;17:1987–1997. [PubMed: 17116651]

Witelson SF. Hand and sex differences in the isthmus and genu of the human corpus callosum. A postmortem morphological study. Brain 1989;112:799–835. [PubMed: 2731030]

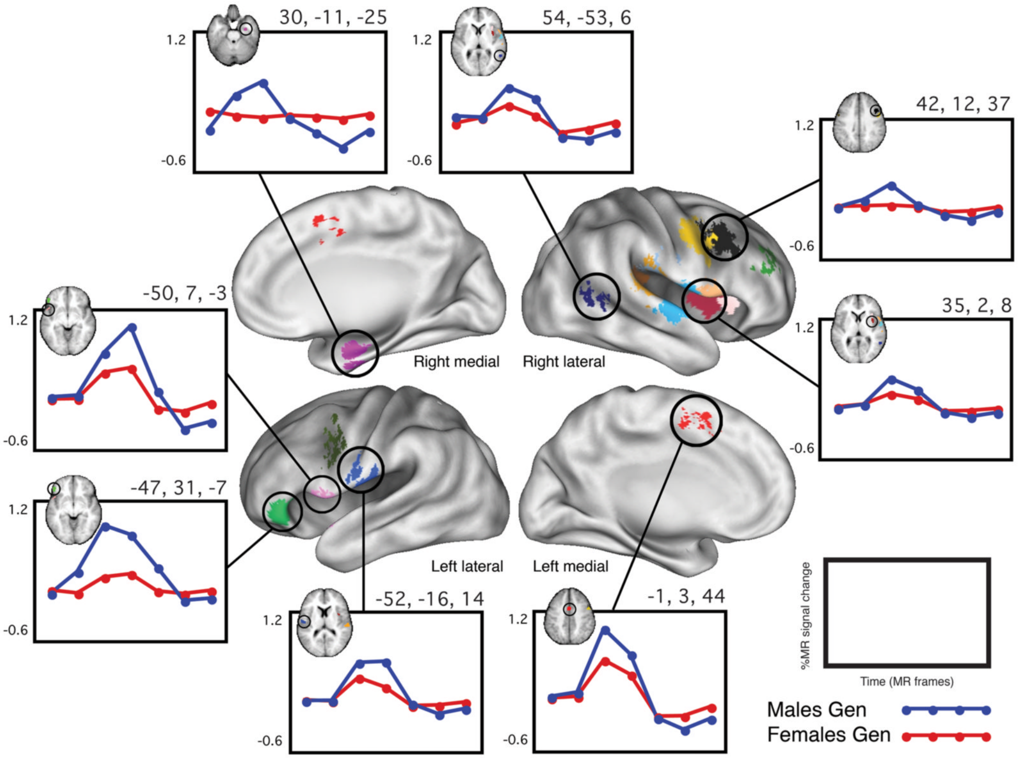

Figure 1. "Sex differences" in Word Generate

Whole-brain surface rendering of the 17 regions in which there was a significant sex x time effect in the Word Generate set that also met the post hoc criteria described in the text. Shown are group-average timecourses for eight representative regions. In all regions, the magnitude of the average timecourse of the males was larger than that of the females. In the upper left-hand corner of each timecourse plot is a depiction of the corresponding region as it appears in a transverse slice of the common MRI anatomical atlas into which the data was transformed. These and all other surface-rendered images were created using CARET software and population-average, landmark-, surface-based atlases (Van Essen, 2005). In this and all subsequent figures, brain surfaces are labeled as right or left and lateral, medial, anterior or posterior. Throughout the figures, circles are used for timecourses from Word Generate task and triangles are used for timecourses from Word/Nonword Read.

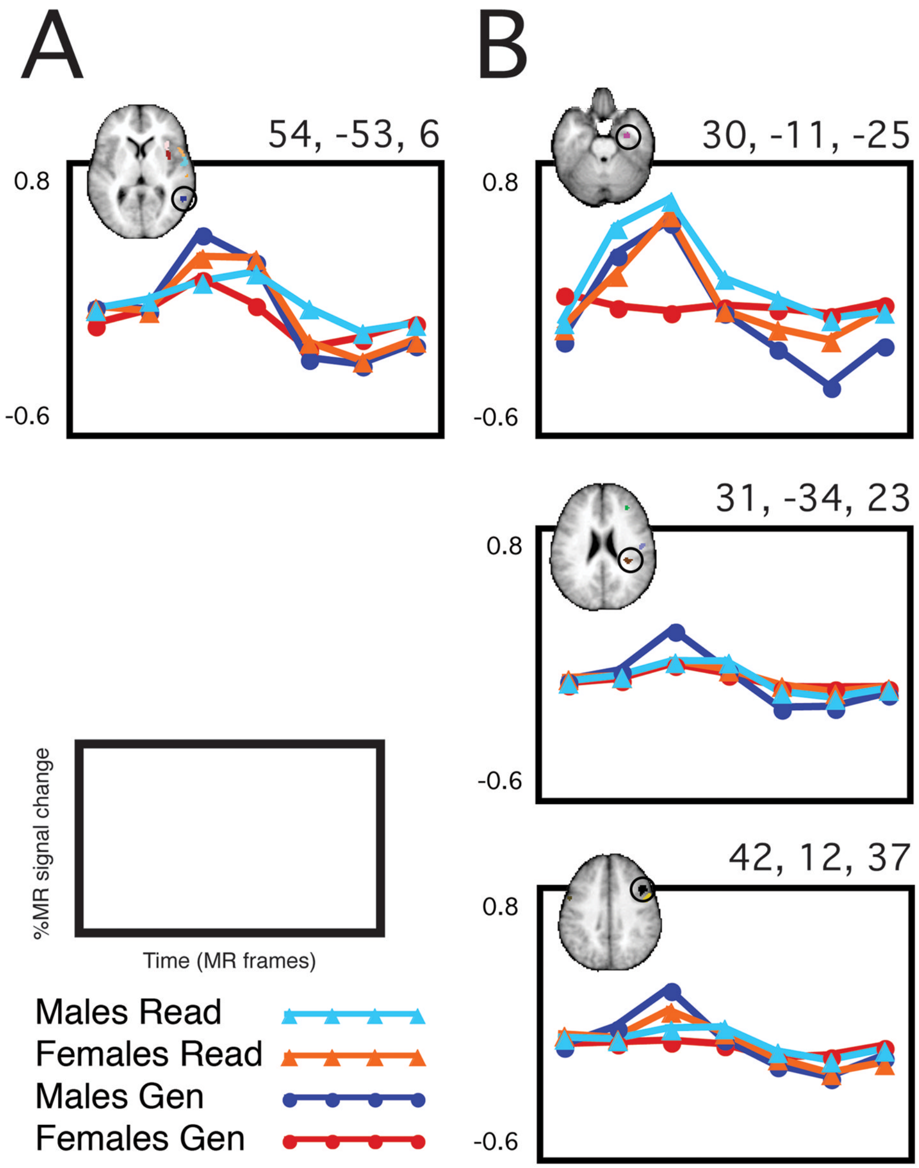

Figure 2. Generalizability of "sex differences"

In column A, the single region is shown in which the sex effect observed in the Word Generate set of tasks is additionally significant in the Word/Nonword Read set of tasks. The region is in the right middle temporal gyrus, and as demonstrated, the direction of the sex effect switches between the two task sets (i.e. M>F in Word Generate, F>M in Word/Nonword Read). In column B, 3 example regions are shown in which the sex difference identified in the Word Generate task set did not show a similarly significant sex difference in Word/Nonword Read.

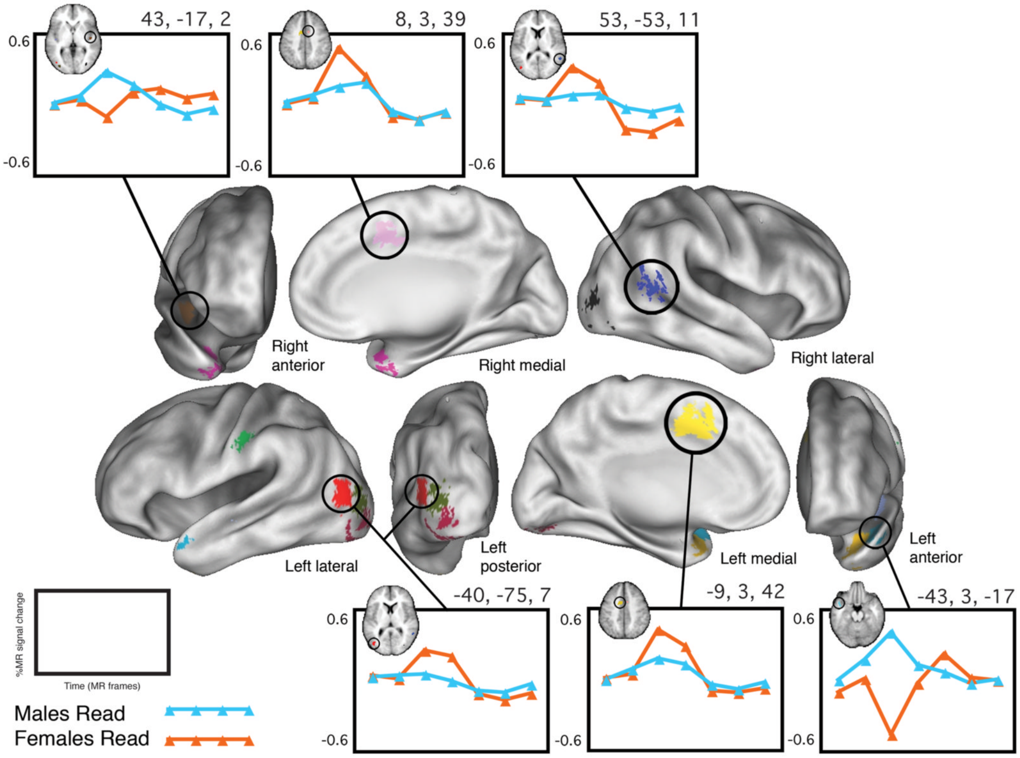

Figure 3. “Sex differences” in Word/Nonword Read

Whole-brain surface rendering of the 13 regions in which there was a significant sex x time effect in the Word/Nonword Read set of tasks that also met the post hoc criteria described in the text. Shown are group-average timecourses for 6 representative regions. In some regions, the magnitude of the average timecourse of the males was larger than that of the females, while in other regions the inverse was true.

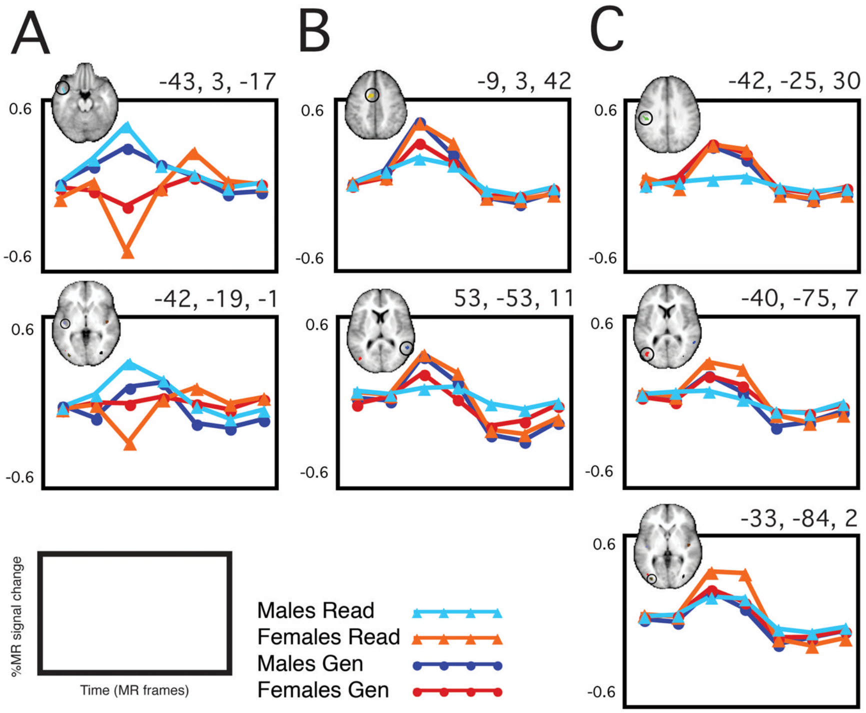

Figure 4. Generalizability of "sex differences"

In column A, 2 regions are shown in which the sex effect observed in Word/Nonword Read is also significant in Word Generate AND the effect was in the same direction in the two task sets (M>F). In column B, 2 regions are shown in which the sex effect observed in Word/Nonword Read was also significant in Word Generate, but in the opposite direction (F>M became M>F). In column C, 3 example regions are shown in which the sex difference identified in Word/Nonword Read did not show a similarly significant sex difference in Word Generate.

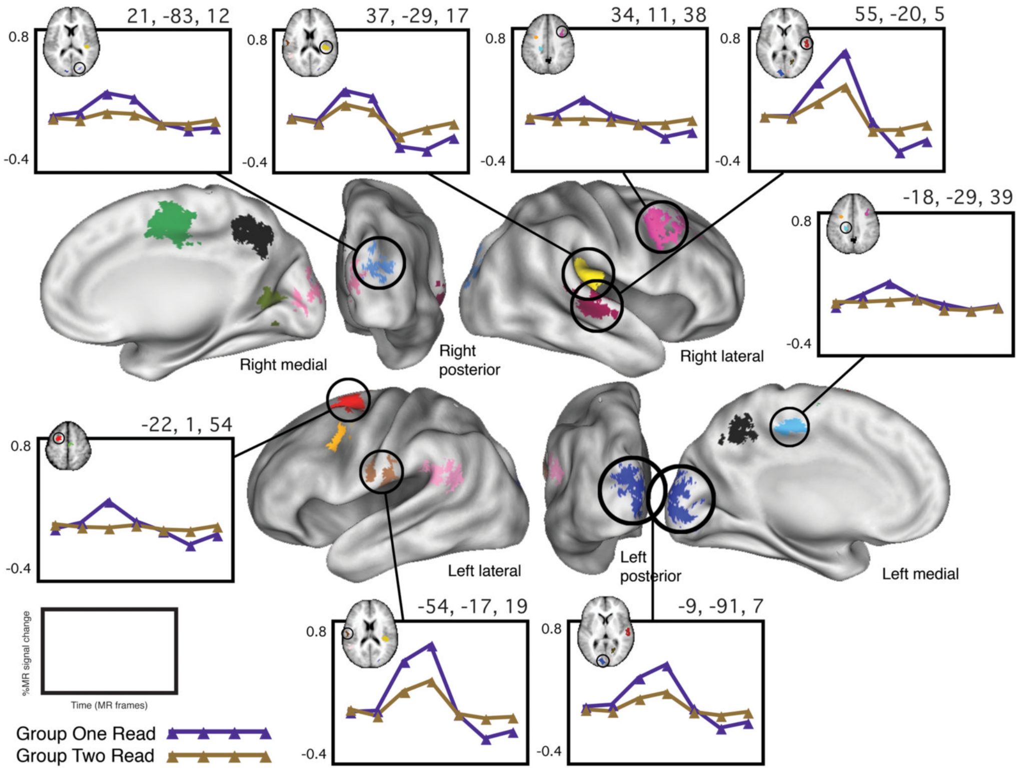

Figure 5. "Group differences" in Word/Nonword Read

Whole-brain surface rendering of the 14 regions in which there was a significant sex x time effect in the Word/Nonword Read set of tasks that also met the post hoc criteria described in the text. Shown are group-average timecourses for 8 representative regions, with group 1 in purple and group 2 in gold. In all regions shown and in fact in all regions identified, the magnitude of the average timecourse of group 1 was larger than that of group 2.

Table 1

Demographic and task performance characteristics of male/female subject groups from both task sets (SD in parentheses)

	Word Generate		Word/Nonword Read	
	Females	Males	Females	Males
n	13	13	10	10
Age (years)	24.6[a] (3.7)	25.2[a] (3.4)	25[a] (3.1)	25.5[a] (1.5)
% Correct	90.5[a] (4.1)	88.9[a] (6.6)	99.2[d] (0.8)	99.3[d] (0.9)
Reaction Time (ms)	1270[a] (222)	1330[a] (211)	710[a] (83)	706[a] (76)
Movement (RMS avg)	0.24[b] (0.04)	0.27[b] (0.07)	0.26[a] (0.00)	0.26[a] (0.00)
Estimated Total Intracranial Volume (cc)	1551[c] (150)	1629[c] (185)	1422 (77)	1542 (66)

[a] Similar across sexes (each task considered separately) by ANOVA, $p > 0.40$

[b] Similar across sexes (each task considered separately) by ANOVA, $p > 0.15$

[c] Similar across sexes (each task considered separately) by ANOVA, $p > 0.20$

[d] Similar across sexes by ANOVA, $p > 0.90$

Table 2

Demographic and task performance characteristics of pseudorandom “control” groups from the Word Generate task set (SD in parentheses)

	Group One	**Group Two**
n Males	7	6
n Females	6	7
Age (years)	25.2[a] (4.0)	24.6[a] (3.0)
% Correct	88.3[a] (7.4)	89.3[a] (5.1)
Reaction Time (ms)	1460[a] (216)	1411[a] (261)
Movement (RMS avg)	0.26[a] (0.07)	0.25[a] (0.05)
Estimated Total Intracranial Volume (cc)	1592[b] (176)	1589[b] (171)

[a] Similar across groups by ANOVA, $p > 0.55$

[b] Similar across groups by ANOVA, $p > 0.95$

Table 3

Demographic and task performance characteristics of pseudorandom "control" groups from the Word/Nonword Read task set (SD in parentheses)

	Group One	**Group Two**
n Males	5	5
n Females	5	5
Age (years)	25.2[a] (2.0)	25.3[a] (2.9)
% Correct	99.1[b] (0.9)	99.4[b] (0.7)
Reaction Time (ms)	710[a] (79)	707[a] (81)
Movement (RMS avg)	0.28[b] (0.06)	0.25[b] (0.07)
Estimated Total Intracranial Volume (cc)	1482[a] (76)	1482[a] (112)

[a] Similar across groups by ANOVA, $p > 0.90$

[b] Similar across groups by ANOVA, $p > 0.30$

Table 4

Regions that exhibited a significant sex x time interaction in the Word Generate task set (for all regions listed, $p < 0.0001$ for sex x time)

X	Y	Z	Location	Approx. BA	Volume (cm^3)	TP3 diff. (F-M)
Left						
−1	3	44	Cingulate Gyrus	24	0.86	−0.47
−51	1	35	Precentral Gyrus	6	0.33	−0.36
−52	−16	14	Postcentral Gyrus	40	0.92	−0.22
−50	7	−3	Superior Temporal Gyrus	22	0.40	−0.29
−47	31	−7	Inferior Frontal Gyrus	47	0.72	−0.77
Right						
42	12	37	Middle Frontal Gyrus	9	0.66	−0.31
48	2	36	Precentral Gyrus	6	1.01	−0.44
31	34	26	Frontal subgyral	--	0.36	−0.29
31	−34	23	Frontal subgyral	--	0.40	−0.2
49	−19	20	Postcentral Gyrus	40	0.69	−0.22
58	−24	13	Superior Temporal Gyrus	22	0.89	−0.22
35	2	8	Claustum	--	0.51	−0.23
54	−53	6	Middle Temporal Gyrus	37	0.27	−0.27
33	15	6	Claustum	--	0.27	−0.27
50	8	5	Precentral Gyrus/Insula	--	0.42	−0.31
56	−9	4	Superior Temporal Gyrus	22	1.00	−0.33
30	−11	−25	Parahippocampal Gyrus	35/36	0.39	−0.53

Table 5

Regions that exhibited a significant sex x time interaction in the Word/Nonword Read task set (for all regions listed, $p < 0.0005$ for sex x time)

X	Y	Z	Location	Approx. BA	Volume (cm^3)	TP3 diff. (F-M)
Left						
−9	3	42	Cingulate Gyrus	24	0.88	0.29
−42	−25	30	Postcentral Gyrus	2	0.35	0.28
−40	−75	7	Medial Occipital Gyrus	19	0.41	0.23
−33	−84	2	Occipital Gyrus	18	0.23	0.2
−42	−19	−1	Posterior Insula	--	0.27	−0.65
−26	−82	−8	Inferior Occipital Gyrus	18	0.57	0.23
−43	3	−17	Middle Temporal Gyrus	21	0.38	−1.02
−36	0	−25	Middle Temporal Gyrus	21	0.44	−0.93
Right						
8	3	39	Cingulate Gyrus	24/32	0.27	0.38
53	−53	11	Middle Temporal Gyrus	39	0.27	0.28
43	−17	2	Posterior Insula	--	0.22	−0.45
32	−79	−1	Inferior Occipital Gyrus	18	0.34	0.14
33	4	−28	Superior Temporal Gyrus	38	0.38	−1.22

Table 6

Classification of regions of sex x time interaction from each study (WG = Word Generate, W/NW R = Word/Nonword Read) into four categories of observation, along with the hypotheses supported by each category

Category	WG[a]	W/NW R	Hypotheses Supported
M>F in the left	4	3	M more L-lateralized and/or F more R-lateralized
F>M in the right	0	3	F more R-lateralized and/or M more L-lateralized
M>F in the right	12	2	M more R-lateralized and/or F more L-lateralized
F>M in the left	0	5	F more L-lateralized and/or M more R-lateralized

[a] The Word Generate task set generated 17 regions of sex difference, but one of these regions was located on the midline and is therefore not included in the categorization.

Table 7

Regions that exhibited a significant “group” x time interaction in the Word/Nonword Read task set (for all regions listed, $p < 0.0001$ for group x time)

X	Y	Z	Location	Approx. BA	Volume (cm^3)	T3 diff. (1–2)
Left						
−22	1	54	Middle Frontal Gyrus	6	1.05	0.2605
−18	−29	39	Cingulate Gyrus	31	0.53	0.1781
−29	0	36	Subgyral	--	0.44	0.2234
−48	−48	20	Superior Temporal Gyrus	22	0.78	0.1516
−54	−17	19	Postcentral Gyrus	40	0.42	0.2945
−9	−91	7	Cuneus	17	1.22	0.2025
Right						
2	−55	41	Precuneus	7	1.52	0.1897
5	−14	48	Medial Frontal Gyrus	6	1.70	0.2032
9	−88	9	Cuneus	17	0.30	0.1569
19	−65	5	Lingual Gyrus	18	0.53	0.1312
21	−83	12	Cuneus	18	0.48	0.1904
34	11	38	Middle Frontal Gyrus	9	0.87	0.1999
37	−29	17	Superior Temporal Gyrus/Lateral Sulcus	42	2.13	0.1362
55	−20	5	Superior Temporal Gyrus	22	1.42	0.2004