NeuroImage 55 (2011) 705–712

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

fMRI item analysis in a theory of mind task

David Dodell-Feder [a,*], Jorie Koster-Hale [b], Marina Bedny [b], Rebecca Saxe [b]

[a] Department of Psychology, Harvard University, 33 Kirkland St., Cambridge, MA 02138, USA
[b] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, 43 Vassar Street, Cambridge, MA 02139, USA

ARTICLE INFO

Article history:
Received 6 August 2010
Revised 8 December 2010
Accepted 13 December 2010
Available online 21 December 2010

Keywords:
Item analysis
Theory of mind
Functional magnetic resonance imaging
Social cognition
Temporo-parietal junction

ABSTRACT

Conventional analyses of functional magnetic resonance imaging (fMRI) data compare the brain's response to stimulus categories (e.g., pictures of faces, stories about beliefs) across participants. In order to infer that effects observed with the specific items (a particular set of pictures or stories) are generalizable to the entire population (all faces, or all stories about beliefs), it is necessary to perform an "item analysis." Item analyses may also reveal relationships between secondary (non-hypothesized) features of the items and functional activity. Here, we perform an item analysis on a set of stories commonly used for localizing brain regions putatively involved in Theory of Mind (ToM): right and left temporo-parietal junction (RTPJ/LTPJ), precuneus (PC), superior temporal sulcus (STS) and medial prefrontal cortex (MPFC). We address the following questions: Do brain regions that comprise the ToM network respond reliably across items (i.e. different stories about beliefs)? Do these brain regions demonstrate reliable preferences for items within the category? Can we predict any region's response to individual items, by using other features of the stimuli? We find that the ToM network responds reliably to stories about beliefs, generalizing across items as well as subjects. In addition, several regions in the ToM network have reliable preferences for individual items. Linguistic features of the stimuli did not predict these item preferences.

Introduction

Consider the following scenario: a researcher wishes to investigate brain regions recruited for Theory of Mind (ToM), i.e. the ability to attribute and reason about the mental states of other individuals. To that end, she creates two sets of stories, one set describing beliefs held by different protagonists and a set of control stories not including beliefs. Using a standard analysis strategy, each brain region's response to belief versus control stories is evaluated for significance, by comparing the average effect size (belief>control) to the variability of the effect across subjects. The researcher concludes that the resulting brain network is recruited more for processing/representing the *category* of stories about mental states than the *category* of control stories. This conclusion, however, goes beyond what was explicitly tested. From a standard analysis, she can only conclude that a contrast between these exact stimuli will on average reveal the same brain regions in a different group of subjects. She cannot conclude that these brain regions will reliably be recruited for other (or all) stories about beliefs.

This example illustrates the "fixed-effects fallacy" (Clark, 1973) or the unfounded inference that conclusions about items sampled from a population generalize to the entire item population. Up until the late 1990s, neuroimagers fell victim to this same problem with subject-wise analyses by treating subjects as "fixed" variables. In treating subjects as a "fixed" variable, the variability of an effect across subjects is not taken into account, which is especially problematic if there is substantial variability in the effect size across participants. Consider an instance in which only one out of five subjects shows a very large effect for some condition, and the other subjects show no effect. Averaging the effect across subjects, the entire group will appear to exhibit a medium-sized effect. Therefore, in modern neuroimaging analyses, random-effects analysis is used, which compares an effect to its variability across subjects. This strategy allows researchers to test whether their findings will generalize to the population from which the subjects are sampled (Friston et al., 1999; Holmes and Friston, 1998). In the example above, treating subjects as a random variable would reveal that the apparent medium-sized average effect is not reliable across subjects.

This same issue exists at the item level. Perhaps only one or two of the stories about beliefs recruit a brain region very strongly (due to some theoretically irrelevant feature), while the remaining stories have no effect. Averaging across items, the group of belief stories will appear to recruit this brain region to a moderate degree. In order to make (theoretically more important) generalizations about the category to which an item belongs, one needs to evaluate the effect size relative to the variation of the effect across items, in an item-wise random-effects analysis. Only upon doing so can one validly conclude that the given brain network is reliably recruited for the item *category*.

* Corresponding author.
E-mail address: feder@fas.harvard.edu (D. Dodell-Feder).

1053-8119/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2010.12.040

Item analysis has yet to become common practice for neuroimagers despite the fact that it has the potential to reveal theoretically relevant distinctions and is known to be feasible. For example, Bedny et al. (2007) tested the hypothesis that distinct brain regions in the frontal and temporal cortex process nouns and verbs (grammatical-class effect). Using subject-wise analysis, they found a differential brain response to verbs versus nouns in the left posterior temporal lobe and left inferior frontal lobe. Conversely, item analysis revealed that only the effect in the temporal lobe was reliable across items, whereas the effect in the frontal lobe was not. This finding suggests that the grammatical-class effect in the frontal region was specific to the nouns and verbs used in the experiment; the effect would not necessarily generalize to other items from the same categories. Therefore, it would be erroneous to conclude that a region within the frontal lobe was specialized for processing verbs versus nouns, despite the significant subject-wise effect. In that case, the item analysis reconciled conflicting findings regarding the role of the prefrontal cortex in processing specific aspects of language (Bedny and Thompson-Schill, 2006; Davis et al., 2004; Shapiro et al., 2006; Tyler et al., 2001, 2004).

The first goal of the current study was therefore to apply item-wise random effects analyses to an experimental paradigm frequently used to identify brain regions involved in ToM. Following in the tradition of developmental investigations of ToM (Wimmer and Perner, 1983), neuroimaging studies have often used "false belief" stories to test belief reasoning. In these stories, a protagonist performs an action based on a belief that is false (e.g., Maxi believes his chocolate is in the green drawer, but his mother moved it to the blue drawer). Participants reading these stories are thus required to represent the outdated belief of the protagonist in order to understand their actions (e.g., looking in the green drawer even though the chocolate is actually in the blue drawer). These stories are contrasted with "false photograph" stories, which also require the representation of false or outdated content (e.g., an old photograph that no longer accurately depicts the landscape of a burgeoning city). False belief and false photograph stories are therefore matched in their general difficulty, logical complexity, and inhibitory demands, but differ in the need to think about someone's thoughts. Accordingly, a set of stories about false beliefs and false photographs (Saxe and Kanwisher, 2003) is commonly employed across a range of studies to identify brain regions in the so-called "ToM network": right and left temporo-parietal junction (RTPJ/LTPJ), superior temporal sulcus (STS), precuneus (PC) and medial prefrontal cortex (MPFC) (Kliemann et al., 2008; Mitchell, 2008; Saxe and Powell, 2006; Saxe et al., 2006; Saxe and Wexler, 2005; Scholz et al., 2009; Young et al., 2007, 2010a). It is therefore theoretically important to establish that these regions' recruitment generalizes beyond the specific commonly used stimuli. Here, we used an item analysis to formally test whether the brain response to these specific stories about false beliefs can be generalized to the category of such stories.

Item analyses also have a second advantage. An item analysis produces an estimate of the response in each brain region, to each specific stimulus. If a region has a reliable preference for specific items within a category, this preference may provide a clue about the region's function. Every item in an fMRI experiment can be characterized on multiple different dimensions or features (e.g., for stories about beliefs, the number of people or mental states mentioned, the degree of syntactic complexity, the specific context of the story, etc.). It may be possible to determine which item dimensions or features best predict each region's response. These dimensions or features may be confounds, which explain away previous categorical effects, or they may confirm and expand prior results, by allowing a higher-resolution picture of the region's processing, within stimulus categories. One specific concern is that activity in the ToM network is best accounted for by linguistic features of the stories that are concomitant with the presence of belief information. The extent to which these factors account for activity in ToM brain regions can be evaluated within a single paradigm by analyzing data at the item level.

In the current paper, we used item analyses to investigate the ToM network. We ask the following questions: (1) Does item-wise analysis replicate subject-wise analysis? That is, does the response in ToM brain regions generalize across items within each stimulus category (false beliefs and false photographs)? (2) Do ToM brain regions demonstrate a reliable preference for items within each category? And, (3) what features of the items account for activity/stable preference for items in these regions? We characterized the stimuli used by Saxe and colleagues (e.g., Saxe and Powell, 2006; Saxe and Wexler, 2005; Young et al., 2007, 2010b) on several dimensions (words per story, number of people per story, Flesch reading ease level, visualizability and several other linguistic aspects), and asked whether any of these features could predict differences between items in the response of ToM brain regions.

Methods

Participants

Sixty-two right-handed naïve adults (M age $= 22 \pm 4$ years, range = 18–35, 35 females) participated in the experiment for payment. All participants were native English speakers, had normal or corrected-to-normal vision, and gave written informed consent in accordance with the requirements of the internal review board at MIT. For a portion of the analyses, the participants were split into two independent groups. The two groups did not differ in age or gender (Group 1: $n = 32$, M age $= 22 \pm 3$, 17 females; Group 2: $n = 30$, M age $= 23 \pm 5$, 18 females; Age: $t(60) = 1.18$ $p = .25$; Gender: $\chi^2(1) = .298$ $p = .59$).

Stimuli

Stimuli consisted of 20 stories in each of two conditions: (1) stories describing false beliefs (BELIEF) and (2) stories describing outdated (i.e. false) photographs and maps (PHOTO; Table 1; Saxe and Kanwisher, 2003, Experiment 2. See Supplementary Table 1 for a complete list of BELIEF and PHOTO items along with the corresponding beta values from the RTPJ, LTPJ, PC, RpSTS, RTP, DMPFC and MMPFC). Both sets of stories required participants to represent false content; the critical difference was in the type of false content represented (i.e., a belief versus a photograph/map). Stories were followed by a true/false question that referred either to the situation in reality or to the false representation. There were an equal number of questions that referred to the reality and representation in each condition, the order of which was counterbalanced within and across runs. Participants responded to the question with a button response. Reaction time (RT) data were collected during the scan.

In the scanner, stories were presented visually for 10 s, followed by the true/false question for 4 s and finally 12 s of rest (a black screen). Stories were presented in a pseudo-random order with the order of conditions counterbalanced across runs and participants. Eight stories were presented in each of 4 runs (4 stories per condition per run) for a total run time of 14 min and 24 s. The text of each story was presented in 30-point white font on a black background via Matlab 7.6 running on an Apple MacBook Pro.

fMRI data acquisition and analysis

Participants were scanned at 3 T (at the MIT scanning facility in Cambridge, MA) with a 12-channel head coil using thirty 4-mm-thick near axial slices covering the whole brain. Standard echoplanar imaging procedures were used (TR = 2 s, TE = 30 ms, flip angle = 90°). The first

Table 1
BELIEF and PHOTO items which elicited the highest and lowest response from the right temporo-parietal junction (RTPJ).

	Item	RTPJ beta value
BELIEF highest	The morning of high school dance Sarah placed her high heel shoes under her dress and then went shopping. That afternoon, her sister borrowed the shoes and later put them under Sarah's bed. *Sarah gets ready assuming her shoes are under the dress.*	.446
BELIEF lowest	When Jeff got ready this morning, he put on a light pink shirt instead of a white one. Jeff is colorblind, so he can't tell the difference between subtle shades of color. *In reality, Jeff's shirt is pink.*	−.059
PHOTO highest	The traffic camera snapped an image of the black car as it sped through the stoplight. Soon after, the car was painted red and the license plates were changed. *According to the traffic camera, the car is black.*	.241
PHOTO lowest	Old maps of the islands near Titan are displayed in the Maritime museum. Erosion has since taken its toll, leaving only the three largest islands. *Near Titan today there are many islands.*	−.470

Each item was followed by a statement (italicized text) which subjects evaluated as "True" or "False" with a button press response.

4 volumes of each run consisted of dummy scans, which were not analyzed, to insure steady-state magnetization. These sequences used PACE online motion correction, which adjusts the slice acquisitions during scanning to correct for head movement up to 8 mm.

FMRI data were analyzed using SPM2 (http://www.fil.ion.ucl.ac.uk/spm) and custom software. Each participant's data were motion corrected and normalized onto a common brain space (Montreal Neurological Institute, MNI template). Data were smoothed using a Gaussian filter (full width half maximum = 5 mm) and were high-pass filtered during analysis. The experiment used a block design and was modeled using a boxcar regressor convolved with a standard hemodynamic response. We performed both subject- and item-wise, whole-brain and ROI analyses following the general procedure of Bedny et al. (2007).

For the subject-wise analysis, individual subject first-level models were created using a general linear model with conditions (BELIEF, PHOTO) as covariates of interest. Second-level random effects analysis was performed on the contrast images generated from the first-level models for BELIEF>PHOTO. Here, the variance between the effect sizes for BELIEF>PHOTO across subjects was used to calculate a t statistic at each voxel, treating subjects as a random variable. The whole-brain BELIEF>PHOTO contrast was FDR corrected for multiple comparisons at $p<.01$.

In the item-wise analysis, individual subject first-level models were creased using a general linear model with each of the 40 items entered as a covariate of interest. This yielded 40 beta maps per subject, which were averaged across subjects to obtain a single beta map per item. Second-level random effects analysis used these values to compute condition differences (BELIEF>PHOTO) via a two-sample t-test at every voxel, treating the items as a random variable. The whole-brain BELIEF>PHOTO contrast was FDR corrected for multiple comparisons at $p<.01$.

For ROI analysis, participants were divided into two independent groups (Group 1 $n=32$, Group 2 $n=30$). Regions identified from the subject-wise random effects analysis of Group 1 were used to extract data from Group 2 participants. Each of the 7 ROIs – RTPJ, LTPJ, PC, right posterior superior temporal sulcus (RpSTS), right temporal pole (RTP), dorsal medial prefrontal cortex (DMPFC), and middle medial prefrontal cortex (MMPFC) – was defined as a sphere of contiguous voxels 9 mm from the peak that were significantly more active ($p<.0001$, uncorrected, $k>10$) for the BELIEF>PHOTO contrast in Group 1. To investigate whether item analysis replicated the findings of previous subject-wise analysis, t-tests were conducted on the average beta values for all BELIEF versus all PHOTO stories in Group 2, in each ROI. All peak voxels are reported in MNI coordinates.

Testing the reliability of the ROI item response: item-wise correlation analysis

To investigate whether a larger response to particular items was consistent across groups of participants, we performed an item-wise correlation analysis. In this procedure, ROIs (RTPJ, LTPJ, PC, RpSTS, RTP, DMPFC, MMPFC) were defined from the 62 subject subject-wise random effects analysis of BELIEF>PHOTO (as a sphere of contiguous voxels 9 mm from the peak at $p<.0001$, uncorrected, $k>10$), and used to extract beta values for each story for each individual. Then, the 62 participants were randomly split into two groups 100 times, and for each iteration, the average beta per item was calculated in each group. These beta values were correlated across groups within each ROI, separately by condition, creating 100 Pearson r values, which allow us to estimate the mean and variance of the reliability of item preferences within condition, across participants. Note that for this procedure, we are measuring the item response correlation between the two groups *within* each condition. Given that the ROIs are defined based on between condition differences (i.e., BELIEF>PHOTO), the measured correlation is independent of the ROI selection procedure.

Item coding and regression analyses

Beta values from Group 2's data (extracted from ROIs defined in Group 1) only were used for all subsequent analyses. To address the question of which features of the item accounted for activity in the ROIs, stories were coded for several linguistic, social and general conceptual as well as perceptual features. Thirteen linguistic features were coded: number of words per story, Flesch reading ease, anaphor reference, causal content, causal cohesion, lexical concreteness, negation, noun-phrase modification, higher-level constituency, number of words before the main verb, intentional content, attitude predication and modality; four social features: number of people per story and the extent to which the items made readers think about the mental states, deception, and social status; and two general conceptual and perceptual features: the extent to which the items made readers think about physical causality and imagine/visualize the events of the story while reading. Altogether, 19 item features (detailed below) were coded.

First, a group of features of each story was estimated by collecting independent participants' ratings of the stories on five dimensions evaluating the extent to which our items made readers think about (i) mental states, (ii) deception, (iii) social status, (iv) physical causality and (v) the extent to which the story made participants image/visualize events. These data were collected via an online study using Amazon's Mechanical Turk (M-Turk). Participants were instructed to answer one of the following questions using a Likert scale from 1 (very little) to 7 (very much): (i) "To what extent did this story make you think about someone's experiences, thoughts, beliefs, desires, and/or emotions?", (ii) "To what extent did this story make you think about someone being deceived or fooled by someone or something?", (iii) "To what extent did this story make you think about someone's appearance, social status, or role in society?", (iv) "To what extent did this story make you think about physical objects, and physical causal interactions?", and (v) "To what extent did you

picture or imagine the events of the story happening as you read?" These five questions were individually paired with each of the 40 stories, for 160 separate single-question surveys. Participants were allowed to do as many surveys as they wanted, for a total of 20 participants per survey question (except question (v), which was answered by 50 participants).

The number of words in each story, and the number of people mentioned in each story, were counted by the experimenters.

To characterize the other linguistic features of the items, we used the Coh-Metrix, a metric designed to measure computational cohesion and text difficulty, based on measures of syntactic, semantics, and representational difficulty (Graesser et al., 2004; McNamara et al., 2002, 2006). The Coh-Metrix includes 54 linguistic features. We chose twelve specific features to analyze in this experiment, using two criteria. First, the feature must be variable across the stimulus items. For example, we excluded any feature that was absent from all of the items. Second, we prioritized features that are hypothesized to modulate activity in the ToM network (e.g., amount of intentional information, number of verbs expressing belief or desire information that use sentential complement clause syntax, general text comprehensibility/difficulty). Below we describe each of the final eleven measures in detail.

One of the features, *Flesch reading ease*, measures comprehension difficulty by calculating average sentence length and the number of syllables per word. Higher Flesch reading ease scores indicates easier text and increased readability.

Four of the features measure aspects of semantic cohesion: anaphor reference, causal content, causal cohesion, and lexical concreteness. Increased semantic cohesion has been shown to correlate with increased ease of processing, mental modeling, and overall comprehension. *Anaphor reference* measures the number of times a single idea, item, or action is referred to throughout the text, indexed by the number of anaphors in each story (such as pronouns *it*, *he*, *she*, and ellipsis markers, *did*, *was*) that refer to a constituent that appeared previously in the story. *Causal content* is a measure of the extent to which a story conveys causal information, which is correlated with coherence, comprehensibility and ease of mental modeling (Graesser et al., 1997). This index is estimated by counting the number of main verbs in the text that are categorized as causal in WordNet (Fellbaum, 1998; Miller et al., 1990). Causal information alone does not guarantee comprehensibility — a reader must also be able to coherently connect causal events and actors. Thus, we also measured *causal cohesion*, the ratio of causal particles to causal verbs, for each story. Causal particles include *because*, *due to*, *if*, *thus*, *unless*. A low causal cohesion score correlates with lack of general textual cohesion and increased difficulty of comprehension. Finally, *lexical concreteness* is a measure of the overall concreteness of the words in the story. Concreteness was computed by finding the mean concreteness of the content words (primarily nouns and verbs) in the stimuli, based on human rating data from the MRC Psycholinguistics Database (Coltheart, 1981).

Four of the features measure aspects of syntactic complexity: negation, noun phrase modification, higher-level constituency, and words before the main verb. Increased syntactic complexity is correlated with increased working memory load, slower processing, and difficulty in comprehension. *Negation* is one index of syntactic complexity, measured by counting the negative expressions in the text, such as *no*, *not*, *un-*, *without*. *Noun phrase modification*, measures the mean number of modifiers, such as adjectives, adverbs, and determiners, per noun phrase, indexing the difference between *most of the very fluffy and drooling puppies* (seven modifiers) and *the puppies* (one modifier). A more general feature, *higher-level constituency*, counts the number of complex units per sentence (such as phrases and clauses) controlling for word count. Sentences with a high number of high-level constituents are often structurally dense, with unusual syntax or embedding, and are generally harder to process. Fourth, sentences with a high number of *words before the main verb* have been shown to be taxing on working memory, making this an additional index of the working memory load for a story.

Finally, three linguistic features were specifically chosen to be relevant for ToM: intentional content, attitude predication, and modality. First, *Intentional content*, like causal content, is a measure of coherence within the text. However, rather than looking at causal relationships, this index measures the incidence of intentional actions and events based on the number of intentional main verbs, categorized based on WordNet ratings. The higher the incidence of intentional actions in a text, the more likely the text is to convey goal-driven content. If ToM regions are sensitive to goal-driven action or desire, they might show sensitivity to variation along this dimension. Second, *Attitude Predication* measures the number of verbs (predicates) that express thoughts and desires, such as *believe*, *want*, and *think*, per 100 words. Acquisition of these verbs has been implicated in development of ToM, both because they allow the expression of mental states and because of their syntax (e.g. opacity: a sentence containing an attitude predicate can be false without affecting the truth of the full sentence, as in *John thinks that there are unicorns is his yard*; de Villiers and Pyers, 2002). Third, *Modality* measures the number of operators that deal with possibility and necessity, such as the words *possibly*, *should*, *might*, and *must*, per 100 words. This class of words has a very similar set of properties as attitude predicates, including opacity and sentential embedding. However, this class crucially does not deal with mental states or desires, and thus serve as a useful comparison for effects of attitude predication.

(For a correlation matrix of pair-wise correlations among all of these features, see Supplementary Table 3).

We performed stepwise forward/backward regression analyses with ROI item response. In stepwise regression, at each step, one predictor is added to the regression, which most improves the model's fit to the data (the forward step). The overall model fit is penalized for each additional predictor, to avoid over-fitting. Predictors are then excluded if their contribution to predicting the outcome becomes non-significant after other predictors are included in the model (the backward step). The process is iterated until adding an additional predictor would not significantly improve the model fit. Each ROI's response to the items was the outcome (dependent) variable and the item features and condition were the predictor (independent) variables. Given the large number of predictors (20) entered into the regression, we used a fairly stringent entry and removal criteria (entry = .01, removal = .05; see Supplementary Table 2 for regression results performed with entry $p = .05$ and removal $p = .10$).

Results

Subject- and item-wise whole brain random effects analysis

We first asked whether brain regions that reliably respond to belief information across subjects also respond reliably across items. To that end, we performed whole brain subject- and item-wise analysis on the same data set. Fig. 1 depicts the brain regions significantly more active (FDR corrected, $p<.01$, $k>10$) for the BELIEF versus PHOTO stories across subjects, and across items, in the whole brain (see Table 2 for list of brain regions). All of the brain regions thought to comprise the ToM network (RTPJ, LTPJ, PC, DMPFC, MMPFC, RSTS, left superior temporal sulcus [LSTS], RTP) were reliably active in both subject- and item-wise analysis indicating that the ToM network also has a reliable response across items.

Brain regions active for subject-wise analysis and not item-wise analysis were the left temporal pole (LTP), right posterior middle frontal gyrus, calcarine sulcus, middle cingulate gyrus, left inferior frontal gyrus and orbito-frontal cortex, indicating that these regions are not reliably activated across false-belief items. That is, these brain regions (not considered part of the classic ToM network) may appear in the subject-wise analyses due to features of a small number of

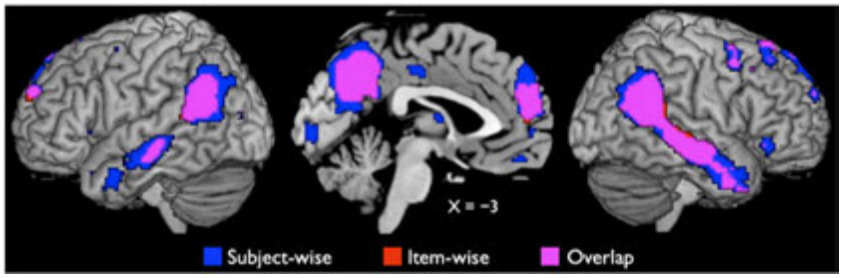

Fig. 1. Whole brain subject- and item-wise random effects analysis of BELIEF>PHOTO. Whole brain subject-wise (blue activations; N = 62 participants) and item-wise (red activations; N = 40 items; 20 BELIEF and 20 PHOTO items) random effects analysis of BELIEF>PHOTO (FDR corrected, $p<.01$) revealed overlap (magenta activations) in commonly observed ToM brain regions: RTPJ, LTPJ, PC, RSTS, LSTS, RTP, and MPFC. Activations are displayed on a canonical brain image.

items; they would not necessarily be activated for other sets of false-belief stories.

ROI analysis of the item-wise data

Whole brain item-wise random effects analysis demonstrated that the ToM network is reliably recruited across items. We sought converging evidence for this finding with ROI analyses. To that end, ROIs indentified in the BELIEF versus PHOTO contrast in the subject-wise analysis (RTPJ, LTPJ, PC, RpSTS, RTP, DMPFC, MMPFC; Fig. 2) for Group 1 were used to extract beta values for each item from Group 2's data. Average beta values for the BELIEF and PHOTO items were compared with *t*-tests (Fig. 2). In all of the ROIs, activity was significantly greater for the BELIEF versus PHOTO stories (RTPJ: $t(38) = 6.13$ $p<.001$; LTPJ: $t(38) = 4.89$ $p<.001$; PC: $t(38) = 5.99$ $p<.001$; RpSTS: $t(38) = 5.19$ $p<.001$; RTP: $t(38) = 3.02$ $p = .005$; DMPFC: $t(38) = 2.09$ $p = .043$; MMPFC: $t(38) = 2.39$ $p = .022$) providing converging evidence for the whole brain analyses, and replicating the results of previous subject-wise analyses of BELIEF versus PHOTO stories (e.g., Saxe and Kanwisher, 2003).

Table 2
Whole brain subject- and item-wise random effects results for BELIEF>PHOTO.

	x	y	z	Peak voxel *t*-value
Subject-wise				
Precuneus	−2	−54	38	14.43
Right temporo-parietal junction	52	−52	22	11.10
Right anterior superior temporal sulcus	58	−20	−14	9.53
Left temporo-parietal junction	−50	−58	20	9.36
Left anterior superior temporal sulcus	−60	−24	−8	7.57
Right temporal pole	50	6	−34	7.54
Middle medial prefrontal cortex	2	60	18	6.43
Right posterior superior frontal gyrus	10	34	62	5.85
Dorsal medial prefrontal cortex	0	56	26	5.70
Left temporal pole	−52	4	−28	5.27
Left superior frontal gyrus	−20	60	30	5.26
Right posterior middle frontal gyrus	46	8	48	4.83
Calcarine sulcus	2	−92	0	4.81
Middle cingulate gyrus	2	−22	42	4.74
Right inferior frontal gyrus	52	30	−6	4.63
Orbito-frontal cortex	2	48	−18	4.59
Item-wise				
Right posterior superior temporal sulcus	62	−32	0	7.55
Right temporo-parietal junction	56	−48	24	7.38
Right anterior superior temporal sulcus	58	−22	−10	7.22
Left temporo-parietal junction	−50	−50	24	7.14
Precuneus	8	−58	40	6.79
Middle medial prefrontal cortex	0	60	18	6.07
Left superior frontal gyrus	−18	60	28	5.93
Right posterior superior frontal gyrus	12	34	62	5.48
Dormal medial prefrontal cortex	2	58	30	5.07
Left anterior superior temporal sulcus	62	−24	−10	4.75
Right temporal pole	52	2	−36	4.53

Peak voxels for ROIs in Montreal Neurological Institute [MNI] coordinates (FDR corrected, $p<.01$, $k>10$).

Reliability of ROI response to items within conditions: item-wise correlation analysis

The previous analyses indicate that ToM brain regions generalize across subjects and items. A separate question is whether any of these brain regions – RTPJ, LTPJ, PC, RpSTS, RTP, DMPFC, MMPFC – demonstrate a systematically higher response to individual stories *within* conditions. If so, this would suggest that the region is responding to specific features of the item. To address this question, we performed an item-wise correlation analysis, correlating ROI response to the items across groups created by randomly splitting the 62 subjects into two groups 100 times. Mean Pearson *r* values across the 100 iterations along with 95% confidence intervals (CI) of the mean *r* value are reported in Table 3 (data from Group 1 and Group 2 are depicted in Fig. 2). The range of mean correlation values was .25 (MMPFC) to .82 (PC). Reliability estimates of these correlations (95% CI of the mean) ranged from ±.10 (PC) to ±.35 (MMPFC).

The within-condition, cross-group correlations demonstrate that most of the ROIs in the ToM network have reliable preferences for items within each condition. An interesting question is whether all the regions in the ToM network have the same item preferences, or whether different brain regions prefer different items. To address this question, we correlated item beta values from Group 2 across ROIs (defined from Group 1), separately for BELIEF and PHOTO items (Table 4). The regions mostly showed uncorrelated preferences among the BELIEF items; the only two significant correlations were between the RTPJ and each of the PC and RTP (PC and RTP response were not correlated with each other). For the PHOTO items, by contrast, many of the regions shared similar preferences: item-wise preferences for PHOTO items were correlated among the RTPJ, LTPJ, PC, RpSTS, and RTP, and between the DMPFC and MMPFC, for example (see Table 4 for the complete list).

In sum, most of the ToM regions had reliable preferences for specific items within the BELIEF and PHOTO conditions, and these preferences were not identical across regions, especially for stories about false beliefs. These results raise the important question of which features that vary between the individual items predict the item-specific responses in each of these regions.

Using item features to predict ROI item response

To test the predictive power of many features simultaneously, we performed stepwise forward/backward regressions. In the RTPJ and LTPJ, condition (RTPJ: standardized $\beta = .695$, $t(39) = 5.97$, $p<.001$; LTPJ: $\beta = .592$, $t(39) = 4.53$, $p<.001$) emerged as the only significant predictor of activity. The resulting model accounted for 47.0% (RTPJ) and 33.3% (LTPJ) of the variance in the response across items (adjusted for the number of predictors). In the PC, RpSTS and RTP,

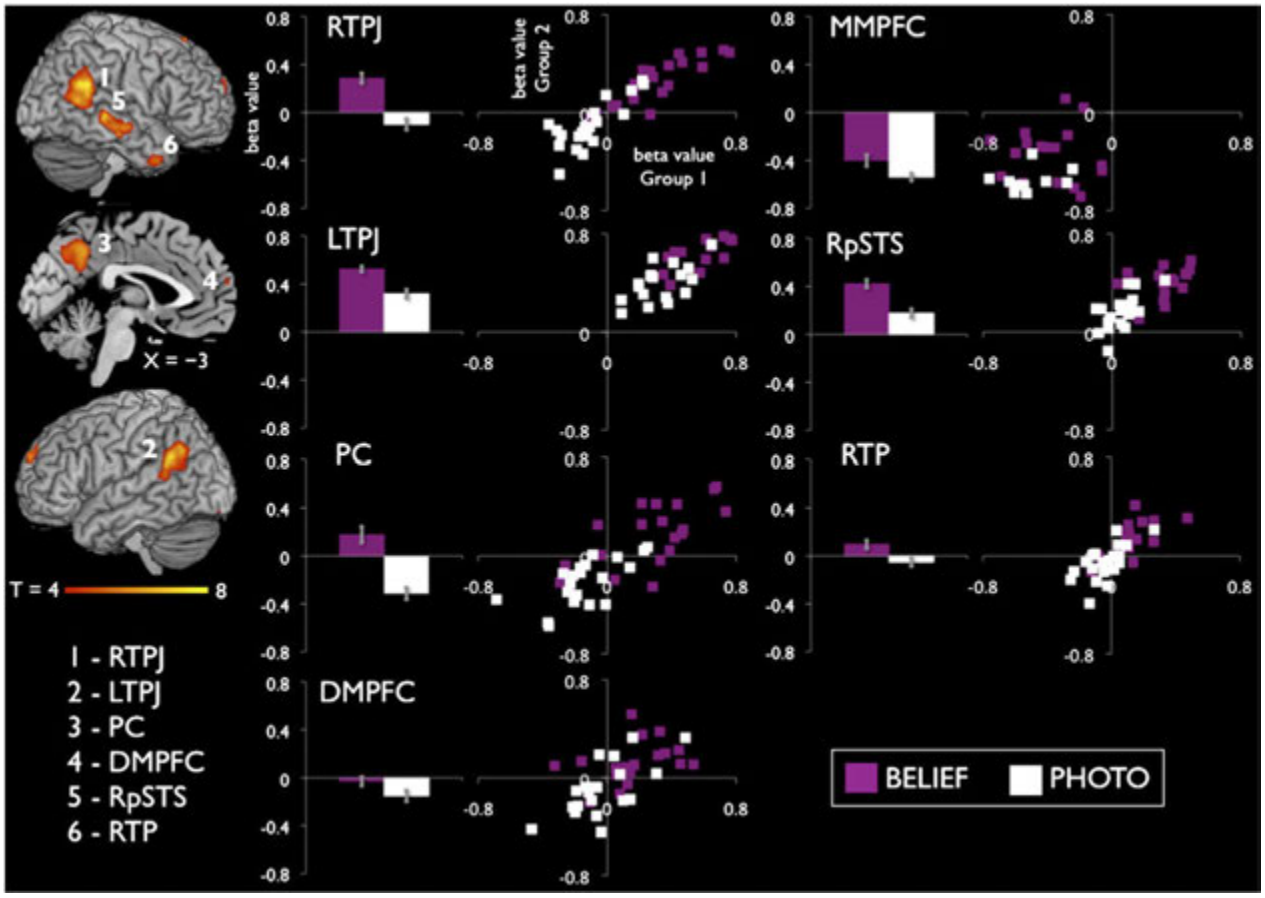

Fig. 2. Item-wise ROI analyses and item by group response correlations within each ROI. Whole brain subject-wise random effects analysis of Group 1 (n = 32) which was used to extract item beta values for Group 2 (n = 30) for BELIEF>PHOTO ($p<.0001$, uncorrected, $k>10$). Seven ROIs were defined from this analysis (RTPJ, LTPJ, PC, RpSTS, RTP, DMPFC and MMPFC [not visible here]). Activations are displayed on a canonical brain image. Corresponding ROI plots of beta values extracted from Group 2 using the ROIs defined in Group 1. Error bars represent standard error of the mean. To the right of the bar graphs are scatterplots of the beta values from Group 1 and Group 2 for each item (see item-wise correlation analysis results and Table 3 for mean correlation values and 95% confidence intervals). Purple data points depict BELIEF items; white data points depict PHOTO items.

condition was not a reliable predictor, once other features were included in the model. In the PC, only number of people ($\beta = .729$, $t(39) = 6.57$, $p<.001$) was included in the model, accounting for 51.9% of the variance. In the RpSTS, only the extent to which M-Turk participants reported that they considered the thoughts, beliefs and desires of someone in the story ($\beta = .680$, $t(39) = 5.71$, $p<.001$) reliably predicted BOLD signal, accounting for 44.8% of the variance. In the RTP, only number of words per story ($\beta = .525$, $t(39) = 3.80$, $p<.001$) reliably predicted BOLD signal, accounting for 25.6% of the variance. In the DMPFC and MMPFC, none of the factors reliably predicted BOLD signal. The other linguistic features were not included in the final model of the response in any of the ToM regions.

Discussion

Inferences about the cognitive function of brain regions based on fMRI data can be significantly strengthened by item analyses, to show that effects of specific items generalize to the entire item category. However, item analyses remain rare, and have not previously been used in social cognitive neuroscience. Our first objective was to determine whether activity in the ToM network generalizes from a commonly used set of false beliefs stories to the category of such stories about false beliefs. All of the brain regions thought to comprise the ToM network (i.e. RTPJ, LTPJ, PC, DMPFC, MMPFC, RSTS, LSTS and

Table 3
Item-wise correlation analysis: mean correlation (*r*-value) and 95% confidence interval of the mean.

	RTPJ	LTPJ	PC	RpSTS	RTP	DMPFC	MMPFC
Belief	.81 ± .12	.68 ± .19	.82 ± .10	.56 ± .21	.66 ± .18	.45 ± .26	.32 ± .31
Photo	.78 ± .12	.77 ± .16	.68 ± .22	.45 ± .32	.62 ± .21	.58 ± .26	.25 ± .35

Table 4
Item beta value correlations (*r*-value) across ROIs separately for BELIEF and PHOTO.

BELIEF	RTPJ	LTPJ	PC	RpSTS	RTP	DMPFC	MMPFC
RTPJ		.48[a]	.79*	.44	.65*	−.31	.50[a]
LTPJ			.12	.18	.33	.18	.06
PC				.27	.33	−.51[a]	.42
RpSTS					.50[a]	.37	.30
RTP						.00	.55
DMPFC							.00
PHOTO	**RTPJ**	**LTPJ**	**PC**	**RpSTS**	**RTP**	**DMPFC**	**MMPFC**
RTPJ		.84*	.78*	.66*	.67*	.48[a]	.27
LTPJ			.63*	.76*	.49[a]	.57[a]	.16
PC				.54[a]	.69*	.59*	.15
RpSTS					.52[a]	.58[a]	.31
RTP						.55[a]	.20
DMPFC							.62*

RTPJ = right temporo-parietal junction, LTPJ = left temporo-parietal junction, PC = precuneus, RpSTS = right posterior superior temporal sulcus, RTP = right temporal pole, DMPFC = dorsal medial prefrontal cortex, MMPFC = middle medial prefrontal cortex.

* Bonferonni corrected $p<.008$.

[a] These values were significant at an uncorrected level of $p<.05$.

RTP, e.g., Aichhorn et al., 2009; Mitchell, 2008; Saxe and Kanwisher, 2003; Saxe and Powell, 2006; Saxe and Wexler, 2005) showed significantly higher responses to BELIEF than PHOTO stories, in item-wise whole brain random effects and in ROI analyses. These results provide formal evidence that ToM network activity in response to these stimuli is not due to idiosyncratic features of a small number of these stimuli, but instead generalizes to the category of such stories about false beliefs. By contrast, several brain regions were revealed only in the subject-, but not item-wise analysis: left temporal pole (LTP), right posterior middle frontal gyrus, calcarine sulcus, middle cingulate gyrus, left inferior frontal gyrus and orbito-frontal cortex. The absence of these regions in the item analysis suggests that they would not reliably be activated by other ToM stimuli, and indeed, these regions are not reliably observed across other experiments testing ToM.

Inferences from these data must be qualified, however. The power of item analyses is limited by the sample of stimuli, and the method by which they were generated. Item analysis supports generalization to a whole population specifically when experimental items were randomly sampled from that population. This limit is analogous to limits on generalizing the results of standard subject-wise analyses: in general, those results should be generalized only to the population from which the subjects were sampled, usually college-aged, highly educated, middle-class/wealthy individuals (as in the current experiment; see Henrich et al., 2010). Given the nature of the items of interest here (verbal narratives), it is hard to quantify the population from which they constitute a random sample. Also, the current stories are not a representative sample of all stories about people's thoughts, because these stories did not contain many descriptions of other mental states like true beliefs, desires, or emotions. Stronger conclusions could be reached in future studies, using item analyses with a larger and more variable set of stimuli, and a better characterized method for generating a 'random sample' of such stimuli (e.g. by selecting random passages from published short stories).

The true scope of the response in regions of the ToM network is suggested by the wide range of prior studies that have reported activity in these regions, using verbal (e.g., Aichhorn et al., 2009; Fletcher et al., 1995; Gallagher et al., 2000; Saxe and Kanwisher, 2003) and non-verbal stimuli (e.g., Castelli et al., 2000; Gallagher et al., 2000; German et al., 2004; Walter et al., 2010), stimuli depicting true beliefs (Saxe et al., 2009; Saxe and Powell, 2006; Young et al., 2010a) and false beliefs (Fletcher et al., 1995; Gallagher et al., 2000; Saxe and Kanwisher, 2003; Vogeley et al., 2001), stimuli in English and non-English languages (Kobayashi et al., 2007), stimuli describing beliefs and preferences (Jenkins and Mitchell, 2010), and stimuli describing affective as opposed to purely epistemic states (Vollm et al., 2006; Walter et al., 2010). Taken together with the current findings, the response of the ToM network does indeed appear to generalize not just to stories about false beliefs, but to the entire category of mental states (or stimuli that lead to mentalizing).

The second purpose of the item analysis was to produce an estimate of the response in each brain region, to each specific stimulus. If a region has a reliable preference for specific items within a category, this preference may provide a clue about the region's function. We found that the RTPJ, LTPJ, PC, RpSTS, RTP and DMPFC had reliable preferences for specific items within conditions, across participants. Furthermore, item preferences were not highly correlated across brain regions, especially for BELIEF stories. Thus, these item preferences could reveal the distinct functions of individual regions within the ToM network.

We therefore attempted to predict the item-specific responses of each region, using twenty different features of the items. The features we coded included belief-related aspects of the story (the extent to which the story made participants consider somebody's thoughts or desires, or to which someone was being deceived), social features (number of people per story and the extent to which someone's social status, role or appearance was considered), causal content (including the extent to which the story made participants consider physical causal interactions) and text difficultly/comprehensibility (e.g. number of words, and various measures of syntactic complexity). The linguistic features were of particular interest, since previous authors have suggested that ToM depends distinctively on syntactic representations of embedded sentence complements (e.g. "Hank thinks *that his saxophone is in the closet*"; de Villiers and Pyers, 2002), and that linguistic features might account for the activity in ToM brain regions (Ferstl and von Cramon, 2002; Ferstl et al., 2008).

We found that simple models, including only one or two factors, best accounted for the cross-item variability in each region's response. The best model of the RTPJ and LTPJ, in particular, included only the single factor of condition (BELIEF versus PHOTO stories), which accounted for almost half of the overall variance in the RTPJ and a third of the variance in the LTPJ. In the PC, the best predictor was the number of people in the story: there was greater activity in the PC for stories that mentioned more people. In the RpSTS the best predictor was the independent ratings of another group of participants, of how much each item made them consider the "thoughts, beliefs, desires or emotions" of the protagonist.

These differences between regions are intriguing, but should be interpreted with caution: in the current relatively small set of stimuli, the number of people in the story and these M-Turk ratings, were both highly correlated with condition. When two predictors are highly correlated, a step-wise regression assigns all of the associated variance in the output variable to one of the two predictor variables: whichever one initially accounts for the most variance. Thus, small differences in the predictive power of two correlated input variables can lead to different-looking models. Future research using regressions over item analyses should use larger sets of stimuli, in which the predictor variables can be better disentangled.

Perhaps the most intriguing result of the regression analyses was the factors that were not able to predict the ToM regions' preferences for items. Linguistic features of the stimuli, and other features we coded, like visualizibility, textual difficulty, and causal content, apparently do not explain away the neural response in any of the ToM regions. It remains an open challenge to account for the within-condition preferences in these regions; we welcome suggestions (the Supplementary Materials includes Supplementary Table 1, showing the response of each brain region to each item).

One practical implication of these results is that it should allow for easier identification of ToM brain regions in new subjects by using the BELIEF stories that produce the greatest activation in the ToM network, and the PHOTO stories which produce the least amount of activation in the ToM network. We confirmed this prediction and report, in the Supplementary Materials (see "ToM Superlocalizer") the results of two naïve subjects tested with an extremely more efficient version of the false belief task: just 5 stories from each condition. This new version produced reliable activation in the ToM network in just 5 min of scanning. Thus, item analyses afford the additional advantage of providing data that can inform the design of more efficient scanner paradigms.

Conclusion

In summary, item analysis provides two important features beyond those of conventional fMRI analyses: They allow the generalization of effects from items employed in a specific experiment to entire categories of items and can provide insight into the more subtle relationships between functional activity and cognitive processes that would normally be obscured by analysis at the category level. In consideration of these ideas, we employed item analysis on a false belief task for ToM brain regions. Our findings demonstrate that (1) activity in the ToM network generalizes to the category of false-belief stories, and (2) some regions within the network have a reliable

preference for specific items within conditions. However, the item-specific responses on these regions were not predicted by plausible confounding factors like the visual vividness or linguistic complexity of the items. Thus, item analyses may reveal a previously unknown division of labor in the ToM network for processing specific stories about beliefs. These hypotheses, suggested by exploratory item analyses, must be confirmed by direct experimental manipulation, and eventually, to get beyond the correlational analyses of fMRI, using tools like transcranial magnetic stimulation, that allow researchers to test a brain region's causal role (Young et al., 2010b).

Acknowledgments

The authors would like to thank Jacqueline Pigeon, Adrianna Jenkins, Hyowon Gweon and Emile Bruneau for help with data collection, Steven T. Piantadosi for advice on data analysis, and Elizabeth Redcay for comments on the manuscript. This work was supported by a John Merck Scholars Grant, the Ellison Medical Foundation and the Office of Naval Research. Data were collected at the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research, MIT.

Appendix A. Supplementary data

Supplementary data to this article can be found online at doi:10.1016/j.neuroimage.2010.12.040.

References

Aichhorn, M., Perner, J., Weiss, B., Kronbichler, M., Staffen, W., Ladurner, G., 2009. Temporo-parietal junction activity in theory-of-mind tasks: falseness, beliefs, or attention. Journal of Cognitive Neuroscience 21 (6), 1179–1192.

Bedny, M., Thompson-Schill, S.L., 2006. Neuroanatomically separable effects of imageability and grammatical class during single-word comprehension. Brain and Language 98 (2), 127–139.

Bedny, M., Aguirre, G.K., Thompson-Schill, S.L., 2007. Item analysis in functional magnetic resonance imaging. Neuroimage 35 (3), 1093–1102.

Castelli, F., Happe, F., Frith, U., Frith, C., 2000. Movement and mind: a functional imaging study of perception and interpretation of complex intentional movement patterns. Neuroimage 12 (3), 314–325.

Clark, H.H., 1973. The language-as-fixed-effect fallacy: a critique of language statistics in psychological research. Journal of Verbal Learning and Verbal Behavior 12, 335–359.

Coltheart, M., 1981. The MRC psycholinguistic database. Quarterly Journal of Experimental Psychology A 33 (4), 497–505.

Davis, M.H., Meunier, M.F., Marslen-Wilson, W.D., 2004. Neural responses to morphological, syntactic, and semantic properties of single words: an fmri study. Brain and Language 89 (3), 439–449.

de Villiers, J., Pyers, J., 2002. Complements to cognition: a longitudinal study of the relationship between complex syntax and false-belief-understanding. Cognitive Development 17 (1), 1037–1060.

Fellbaum, C., 1998. Wordnet: An Electronic Lexical Database. MIT press, Cambridge, MA.

Ferstl, E.C., von Cramon, D.Y., 2002. What does the frontomedian cortex contribute to language processing: coherence or theory of mind? Neuroimage 17 (3), 1599–1612.

Ferstl, E.C., Neumann, J., Bogler, C., von Cramon, D.Y., 2008. The extended language network: a meta-analysis of neuroimaging studies on text comprehension. Human Brain Mapping 29 (5), 581–593.

Fletcher, P.C., Happé, F., Frith, U., Baker, S.C., Dolan, R.J., Frackowiak, R.S., et al., 1995. Other minds in the brain: a functional imaging study of "theory of mind" in story comprehension. Cognition 57 (2), 109–128.

Friston, K.J., Holmes, A.P., Worsley, K.J., 1999. How many subjects constitute a study? Neuroimage 10 (1), 1–5.

Gallagher, H.L., Happe, F., Brunswick, N., Fletcher, P.C., Frith, U., Frith, C.D., 2000. Reading the mind in cartoons and stories: an fMRI study of 'theory of mind' in verbal and nonverbal tasks. Neuropsychologia 38 (1), 11–21.

German, T.P., Niehaus, J.L., Roarty, M.P., Giesbrecht, B., Miller, M.B., 2004. Neural correlates of detecting pretense: automatic engagement of the intentional stance under covert conditions. Journal of Cognitive Neuroscience 16 (10), 1805–1817.

Graesser, A.C., Millis, K.K., Zwaan, R.A., 1997. Discourse comprehension. Annual Review of Psychology 48, 163–189.

Graesser, A.C., McNamara, D.S., Louwerse, M.M., Cai, Z., 2004. Coh-Metrix: analysis of text on cohesion and language. Behavior Research Methods, Instruments, & Computers 36 (2), 193.

Henrich, J., Heine, S.J., Norenzayan, A., Heine, S.J., Norenzayan, A., 2010. The weirdest people in the world? The Behavioral and Brain Sciences 33 (2–3), 61–83 discussion 83–135.

Holmes, A.P., Friston, K.J., 1998. Generalizability, random effects, and population inference. Neuroimage 7, S754.

Jenkins, A.C., Mitchell, J.P., 2010. Mentalizing under uncertainty: dissociated neural responses to ambiguous and unambiguous mental state inferences. Cerebral Cortex 20 (2), 404–410.

Kliemann, D., Young, L., Scholz, J., Saxe, R., 2008. The influence of prior record on moral judgment. Neuropsychologia 46 (12), 2949–2957.

Kobayashi, C., Glover, G.H., Temple, E., 2007. Cultural and linguistic effects on neural bases of 'Theory of Mind' in American and Japanese children. Brain Research 1164, 95–107.

McNamara, D.S., Louwerse, M.M., Graesser, A.C., 2002. Coh-Metrix: Automated Cohesion and Coherence Scores to Predict Text Readability and Facilitate Comprehension. Grant Proposal. . Available At: http://Cohmetrix.Memphis.Edu/Cohmetrixpr/Publications.html.

McNamara, D.S., Ozuru, Y., Graesser, A.C., Louwerse, M., 2006. Validating Coh-Metrix. In: Sun, R., Miyake, N. (Eds.), Proceedings of the 28th Annual Conference of the Cognitive Science Scociety. Eribaum, Mahwah, NJ, p. 573.

Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.J., 1990. Introduction to wordnet: an on-line lexical database*. International Journal of Lexicography 3 (4), 235.

Mitchell, J.P., 2008. Activity in right temporo-parietal junction is not selective for theory-of mind. Cerebral Cortex 18 (2), 262–271.

Saxe, R., Kanwisher, N., 2003. People thinking about thinking people. The role of the temporo-parietal junction in "theory of mind". Neuroimage 19 (4), 1835–1842.

Saxe, R., Powell, L.J., 2006. It's the thought that counts: specific brain regions for one component of theory of mind. Psychological Science 17 (8), 692–699.

Saxe, R., Wexler, A., 2005. Making sense of another mind: the role of the right temporo parietal junction. Neuropsychologia 43 (10), 1391–1399.

Saxe, R., Schulz, S., Jiang, Y., 2006. Reading minds versus following rules: dissociating theory of mind and executive control in the brain. Social Neuroscience 1 (3/4), 284–298.

Saxe, R.R., Whitfield-Gabrieli, S., Scholz, J., Pelphrey, K.A., 2009. Brain regions for perceiving and reasoning about other people in school-aged children. Child Development 80 (4), 1197–1209.

Scholz, J., Triantafyllou, C., Whitfield-Gabrieli, S., Brown, E.N., Saxe, R., 2009. Distinct regions of right temporo-parietal junction are selective for theory of mind and exogenous attention. PLoS ONE 4 (3), e4869.

Shapiro, K.A., Moo, L.R., Caramazza, A., 2006. Cortical signatures of noun and verb production. Proceedings of the National Academy of Sciences of the United States of America 103 (5), 1644.

Tyler, L.K., Russell, R., Fadili, J., Moss, H.E., 2001. The neural representation of nouns and verbs: PET studies. Brain 124 (8), 1619.

Tyler, L.K., Bright, P., Fletcher, P., Stamatakis, E.A., 2004. Neural processing of nouns and verbs: the role of inflectional morphology. Neuropsychologia 42 (4), 512–523.

Vogeley, K., Bussfeld, P., Newen, A., Herrmann, S., Happe, F., Falkai, P., et al., 2001. Mind reading: neural mechanisms of theory of mind and self-perspective. Neuroimage 14 (1 Pt 1), 170–181.

Vollm, B.A., Taylor, A.N., Richardson, P., Corcoran, R., Stirling, J., McKie, S., et al., 2006. Neuronal correlates of theory of mind and empathy: a functional magnetic resonance imaging study in a nonverbal task. Neuroimage 29 (1), 90–98.

Walter, H., Schnell, K., Erk, S., Arnold, C., Kirsch, P., Esslinger, C., et al., 2010. Effects of a genome-wide supported psychosis risk variant on neural activation during a theory-of-mind task. Molecular Psychiatry. doi:10.1038/mp.2010.18.

Wimmer, H., Perner, J., 1983. Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. Cognition 13 (1), 103–128.

Young, L., Cushman, F., Hauser, M., Saxe, R., 2007. The neural basis of the interaction between theory of mind and moral judgment. Proceedings of the National Academy of Sciences of the United States of America 104 (20), 8235–8240.

Young, L., Camprodon, J., Hauser, M., Pascual-Leone, A., Saxe, R., 2010a. Disruption of the right temporo-parietal junction with transcranail magnetic stimulation reduces the role of beliefs in moral judgment. PNAS 107 (15), 6753–6758.

Young, L., Dodell-Feder, D., Saxe, R., 2010b. What gets the attention of the temporo-parietal junction? An fMRI investigation of attention and theory of mind. Neuropsychologia 48 (9), 2658–2664.