NeuroImage 99 (2014) 301–311

Contents lists available at ScienceDirect

NeuroImage

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

Validating the Why/How contrast for functional MRI studies of Theory of Mind

Robert P. Spunt *, Ralph Adolphs

California Institute of Technology, USA

ARTICLE INFO

Article history:
Accepted 11 May 2014
Available online 17 May 2014

Keywords:
Social cognition
Theory of Mind
Action understanding
Attribution
False belief
Mentalizing
Localizer
fMRI

ABSTRACT

The ability to impute mental states to others, or Theory of Mind (ToM), has been the subject of hundreds of neuroimaging studies. Although reviews and meta-analyses of these studies have concluded that ToM recruits a coherent brain network, mounting evidence suggests that this network is an abstraction based on pooling data from numerous studies, most of which use different behavioral tasks to investigate ToM. Problematically, this means that no single behavioral task can be used to reliably measure ToM Network function as currently conceived. To make ToM Network function scientifically tractable, we need standardized tasks capable of reliably measuring specific aspects of its functioning. Here, our goal is to validate the Why/How Task for this purpose. Several prior studies have found that when compared to answering how-questions about another person's behavior, answering why-questions about that same behavior activates a network that is anatomically consistent with meta-analytic definitions of the ToM Network. In the version of the Why/How Task presented here, participants answer yes/no Why (e.g., Is the person helping someone?) and How (e.g., Is the person lifting something?) questions about pretested photographs of naturalistic human behaviors. Across three fMRI studies, we show that the task elicits reliable performance measurements and modulates a left-lateralized network that is consistently localized across studies. While this network is convergent with meta-analyses of ToM studies, it is largely distinct from the network identified by the widely used False-Belief Localizer, the most common ToM task. Our new task is publicly available, and can be used as an efficient functional localizer to provide reliable identification of single-subject responses in most regions of the network. Our results validate the Why/How Task, both as a standardized protocol capable of producing maximally comparable data across studies, and as a flexible foundation for programmatic research on the neurobiological foundations of a basic manifestation of human ToM.

Introduction

Human social cognition makes constant use of a Theory of Mind (ToM), an ability that encompasses conceiving of behavior as driven by unobservable mental states, and appreciating that the mental states of others may differ from one's own (Apperly, 2012; Dennett, 1989; Gopnik and Wellman, 1992; Leslie et al., 2004; Premack and Woodruff, 1978; Wellman et al., 2001). This ability is likely necessary for survival in a complex social world, is thought to be severely impaired in numerous psychopathologies (e.g., autism), and may be unique to humans in degree if not in kind (Kennedy and Adolphs, 2012). Moreover, hundreds of neuroimaging studies have already examined the correlates of ToM in the human brain. Collectively, these studies suggest that the use of ToM is reliably associated with a set of regions now commonly known as the ToM Network. The present studies were motivated by two interrelated problems facing neuroimaging studies on ToM.

Problems with prior research

The first problem is a significant lack of standardized tasks for investigating specific uses of ToM. As noted by numerous meta-analyses, the hundreds of neuroimaging studies of ToM feature enormous variability in the manner by which ToM is operationally defined (Carrington and Bailey, 2009; Denny et al., 2012; Lieberman, 2010; Mar, 2011; Van Overwalle and Baetens, 2009). This is not surprising: The broad ability called ToM spans the flexible use of a wide range of mental representations (e.g., belief vs. desire) to understand a diverse array of stimuli (e.g., verbal vs. nonverbal) in the service of a variety of goals (e.g., deception vs. empathic understanding). For example, many neuroimaging studies have investigated ToM through the lens of the false-belief localizer (Saxe et al., 2004), which requires participants to comprehend verbal narratives and make a prediction about a character's future behavior based on a representation of their belief. Other neuroimaging studies have investigated ToM through a different lens, using simple nonverbal geometric animations (Heider and Simmel, 1944) to evoke inferences about motive and intent (e.g., Schultz et al., 2003). Perhaps unsurprisingly, the one empirical study to formally compare these two tasks concluded that they modulate largely distinct neural systems

* Corresponding author.
E-mail addresses: spunt@caltech.edu (R.P. Spunt), radolphs@caltech.edu (R. Adolphs).

http://dx.doi.org/10.1016/j.neuroimage.2014.05.023
1053-8119/

(Gobbini et al., 2007). This is not itself problematic, since it is natural to expect that a cognitive construct as broad and complex as ToM would be decomposable into multiple distinct processes, each of which would require a distinct methodology to investigate scientifically. Importantly, programmatic scientific research necessitates the existence of standardized protocols that are generally accepted by the research community (or in the least multiple research groups) as a valid, reliable, and distinctive operational definition of a theoretical construct. In the absence of such protocols, findings of different studies are often extremely difficult to compare, even if those studies claim to be investigating the same theoretical construct. Ultimately, this impedes scientific progress by preventing cumulative research.

An adverse consequence of a lack of standardization is illustrated by the second problem this study aims to help address: anatomical delineations of the ToM Network remain imprecise. The regions included in the definition of the network vary across different literature reviews, and even large meta-analyses that include hundreds of studies fail to converge on a precise definition (Carrington and Bailey, 2009; Denny et al., 2012; Lieberman, 2010; Mar, 2011; Van Overwalle and Baetens, 2009). When convergence does occur, it is often explained by the fact that the labels used to define the regions of the network are themselves anatomically imprecise. For instance, the labels used to define the two regions most reliably associated with ToM – the dorsomedial prefrontal cortex (dmPFC) and the temporoparietal junction (TPJ) – can both be used to refer to large areas of cortex that are known to exhibit both structural and functional heterogeneities. Because of this, the same label is often used to report areas of activation that are clearly different; this, in turn, blurs out potentially meaningful distinctions at both the neural and cognitive levels of analysis. In sum, the search for a single network in the human brain subserving ToM is probably misguided.

The value of standardization

Methodological variability must be balanced with methodological standardization, because only with the latter is programmatic research possible. This can be illustrated with reference to the single currently standardized protocol for investigating the neural bases of ToM. This is the False-Belief Localizer (often referred to, in fact, as the Theory-of-Mind Localizer) developed by Rebecca Saxe and colleagues (Saxe and Kanwisher, 2003; Saxe and Powell, 2006). The task uses brief verbal narratives to manipulate the demand to represent another person's false belief about reality. Two types of verbal narratives are contrasted to isolate the neural bases of representing false-belief: Stories in which a character comes to have a false belief about the world, and stories in which a physical record of the world (e.g.,. a photograph, map, or painting) becomes outdated or misleading. Compared to False-Photograph stories, False-Belief stories reliably evoke an increased response in a set of brain regions that are anatomically similar to the putative ToM Network. In fact, these regions can be reliably localized in individual participants using an empirically validated version of the protocol that takes less than 10 min to run and is publicly available (Dodell-feder et al., 2011). Given the consistency of the basic operational definition across studies, it is now reasonable to aggregate data across numerous studies in order to establish reliability and produce an empirical distribution against which new data can be evaluated. By using such an empirical distribution, Dufour et al. (2013) recently demonstrated that a small sample of adults with an autism spectrum disorder (ASD) showed a response to the Belief > Photo contrast that was within normal ranges. Given the programmatic nature of research using the Belief > Photo contrast to probe ToM Network function, this finding has clear implications for past and future research using a version of the Belief > Photo contrast. Critically, this is not because researchers have and will likely continue to share an interest in ToM; rather, what is critical is that researchers have and will likely continue to share an operational definition of ToM and to use consistently a particular localizer task. Without such a common ground, the findings from different studies are often difficult and sometimes impossible to compare.

Motivation for the present studies

Without standardized protocols for generating a body of data that is comparable across studies, programmatic research is virtually impossible. Unfortunately, the False-Belief Localizer is at present the only standardized protocol for manipulating a use of ToM and probing its underlying brain systems. Here, we follow the example it sets by introducing and validating a standardized contrast for investigating the human ability to explain behavior (Heider, 1958; Jones and Davis, 1965; Kelley, 1973). In prior work, we have investigated the neural correlates of this ability by instructing subjects to freely think of answers to why-questions about observable human behaviors. In a second condition, participants observe the same behaviors, but instead think of answers to how-questions about those behaviors. Across several studies examining variants of this attentional manipulation, we have observed that the Why > How contrast evokes a response in a set of brain regions that, like the Photo > Belief contrast, shows a high degree of qualitative correspondence with meta-analytically and review-based definitions of the ToM Network (Spunt and Lieberman, 2012a,b, 2013; Spunt et al., 2010, 2011).

The present study was motivated to validate and standardize a novel implementation of this contrast that significantly improves upon past research. In light of the problems identified above, our central aim was not to make a theoretical contribution, but a methodological one. There is no poverty of theory about what ToM entails, but there remains a significant poverty of validated methods for manipulating ToM in the context of a neuroimaging experiment. In Study 1, we introduce the method for achieving the Why/How contrast and present its behavioral and neural effects. In Study 2, we evaluate the test–retest reliability of the Why/How contrast in the same participants, and formally compare it to the Belief/Photo contrast obtained in the commonly used False-Belief Localizer in order to establish its discriminant validity. In Study 3, we introduce an efficient version of the new Why/How contrast and make this publicly available for use in neuroimaging research on ToM.

Study 1

Materials and methods

Participants

Participants were twenty-nine right-handed adults (19 males, 10 females; mean age = 27.10, age range = 19–38), all native English-speaking citizens of the United States. Each participant was neurologically and psychiatrically healthy, had normal or corrected-to-normal vision, spoke English fluently, had IQ in the normal range (as assessed using the Wechsler Abbreviated Scales of Intelligence), and was not pregnant or taking any psychotropic medications. Each participant provided written informed consent according to a protocol approved by the Institutional Review Board of the California Institute of Technology, and received financial compensation for participating.

Yes/No Why/How Task

The version of the Why/How contrast (Fig. 1) introduced here builds on the first author's previous work investigating the human brain regions associated with answering *Why* and *How* questions about human behavior (Spunt and Lieberman, 2012a,b, 2013; Spunt et al., 2010, 2011). Participants in these prior studies spontaneously and silently generated their own responses to these questions. Although this elicitation method features high ecological validity, it comes at a cost of experimental control and performance measurement. To address this limitation, we designed a version of the task that manipulates attention to “why” versus “how” by having participants answer pre-tested yes/no questions about naturalistic human behaviors shown in photographs. This provides a behavioral measure of both accuracy and

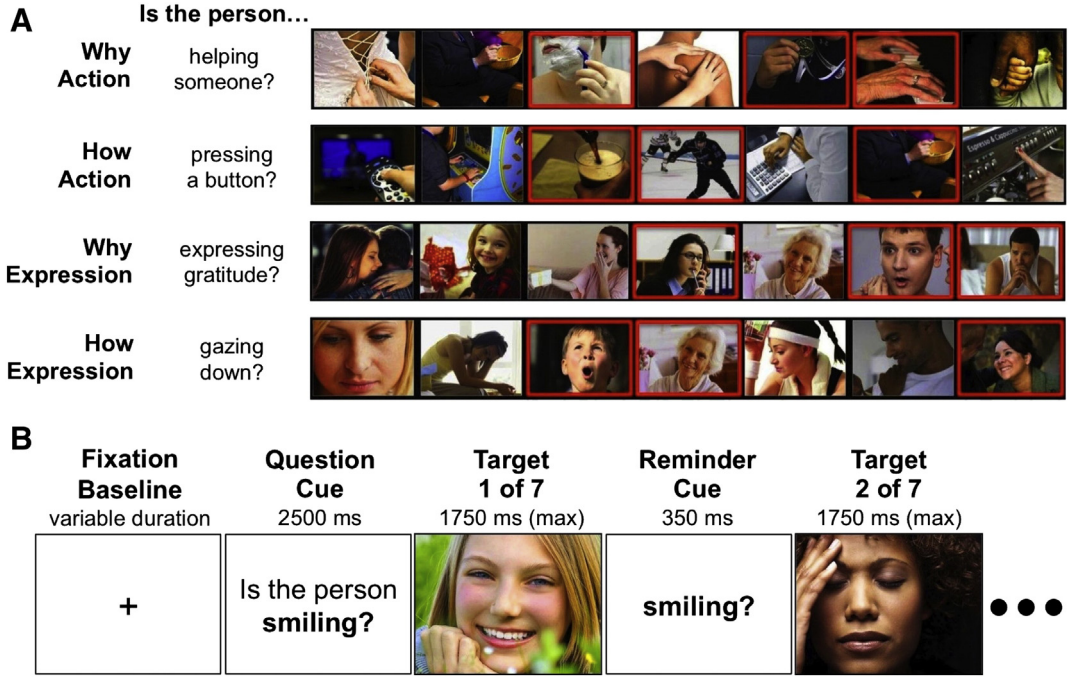

Fig. 1. Design of the Yes/No Why/How Task. (A) Examples of four blocks created by pairing either a question about motive (why) or implementation (how) with a set of photographs featuring either intentional actions or emotional expressions. Independently acquired normative data is used to ensure that every photo selected has an unambiguous (i.e., consensus) response. In the example blocks shown, the photographs outlined in red elicited a consensus response of 'no', while the remaining photographs elicited a consensus response of 'yes'. (B) Schematic showing the task timing. Each block begins with question presentation, and is followed by a set of photographs paired with that question. Between each photograph is a brief reminder of the question for that block. For each photograph, participants have 1750 ms to respond. If they fail to respond by that time, the task advances. Responding before the end of the 1750 ms ends the trial and advances to the next trial. Hence, block durations were contingent on response times. However, total task duration was not, as block onsets were fixed. As described in the main text, the versions of the Yes/No Why/How Task used in Studies 2 and 3 featured only trivial differences to what is presented here, which corresponds to the version used in Study 1.

response time, which can be used to validate that participants are in fact performing the task, as well as to explore individual differences and further associations of behavioral performance variability with brain activation. As in the original Why/How task, each photograph appears twice, once as the object of a question designed to focus attention on why it is being performed, and once as the object of a question designed to focus attention on how it is being performed.

The final set of photographs featured 42 photographs of familiar actions of the hand, and 42 photographs of familiar facial expressions. Table 1 displays the 24 questions featured in the study, broken down by condition (Why vs. How) and behavior category (Hand Actions vs. Facial Expressions). Each question was paired with 4 photographs designed to elicit the response 'yes', and 3 photographs designed to elicit the response 'no'. These pairings were selected based on the responses of an independent sample of respondents recruited through Amazon.com's web service Mechanical Turk. Each pairing was evaluated by at least 25 native English speaking U.S. citizens. We selected question–photo pairs with answers that elicited a consensus of at least 80.00% across participants. The average consensus of the final stimulus was 93.66% (SD = 6.37%) and did not differ significantly across the experimental manipulation of Why versus How.

During MRI scanning, items were presented to participants in blocks of 7 corresponding to each of the 24 questions (Fig. 1). The order of question-blocks was optimized to maximize the efficiency of estimating the Why > How contrast. This was achieved by generating the design matrices for one million pseudo-randomly generated orders, and for each calculating the efficiency of estimating the contrast of the regressors corresponding to Why and How question blocks. The two most efficient orders were retained, and one was randomly assigned to each participant. Prior to performing the Why/How localizer, participants were told they would be performing a "Photograph Judgment Test" in which they would answer yes/no questions about photographs of people. They were then shown two example trials and were invited to ask the experimenter questions if they did not fully understand the task. Finally, they were told that they would have a limited amount of time to respond to each photograph, and that if they were not sure about any answer, they should make their best guess. Total runtime of the task was 7 min, 5 s (Fig. 1 provides details for the timing of trials).

Stimulus presentation and response recording

In all three studies, stimulus presentation and response recording was achieved using the Psychophysics Toolbox (version 3.0.9; Brainard, 1997)

Table 1
The questions used in the Yes/No Why/How Task to manipulate and measure attention to "why" versus "how" for actions and expressions. All questions began with the string "Is the person ". The questions used in Study 3 are marked with an asterisk.

Why		How	
Intentional actions	Emotional expressions	Intentional actions	Emotional expressions
Competing against others?*	Admiring someone?*	Holding a ball?	Gazing down?
Concerned with their health?*	Being affectionate?	Lifting something?*	Looking at the camera?*
Having fun?	Expressing gratitude?	Pressing a button?*	Looking to their side?*
Helping someone?*	Expressing self-doubt?*	Reaching for something?*	Opening their mouth?*
Protecting themselves?*	In an argument?*	Using a writing utensil?	Showing their teeth?
Sharing knowledge?	Proud of themselves?*	Using both hands?*	Smiling?*

operating in MATLAB (version 2012a; MathWorks Inc., Natick, MA, USA). An LCD projector showed stimuli on a rear-projection screen. Participants made their responses using their right hand index and middle fingers on a button box.

Image acquisition

All imaging data were acquired at the Caltech Brain Imaging Center using a Siemens Trio 3.0 Tesla MRI Scanner outfitted with a 32 channel phased-array headcoil. We acquired 170 T2*-weighted echoplanar image volumes (EPIs; slice thickness = 3 mm, 47 slices, TR = 2500 ms, TE = 30 ms, flip angle = 85°, matrix = 64 × 64, FOV = 192 mm). Moreover, we also acquired a high-resolution anatomical T1-weighted image (1 mm isotropic) and field maps for each participant.

Image analysis

Functional data were analyzed using a combination of custom code and the MATLAB-based software package Statistical Parametric Mapping (SPM8, Wellcome Department of Cognitive Neurology, London, UK). Prior to statistical analysis, the first two EPI volumes from each run were discarded to account for T1 equilibration, and the remaining volumes were subjected to the following preprocessing steps: (1) each EPI volume was realigned to the first EPI volume of the run and simultaneously unwarped based on the fieldmap volumes; (2) the T1 structural volume was co-registered to the mean EPI; (3) the group-wise DARTEL registration method included in SPM8 (Ashburner, 2007) was used to normalize the T1 structural volume to a common group-specific space (with subsequent affine registration to MNI space); and (4) normalization of all EPI volumes to MNI space using the deformation flow fields generated in the previous step, which simultaneously re-sampled volumes (2 mm isotropic) and applied spatial smoothing (Gaussian kernel of 6 × 6 × 6 mm, full width at half maximum).

Single-subject effects were estimated using a General Linear Model. The hemodynamic response was modeled using the canonical (double-gamma) response function, and the predicted and actual signals were high-pass filtered at 0.01 Hz. As covariates of no interest, all models included the 6 motion parameter estimates from image realignment, and regressors indicating timepoints where in-brain global signal change (GSC) exceeded 2.5 SDs of the mean GSC or where estimated motion exceeds 0.5 mm of translation or 0.5 degrees of rotation. Finally, all models were estimated using the robust weighted least-squares algorithm implemented in the SPM8 RobustWLS toolbox (Diedrichsen and Shadmehr, 2005).

Each single-subject model included effects for the two conditions of interest: *Why* and *How*. Conditions were modeled as variable epochs (Grinband et al., 2008), with each epoch spanning onset of the first photograph of each block to the offset of the final photograph. In addition to the covariates of no interest described above, three additional parametric regressors were included. The first modeled variation in the type of behavior (action vs. expression) shown in the photographs across all blocks (a variable of no interest for the present study). The second modeled variation in the total accuracy of the responses within each block and ensures that the Why/How contrast is not confounded with performance accuracy. The third modeled the variation in the total duration of each block (effectively modeling any RT differences, since it was self-paced) and ensures that the Why/How contrast is not confounded with time on task. As we describe below, we include additional analyses in the Supplementary Materials that confirm that performance-related variability does not provide a sufficient explanation of the effects observed in the Why/How contrast.

To investigate the group-level effects, a single image for each participant representing the contrast of the Why and How conditions was entered into a second-level one-sample *t*-test. The resulting *t*-statistic image was corrected for multiple comparisons using cluster-level family-wise error (FWE) rate of .05 with a cluster-forming threshold of $p < .001$. In Table 2, we report only those peaks that survive a voxel-level FWE rate of .05.

Table 2
Group-level results of the Why/How contrast from Study 1 ($N = 29$). All peaks survive a whole-brain search thresholded at a voxel-wise family-wise error rate of .05 and a cluster extent (k) of at least 10 voxels. PFC = prefrontal cortex; IFG = inferior frontal gyrus; OFC = orbitofrontal cortex; STS = superior temporal sulcus; x, y, and z = Montreal Neurological Institute (MNI) coordinates in the left–right, anterior–posterior, and inferior–superior dimensions, respectively.

Contrast name				MNI coordinates		
Region name	L/R	*k*	t-value	x	y	z
Why > How						
Dorsomedial PFC	L	1112	11.863	−6	58	18
	L	–	10.762	−6	56	40
	L	–	9.787	−6	24	64
	R	23	7.953	8	52	40
Ventormedial PFC	L	345	10.269	−2	44	−18
Lateral OFC	L	144	8.604	−46	24	−12
Temporoparietal junction	L	103	8.785	−48	−66	28
Posterior cingulate cortex	L	133	8.179	−2	−48	30
Temporal pole	R	32	7.799	46	14	−34
Anterior STS	L	155	11.498	−54	−8	−18
	R	11	7.371	60	−8	−22
	R	16	7.096	54	−2	−22
Posterior STS	L	36	8.042	−58	−36	0
Cerebellum (posterior lobe)	R	166	11.083	34	−80	−34
	L	13	7.143	−30	−82	−32
How > Why						
Intraparietal sulcus	L	178	10.987	−42	−42	40
	L	21	7.993	−46	−42	58
	R	25	7.151	40	−42	48
Supramarginal gyrus	R	238	8.753	56	−32	50
	R	31	7.801	50	−42	40
	R	16	7.695	60	−24	24
	L	18	6.922	−56	−36	40
Precuneus (dorsal)	L	81	10.150	−8	−62	58
	R	23	7.699	10	−68	52
Superior parietal lobule	L	19	7.497	−26	−62	58
Lateral occipital cortex	L	20	7.048	−24	−70	32

To visualize the consistency of the Why > How contrast with the same contrast from our prior work, we used data from two published studies that used an open response protocol (instead of the yes/no response of the present study) to achieve the Why > How contrast for intentional hand actions (Spunt and Lieberman, 2012a) and emotional facial expressions (Spunt and Lieberman, 2012b). We computed the minimum statistic image from the group-level t-statistic images for the Why > How comparison in each study.

Results

Performance

For the Why/How Localizer, participants were significantly more accurate in their responses when answering How (M = 96.47%, SD = 2.73%) compared to Why (M = 93.39%, SD = 3.88%) questions, $t(28) = 3.671$, $p = .001$, 95% CI [1.361, 4.797]. In addition, participants were faster when answering How (M = 794 ms, SD = 112 ms) compared to Why (M = 909 ms, SD = 122 ms) questions, $t(28) = 12.366$, $p < .001$, 95% CI [96, 135]. Remarkably, all participants demonstrated this RT effect, responding faster to How compared to Why questions.

These data demonstrate that the Why/How contrast is reliably associated with two performance-related effects: Compared to How questions, Why questions elicit lower response accuracy and longer response times (RT). Importantly, we estimated the Why/How contrast using models that simultaneously modeled variance explained by accuracy and latency. In addition to incorporating RT and accuracy into our regression model in the main analyses presented below, we further confirmed that performance-related variability cannot explain the neural responses typically observed in the Why/How contrast, by conducting a secondary set of analyses, which we report in detail in the Supplementary Materials. Briefly, we estimated two additional models for each participant. The first modeled the Why/How contrast across high-accuracy

Why questions and low-accuracy How questions, such that Why questions elicited significantly higher accuracy rates than did How questions. The second modeled the Why/How contrast across the Why questions eliciting the fastest RTs and the How questions eliciting the slowest RTs, such that Why questions elicited significantly faster RTs than did How questions. As listed in Table S2, both analyses strongly replicate the results presented below, demonstrating that performance variability cannot explain the effects reported here.

Brain regions modulated by the Why/How contrast

The Why > How contrast isolated a largely left-lateralized set of cortical regions that are anatomically consistent with meta-analytic definitions of the ToM Network (Fig. 2A) and with the regions observed in our published studies that used an open-answer response protocol to achieve the Why > How contrast for intentional actions and emotional facial expressions (Fig. 2B; Spunt and Lieberman, 2012a,b). These regions span dorsomedial, ventromedial, and lateral orbital areas of the prefrontal cortex (PFC); a medial parietal area spanning the posterior cingulate cortex and precuneus (PCC/PC); the left temporoparietal junction (TPJ); and the anterior superior temporal sulcus (aSTS) bilaterally (Table 2). In addition, we observed a right-lateralized response in the posterior lobe of the cerebellum that is also consistent with our prior work as well as a recently published meta-analysis demonstrating reliable cerebellar responses to higher-order social cognition (Van Overwalle et al., 2013).

As also listed in Table 2, the How > Why comparison isolated a set of cortical regions including an area of the left lateral occipital cortex and left superior parietal lobule, as well as several other areas of the parietal lobe bilaterally, including the intraparietal sulcus, supramarginal gyrus, and dorsal precuneus.

Study 2

Study 1 introduced a novel implementation of the Why/How contrast that overcomes the central limitation of the version used in our previous research, which relied on covert answers to open-ended questions (Spunt and Lieberman, 2012a,b, 2013; Spunt et al., 2010, 2011). The resulting task produces functional contrast in a set of brain regions that converge both with those observed using the original Why/How Task and with meta-analytical definitions of the ToM Network (Fig. 2). The primary motivation of Study 2 was to determine the extent to which the Why/How contrast is distinct when compared to the Belief/Photo contrast provided by the False-Belief Localizer (Dodell-feder et al., 2011; Saxe and Kanwisher, 2003; Saxe and Powell, 2006), which remains the only existing standardized contrast for investigating the neural correlates of ToM. To perform the comparison, we capitalized on the fact that a subset of the participants in Study 1 had earlier participated in a separate fMRI study, which included both the False-Belief Localizer as well as an earlier version of the Yes/No Why/How Task used in Study 1. This allowed us to evaluate the dissimilarity of the Belief/Photo and Why/How contrasts in the same set of participants. If we find that the responses produced by the Why/How contrast are dissimilar from those produced by the Belief/Photo contrast, then the former can be said to feature high discriminant validity, a desirable psychometric property for novel test instruments (Campbell, 1960; Campbell and Fiske, 1959).

Materials and methods

Participants

The data used in the present study was collected as part of a larger study (currently unpublished) that featured a subset of ten of the participants from Study 1 (6 males, 4 females; mean age = 26.50, age range = 21–38), and the procedures for recruiting, screening, consenting, and compensating them were identical to those used in Study 1 (the average amount of time between Study 1 and the present study was 30.33 days). This study included a single fMRI session that included an earlier version of the Yes/No Why/How Task used in Study 1

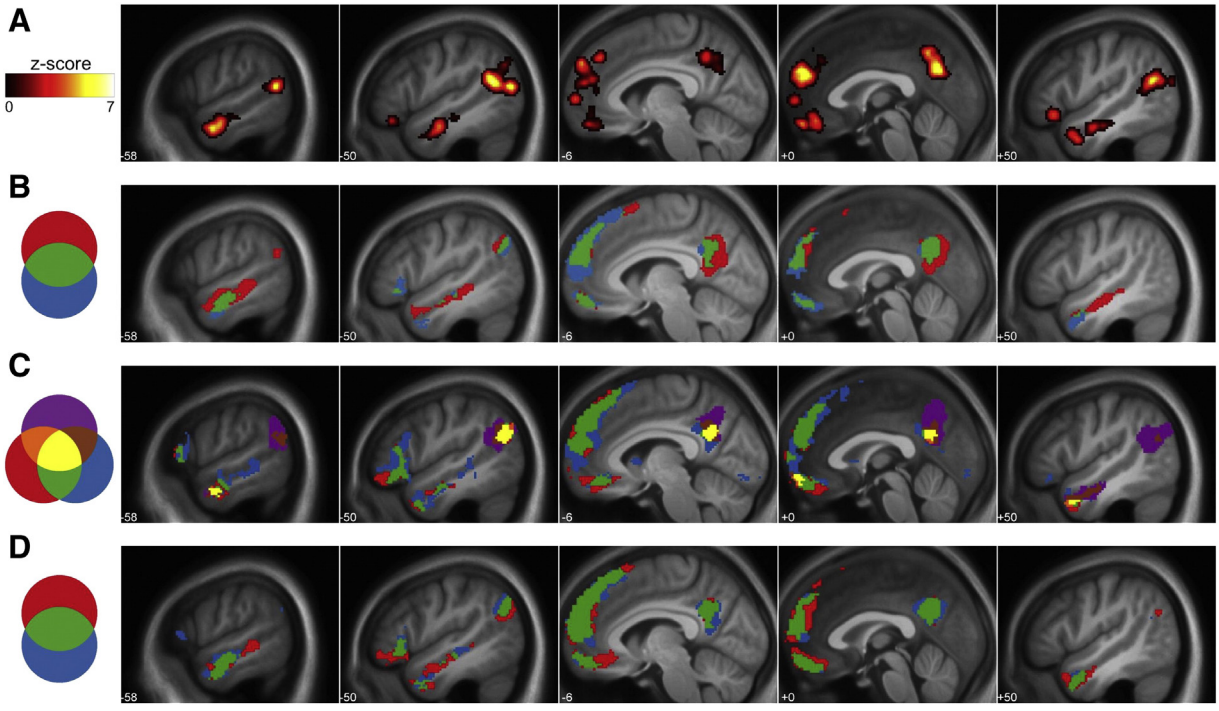

Fig. 2. Sagittal sections displaying (A) regions of the putative Theory-of-Mind (ToM) Network defined using the automated meta-analysis software Neurosynth (Yarkoni et al., 2011); (B) conjunction and disjunction of the top 10% activated voxels in the original (open-ended) implementation of the Why > How contrast (Blue; data from Spunt and Lieberman, 2012a, 2012b) and the new (yes/no) version of the Why > How contrast from Study 1 (Red); (C) conjunction and disjunction of the top 10% activated voxels for the comparisons described in Study 2, where: Red = Why > How Contrast (Study 1), Blue = Why > How Contrast (Study 2), and Purple = Belief > Photo Contrast (Study 2); and (D) conjunction and disjunction of the top 10% activated voxels in the Why/How contrast estimated on the full samples in Study 1 (N = 29; Blue) and Study 3 (N = 21; Red).

(differences described below) and the publicly available version of the False-Belief Localizer (Dodell-Feder et al., 2011).

Why/How contrast

As noted above, the version of the Yes/No Why/How Task differed in several ways from the one introduced in Study 1. These differences were motivated by the specific questions being investigated in the larger study for which it was designed. First, in addition to the two categories of social behavior featured in Study 1 (intentional actions and emotional expression), this version featured why- and how-questions about a third stimulus category showing the effects of non-social processes, for instance, scenes showing the consequences of extreme weather. The comparison of social to non-social stimuli is outside the scope of the present report and will not be discussed further. Second, the blocks for this version each included nine rather than seven trials. Third, this version featured a small number of differences in the specific question–photograph pairs used to achieve the Why/How manipulation. All questions are provided in Table S1. Finally, there were small differences in the timing of the trial elements within each block, and with the average stimulus onset asynchrony. In light of these differences combined with the nonsocial Why/How condition, this version had a total runtime of 16 min, 35 s.

False-Belief Localizer

Participants performed the most recent version of the publicly available False-Belief Localizer (Dodell-feder et al., 2011; http://saxelab.mit.edu/tomloc.zip, version Sept. 7, 2011). Given that the task has been described extensively elsewhere, we only briefly describe it here. The contrast is formed by comparing two conditions, both of which involve reading a short story and judging the veracity of brief statement about the events described in the story. *Belief* stories describe the events that lead one or more characters to form a false belief about the world, while *Photo* stories describe the events that lead a physical representation of the world (e.g., a photograph, map, or sign) to become outdated or misleading. Henceforth, we refer to the comparison of these conditions as the Belief/Photo contrast. Although we made no changes to the original stimuli, we modified the timing of the task so that presentation durations were self-paced within a fixed time window. Prior to performing the task, participants were shown an example trial and were invited to ask questions before beginning. Total run time of the task was 8 min, 50 s.

Image acquisition

Image acquisition parameters differed only in the number of EPI volumes acquired for each task: 398 volumes were acquired for the Why/How contrast, while 212 volumes were acquired for the Belief/Photo contrast.

Image analysis

The image preprocessing pipeline was identical to the one used in Study 1. Why/How model specification differed only in the inclusion of covariates of no interest modeling responses to the non-social stimulus category. The model for the False-Belief Localizer task included effects for the two conditions of interest: Belief and Photo. Each trial was modeled as a variable epoch spanning the onset of Story presentation and the offset of the Judgment period. In addition to the nuisance covariates of no interest described in the methods of Study 1, we also included a single parametric regressor modeling the total duration of each block. This regressor ensures that the Belief > Photo contrast is not confounded with time on task.

To evaluate the claim that the Why/How contrast is distinct from the Belief/Photo contrast, we compared their group-level activation maps. To test for common areas of activation, we used their minimum statistic to test the conjunction null (Nichols et al., 2005). To test for statistically different levels of activation, we entered participants' contrast images for the effects of each condition for both tasks into a single, random-effects analysis using a flexible factorial repeated-measures ANOVA (within-subject factors: Why/How task, condition; blocking factor: subject). Within this model, we tested the Task-by-Condition interaction to determine regions that are differentially modulated in the two contrasts.

To supplement these univariate analyses, we employed an analytical strategy known as representational similarity analysis (Kriegeskorte et al., 2008) in order to evaluate the similarity structure of the multivariate patterns of activity that characterize the Why/How and Belief/Photo contrasts. Activity patterns were extracted from a mask of voxels showing a preferential association with prior neuroimaging studies of theory-of-mind and mentalizing. To create the mask, we used the automated meta-analysis tool Neurosynth (Yarkoni et al., 2011; http://neurosynth.org/features) to download a reverse inference map that shows the likelihood that the term "mentalizing" was used in a study if activation was reported at a particular voxel. We used the term "mentalizing" because (a) it is used interchangeably with the phrase "Theory of Mind", and (b) Neurosynth does not currently offer a map for the phrase "Theory of Mind". When creating the mask, we included only those clusters larger than 75 voxels. Neurosynth was used to define our reference mask for three reasons. First, it is the most unbiased method available, based entirely on automated text mining of 5809 published neuroimaging articles. Two, it is the most transparent method available, in that the data is publicly available for download. Finally, it produces a map that is consistent with published meta-analyses of neuroimaging studies of ToM (Carrington and Bailey, 2009; Denny et al., 2012; Mar, 2011; Schurz et al., 2014; Van Overwalle and Baetens, 2009).

For each of the 10 participants, we extracted the t-statistic values within the mentalizing mask from the voxels achieving threshold in the previously described Why/How contrast estimated in the same session; the same Why/How contrast estimated in a second session; and their Belief/Photo contrast itself. Each of these sets of voxels could then be considered as a vector, and were correlated. The Pearson correlation coefficient thus quantified, for each participant, the consistency of the multivariate activity patterns across the three contrasts. We then used a paired samples *t*-test on the Fisher z-transformed correlations to verify that the two Why/How contrasts were more similar to one another than either were to the belief/photo contrasts. We represented the similarity structure in two ways (Figs. 3B and C). Fig. 3B shows a representational dissimilarity matrix (RDM) showing the degree of pairwise dissimilarity among the three contrasts estimated for each of the ten participants: the Why/How contrast from Study 1 (rows/columns 1–10; Why/How_{S1}); the same contrast from an earlier study (rows/columns 11–20; Why/How_{S2}); and the Belief/Photo contrast (rows/columns 21–30). The dissimilarity measure used is 1 minus the Pearson correlation (r) and ranges from 0 (perfect correlation) to 2 (perfect anti-correlation). Because the order of participants is the same across the three blocks of contrasts, the diagonals within each block represent within-subject pattern dissimilarities, while the off-diagonals represent between-subject dissimilarities. Also shown in Fig. 3C is a two dimensional representation of the similarity structure based on applying multidimensional scaling to the RDM. Each colored circle represents a single contrast image, and contrast images for the same participant are connected by dotted lines. The length of these lines corresponds to the dissimilarity of the multivariate patterns.

Unless otherwise specified, all analyses were interrogated using a cluster-level family-wise error (FWE) rate of .05 with a cluster-forming voxel-level *p*-value of .001. For visual presentation, thresholded t-statistic maps are overlaid on the average of the participants' T1-weighted anatomical images.

Results

Performance

For the Why/How Task, participants were again slightly more accurate in their responses when answering How (M = 92.59%, SD = 5.15%)

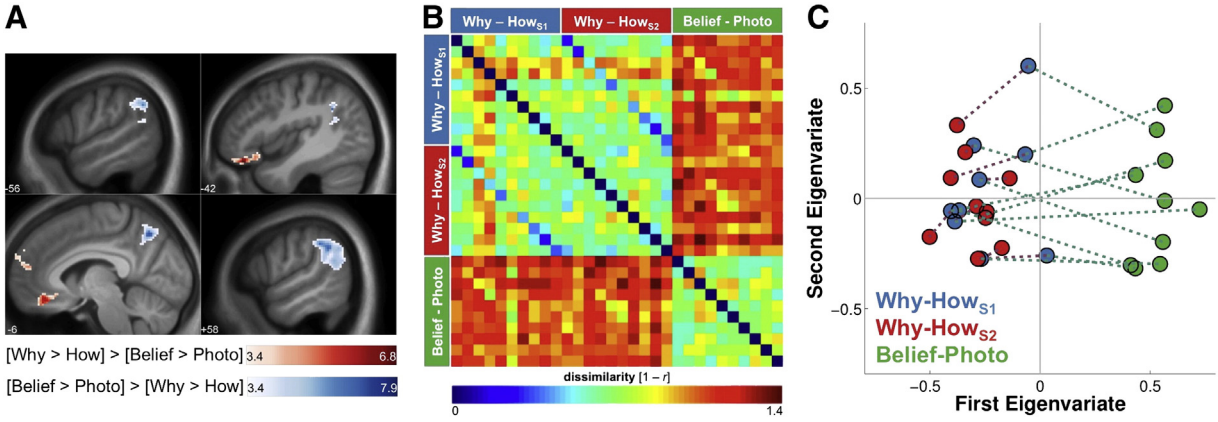

Fig. 3. Univariate and multivariate similarities across tasks. (A) Comparison of the univariate (voxel-wise) responses to the two contrasts from Study 2 (N = 10; cluster-level corrected across the whole-brain). The prefrontal regions depicted with a red colormap showed a stronger response to the Why/How contrast, while the medial parietal and temporoparietal regions depicted with a blue colormap showed a stronger response to the Belief/Photo contrast. (B) Comparison of the multivariate (multivoxel) response patterns produced by the Why/How and Belief/Photo contrasts within the meta-analytically defined regions of the Theory-of-Mind Network shown in Fig. 2a. This panel uses a representational dissimilarity matrix (RDM) to visualize the degree of pairwise dissimilarity among the response patterns produced by the three contrasts estimated for each of the ten participants: the Why/How contrast from Study 1 (rows/columns 1–10; Why/How$_{S1}$); the Why/How contrast from Study 2 (rows/columns 11–20; Why/How$_{S2}$); and the Belief/Photo contrast (rows/columns 21–30). The dissimilarity metric is 1 minus the Pearson correlation (r), where a value of 0 indicates perfect correlation; 1 indicates non-correlation; and 2 indicates perfect anti-correlation. Because the order of participants is constant across the three blocks of contrasts, the diagonals within each block represent within-subject pattern dissimilarities, while the off-diagonals represent between-subject dissimilarities. (C) A two-dimensional representation of the similarity structure based on multidimensional scaling applied to the RDM. Each colored circle represents a single contrast image, and constrast images for the same participant are connected by dashed colored lines. The length of these lines is the Euclidean distance between them, with longer lines representing more dissimilar multivariate patterns.

compared to Why (M = 91.02%, SD = 5.20%) questions, $t(9) = 2.613$, $p = .028$, 95% CI [−2.937, −0.211]. In addition, participants were faster when answering How (M = 831 ms, SD = 128 ms) compared to Why (M = 901 ms, SD = 117 ms) questions, $t(9) = 4.851$, $p = .001$, 95% CI [37, 102]. This replicates the behavioral effects observed in Study 1.

For the False-Belief Localizer, accuracy did not differ across the Belief (M = 73%, SD = 21.108%) and Photo (M = 76%, SD = 15.056%) conditions, $t(9) = -.758$, $p = .468$. Similarly, response time (Story onset to Judgment) did not differ across the Belief (M = 14.381 s, SD = 3.42 s) and Photo (M = 13.608 s, SD = 3.812 s) conditions, $t(9) = 1.719$, $p = .120$. Despite the lack of differences across the conditions, the neuroimaging analysis of the False-Belief Localizer presented below control for variability in trial duration using the same procedures used in the analysis of the Why/How Task data.

Finally, we determine the extent to which performance was correlated across the three tasks. Although accuracy to Why trials was positively correlated across the two versions of the Why/How Task, r(8) = 0.670, p = 0.034, 95% CI [0.070, 0.914], neither was positively correlated with accuracy for Belief trials in the False-Belief Localizer (ps > .589). Similarly, although accuracy for How trials was positively correlated across the two versions of the Why/How Task, r(8) = 0.706, p = 0.022, 95% CI [0.138, 0.925], neither was positively correlated with accuracy for Photo trials in the False-Belief Localizer (ps > .641). This provides behavioral evidence for discriminant validity in the behavior being measured by the two tasks.

Comparison of the Why/How and Belief/Photo contrasts

Table 3 lists the results of the comparison of the Why/How and Belief/Photo contrasts. Only two regions were observed to be jointly activated by both tasks: left temporoparietal junction and posterior cingulate cortex. Of the total number of voxels activated above threshold by each contrast, this common activity accounted for only 11% and 9% in the Why/How and Belief/Photo contrasts, respectively. Moreover, when statistically comparing the magnitude of the two contrasts, a clear pattern emerges: whereas medial and orbital prefrontal regions responded to Why/How more so than to belief/photo, medial parietal and temporoparietal regions responded to belief/photo more so than to Why/How.

The above shows that the two tasks differ in the magnitude of the response in the putative ToM Network. Next, we examined the extent to which they differ in the patterns of spatially distributed activity evoked in an independently defined mask of regions associated with mental state inference in prior work as identified by the automated meta-analysis tool Neurosyth (Yarkoni et al., 2011). If the Why/How contrast produces a pattern of activity that diverges from the one produced by the Belief/Photo contrast, each participant's Why/How contrast should show a pattern of activity that is more similar to a previous Why/How contrast than to their Belief/Photo contrast. To evaluate differences in the pairwise similarities across the three contrasts, we used the Fisher's z-transformed Pearson correlation of the multivariate response pattern across the ToM Network mask. Whereas we observed no evidence for a correlation of the Why/How and Belief/Photo contrasts ($r_{mean} = 0.02$, $r_{sd} = 0.17$), such a correlation was apparent across the two versions of the Why/How contrast ($r_{mean} = 0.72$, $r_{sd} = 0.20$), and the difference between these two sets of correlations was significant, t(9) = 7.091, p < 0.001, 95% CI [0.478, 0.926]. Fig. 3B represents each participants' response pattern in two-dimensions using multidimensional scaling. A similarly significant difference in the

Table 3
Whole-brain comparisons of the Why/How and Belief/Photo contrasts in Study 2 ($N = 10$). All listed regions survive a whole-brain analysis thresholded with a cluster-level family-wise error rate of .05 and a cluster-forming threshold of $p < .001$. PFC = prefrontal cortex; OFC = orbitofrontal cortex; k = voxel extent of the cluster containing the peak; x, y, and z = Montreal Neurological Institute (MNI) coordinates in the left-right, anterior-posterior, and inferior-superior dimensions, respectively.

Contrast name				MNI coordinates		
Region name	L/R	*k*	t-value	x	y	z
[Why > How] & [Belief > Photo]						
Temporoparietal junction	L	204	7.630	−48	−66	26
Posterior cingulate cortex	L	120	5.430	−6	−56	28
[Why > How] > [Belief > Photo]						
Lateral OFC	L	227	6.827	−42	36	−18
Ventromedial PFC	L	187	6.204	−8	38	−16
Dorsomedial PFC	L	107	4.642	−6	54	16
[Belief > Photo] > [Why > How]						
Precuneus	R	165	7.898	2	−58	22
Posterior cingulate cortex	L	597	6.875	−8	−62	46
Temporoparietal junction	R	1105	6.404	46	−50	24
	L	362	6.149	−44	−50	34
Superior frontal gyrus	R	221	5.977	20	10	56

multivoxel patterns of activation evoked by the two contrasts was obtained if the entire gray matter mask was used (Fig. S3).

Study 3

The purpose of Study 3 was threefold. First, we sought to validate a more efficient version of the Why/How Task by eliminating questions and photographs that did not elicit high response consensus in Study 1. Second, we sought to demonstrate in a new group of participants the reproducibility of the behavioral and neural effects observed in Studies 1 and 2. Third, we sought to demonstrate the feasibility of using the optimized Why/How Task, which has a total runtime of just 5 min, as a localizer task for defining functional ROIs in individual participants.

Materials and methods

Participants

- Participants were a completely new set of twenty-one right-handed adults (10 males, 11 females; mean age = 27.62, age range = 19–38) recruited from the greater Los Angeles area. The procedures for recruiting, screening, consenting, and compensating these participants were identical to those used in Studies 1 and 2.

Why/How contrast

The version of the Why/How Task employed here was identical to the one used in Study 1 except for the following changes, all of which were intended to reduce the total runtime of the task. The total number of question blocks was reduced to 16 (see Table 1), and the number of photographs per block was increased to 8. Based on the pilot data used to select the items included in Study 1, for each block the consensus response was 'yes' for 5 photos and 'no' for the remaining 3. In addition, we introduced minor modifications to the timing of the task as depicted in Fig. 1. These modifications were justified by observation from Study 1 that participants were not only quite efficient (mean RT ranged from 574 to 1141 ms) but exhibited near-ceiling accuracy rates (mean accuracy ranged from 86 to 100%). Collectively, these changes yielded a version of the task with a total runtime of 5 min, 12 s. The stimuli and MATLAB code for presenting and scoring the task can be downloaded at http://www.bobspunt.com/whyhow-localizer/.

Image acquisition

Image acquisition procedures differed only in the use of a multiband excitation sequence to acquire 3212 EPI volumes (acceleration factor = 4; slice thickness = 2.5 mm, 56 slices, TR = 1000 ms, TE = 30 ms, flip angle = 60°, matrix = 80 × 80, FOV = 200 mm).

Image analysis

Image preprocessing and model specification aspects of the analysis pipeline were identical to those described in Studies 1 and 2.

Results

Performance

We replicate the behavioral effects observed in Studies 1 and 2: Participants were more accurate in their responses when answering How (M = 95.76%, SD = 3.17%) compared to Why (M = 91.96%, SD = 3.93%) questions, $t(20) = 3.302$, $p = .004$, 95% CI [−6.192, −1.398]. In addition, participants were faster when answering How (M = 611 ms, SD = 87 ms) compared to Why (M = 686 ms, SD = 108 ms) questions, $t(20) = 5.625$, $p < .001$, 95% CI [47, 102].

Brain regions modulated by the Why/How contrast

As shown in Fig. 2D and listed in Table 4, a whole-brain search confirmed that the 5-minute version of the Why/How Task continues to produce a robust, group-level response in the same brain networks observed in Studies 1 and 2.

Reliability of single-subject localization

Finally, we sought evidence pertaining to the feasibility of using the 5-minute version of the Why/How Task as a localizer of functional ROIs in individual participants For each region identified in the whole-brain contrast, we determined the percentage of participants for which a cluster of at least 10 voxel extent could be identified after thresholding each participants' single-subject Why/How contrast using a cluster-level family-wise error rate of .05. As shown in Table 4, this criterion allowed us to detect activity in most regions in at least 80% of participants. This was true for regions both activated or deactivated in the Why > How contrast. This demonstrates the inter-subject consistency of the Why/How contrast, and validates its use as an efficient functional localizer. As described above, we have made this version of the task publicly available under the name Why/How Localizer.

Functional lateralization

As described in more detail in the Supplementary Materials, we used the pooled data from Study 1 and the present study ($N = 50$) to determine the extent to which the degree of lateralization present in the Why > How contrast is statistically reliable. This is motivated by the second problem identified in the Introduction, namely, that anatomical definitions of the ToM Network remain imprecise. If the regions associated with the Why > How contrast show a response that is reliably lateralized, this would further increase the precision of its anatomical definition. The results of this analysis are listed in Table S3: the network evoked by the Why/How localizer was strongly left-lateralized. Of all the cortical regions associated with the Why > How contrast, only the posterior cingulate cortex failed to show left hemisphere selectivity. The single region to show evidence of right hemisphere selectivity was in the posterior lobe of the cerebellum.

Discussion

Taken together, the three studies presented here validate the Why/How contrast for functional MRI studies of ToM. In Study 1, we introduced an improved protocol for achieving the Why/How contrast and showed that it activates a largely left-lateralized network that converges both with our prior work (Spunt and Lieberman, 2012a,b, 2013; Spunt et al., 2010, 2011) and with meta-analytic definitions of the ToM Network. In Study 2, we showed that within the same set of participants, the network activated by the Why/How contrast is reliable across testing sessions, and is clearly distinct from the network activated by the only existing standardized protocol for investigating the neural bases of using ToM, the False-Belief Localizer (Dodell-feder et al., 2011; Saxe and Kanwisher, 2003). In Study 3, we showed that the network is reproducible in a completely new group of participants, demonstrated the feasibility of using the new Why/How protocol as an efficient functional localizer at the single-subject level. Finally, across all studies, we found that the new Why/How Task yields reliable behavioral effects. Taken together, these findings validate a novel instrument for manipulating a distinct use of ToM and assessing both its behavioral and neural correlates.

We believe that this instrument helps solve the two problems with previous neuroimaging work on ToM that were identified in the Introduction. The first problem regarded the fact that despite the enormous number of studies that have been devoted to investigating the neural bases of different uses of ToM (Carrington and Bailey, 2009; Denny et al., 2012; Lieberman, 2010; Mar, 2011; Schurz et al., 2014; Van Overwalle and Baetens, 2009), there has been relatively little attention devoted to the evaluation and standardization of the behavioral methods used in these studies. We hope that the study presented here will help reverse this trend and ultimately define transparent criteria for evaluating the quality of the behavioral methods used in neuroimaging studies.

Table 4
Group-level results of the Why/How contrast from Study 2 ($N = 21$). All peaks survive a whole-brain search thresholded at a voxel-wise family-wise error rate of .05 and a cluster extent (k) of at least 10 voxels. The regions of interest (ROI) used to constrain the search for single-subject voxels were created by growing spheres (radius = 12 mm) around each peak observed at the group-level and intersecting the resulting spherical volume with the group-level t-statistic map thresholded using a voxel-wise p-value of .001. The single-subject columns display the percentage of subjects (of 21) for whom a cluster of at least 10 voxel extent could be identified in each ROI after thresholding the single-subject t-statistic image using cluster-level correction at a family-wise error rate of .05. Values in the column "Mean k" show the average extent of the found clusters. PFC = prefrontal cortex; OFC = orbitofrontal cortex; STS = superior temporal sulcus; MTG = middle temporal gyrus; x, y, and z = Montreal Neurological Institute (MNI) coordinates in the left–right, anterior–posterior, and inferior–superior dimensions, respectively.

Contrast name				MNI coordinates			Single-subject	
Region name	L/R	Extent	t-value	x	y	z	%	Mean k
Why > How								
Dorsomedial PFC	L	1415	14.516	−8	62	22	95	518
	L	–	12.730	−20	34	44	86	315
	L	–	8.004	−6	48	0	57	148
	R	26	8.238	8	58	28	86	338
	R	19	8.072	14	50	40	52	278
Ventromedial PFC	L	115	8.686	−2	46	−18	76	264
Lateral OFC	L	52	9.613	−42	30	−14	67	337
	L	25	7.820	−48	22	−2	71	250
Temporoparietal junction	L	103	9.827	−46	−62	32	100	384
	R	23	8.021	56	−62	24	57	174
Posterior cingulate cortex	L	349	11.261	−4	−50	32	81	430
Temporal pole	L	78	11.223	−46	6	−34	67	235
Anterior STS	L	142	10.592	−60	−12	−16	81	364
	R	57	9.155	54	0	−28	38	256
How > Why								
Intraparietal sulcus	L	134	11.741	−40	−40	44	95	348
	R	49	8.817	46	−34	42	95	412
	R	12	7.676	38	−46	54	86	439
Supramarginal gyrus	L	44	9.144	−60	−28	36	90	389
	R	16	8.547	56	−38	30	76	294
	R	44	8.319	60	−24	38	81	444
Posterior MTG	L	65	10.814	−52	−60	0	86	293
IFG (opercularis)	R	57	9.026	48	10	16	67	154
Precuneus (dorsal)	L	30	8.315	−12	−66	54	76	372
	R	22	7.889	18	−64	56	67	386

The second problem regarded the fact that neuroanatomical definitions of the putative ToM Network remain highly imprecise. The cause of this imprecision is no doubt partially attributable to the first problem, in that the different tasks used to investigate ToM activate different regions of the brain (Gobbini et al., 2007; Schurz et al., 2014). Indeed, we found that with both univariate and multivariate measures, the Why/How contrast is remarkably distinct when compared to the Belief/Photo contrast (discussed further below). Of equal importance is our observation that the neuroanatomical correlates of the Why/How contrast are highly reliable, both within and across participants, and in our right-handed participants showed a reliable left-lateralization. Moreover, our data suggests that by using the publicly available Why/How Localizer, future studies can localize this network in individual participant's in as little as 5 min.

This level of anatomical specificity is largely lacking from extant neuroimaging work on ToM, which has relied almost exclusively on qualitative reviews or large meta-analyses when defining the boundaries of ToM. To be clear, our aim is not to claim that the network identified by the Why/How contrast is a precise representation of the ToM Network. On the contrary, we think that a central part of the problem is the generally well-accepted idea that there is a single network in the human brain that supports a monolithic ToM ability. This idea seems to have encouraged a disproportionate focus on what is common across the many faces of ToM, both in how it is operationally defined and in where it shows up in the brain. The present studies demonstrate that, moving forward, increased attention will need to be paid to conceiving ToM not as a single ability, but as collection of abilities that may function differently depending on the person and the context.

Evaluating the new Why/How task: strengths and limitations

We believe the new implementation of the Why/How contrast has several notable strengths that make it a powerful instrument for probing the neurobiological bases of social cognition. At the same time, we acknowledge its limitations.

The task permits use of complex, naturalistic social stimuli

As in the original implementation of the Why/How contrast, the manipulation is attentional in that the Why and How questions are asked of the same set of photographs. This permits use of complex, naturalistic nonverbal social stimuli while avoiding concerns about the innumerable differences that can emerge across such stimuli, such as differences in low-level visual properties, proportion of particular objects shown, or emotional meaning.

We note two caveats in our definition of the Why/How contrast as an attentional manipulation. The first caveat regards the fact that although the photographs are invariant across the Why and How conditions, the reminder cues briefly presented between each photograph naturally varied as a function of the question being asked. This was seen as a desirable task feature that effectively eliminated any working memory demands caused by having to remember the question for the duration of the block. Given that the reminder cues are presented very briefly (350 ms in the Study 1 version; 300 ms in the Study 3 version), and that the results converge with previous Why/How studies using a pure attentional manipulation, we believe it is highly unlikely that these verbal stimuli provide a sufficient explanation for the effects observed in the new Why/How contrast.

A second caveat regards the possibility that Why versus How questions differentially lead subjects to allocate attention onto, or to fixate, particular features of the nonverbal stimuli. Eyetracking could explore the latter possibility (although it is unlikely to show large differences, given the relatively small visual angle subtended by the stimuli in the first place). However, attentional issues are harder to isolate. In fact, we think it likely that differential allocation of attention onto particular features of the stimulus may be part and parcel of the differential demand of answering why versus how questions. Whether attention is

differentially allocated to features of the images, or to associations we have for those features, surely at some level differential attention will need to come into play. Rather than a confound, we would suggest this could be a fruitful line of research in its own right.

The task constrains response content and measures performance

As described above, the original Why/How Task used open-ended Why and How questions to evoke covert responses to social stimuli. Although this method of responding has the desirable feature of being highly naturalistic, it prevents experimental control of response content and performance measurement. The evaluative response method used in the new Why/How contrast represents a significant improvement in that it is designed to evoke well-normed consensus responses, and therefore yields accuracy and response time (RT) measures. In the present study, this allowed us to identify a reliable behavioral difference across Why and How questions on both accuracy and RT outcomes. With such well-characterized behavioral effects, we were able to conclusively demonstrate that performance-related variability does not provide a sufficient explanation for the response in the cortical regions observed in the Why/How contrast (Table S2).

A potential limitation regards the fact that the accuracy of a given response is based solely on the consensus of an independently acquired group of healthy, English-speaking, American citizens. This is particularly true in the case of understanding answers to Why questions, which typically draw heavily on knowledge that is likely to be culturally specific. Given this, we clarify that the validity of the accuracy measurement assumes that the respondent has the cultural knowledge necessary for arriving at the answer that elicited consensus in the reference normative sample. While posing some degree of methodological limitation, this feature also opens the door for exciting variations on the task. For instance, one could compare consensus responses across different cultures. Or one could investigate responses in clinical populations who have atypical inferences, such as people with autism spectrum disorders (work currently ongoing in our laboratory). In all of these cases, one can reference the respondents' answer to the normative response, to a group-specific response (e.g., obtained from the participants in that study beforehand), and one could even derive individually idiosyncratic responses, allowing investigations of universals, culturally or group-specific processing, and individual differences.

The task has convergent validity

The new Why/How contrast activates a brain network that is convergent with the network typically observed in the original Why/How studies (Fig. 2B). Although suggestive, this is not conclusive evidence that the two versions are interchangeable manipulations of the same underlying process. Indeed, although the two versions are conceptually similar by design, they have obvious differences, the most notable of which is the method of eliciting responses. Given the substantial improvements offered by the new version, we certainly prefer it moving forward, but also suggest that investigating the nature of possible differences in processing demands evoked by the two versions is a worthwhile line for future research.

The task has discriminant validity

We found that the Why/How contrast show very little overlap with the Belief/Photo contrast produced by the False-Belief Localizer, and that even within an objectively-defined meta-analytic mask of the putative ToM Network, the two contrasts show no evidence of a correlation in their spatially distributed activity patterns. In parallel, response accuracy was not correlated across the two tasks. As such, the Why/How contrast demonstrably taps into a process, or set of processes, that are part of our broad set of abilities to think about the internal states of other people, but that are largely separate from those specifically isolated by the Belief/Photo contrast. Importantly, this does not demonstrate that the Why/How contrast is an alternative or improvement upon the Belief/Photo contrast. On the contrary, the data show that the two are in fact complementary, providing methods for targeting different uses of ToM, measuring different behavioral outcomes, and modulating different brain networks.

The task is flexible

Although we have made the Study 3 version of the task publicly available as a standardized functional localizer, we believe it is worthwhile to highlight the adaptability of the task for a wide array of distinct research questions. Such questions fall into roughly three categories corresponding to variation in the stimulus being evaluated (e.g., facial expressions vs. hand actions, as in the present version); variation in the question being answered (e.g., questions about belief vs. motive); and variation in the person answering the question (e.g., clinical populations). Given the adaptability of the basic protocol, the existence of a standardized protocol, and a growing body of normative data using variants of the Why/How contrast, this task provides a rich opportunity for cumulative research on the neurobiological bases of a specific use of ToM.

Conclusion

We believe the Why/How contrast is a method for investigating a natural way in which human beings use their ToM to understand their own and other people's behaviors. It elicits an anatomically circumscribed and highly reproducible response in the healthy human brain. Although this response resembles the putative ToM Network, we intentionally avoid calling it by that name. Moving forward, we encourage the field to relax its dependence on this misleading label that implicitly endorses the tentative view that ToM is a single ability implemented in a single brain network. There may well be some validity to this singular view of ToM, but even if so, it seems unreasonable to assume that its neural implementation and behavioral expression would appear the same across the many different tasks and measures used to study it. The Why/How Task is one such measure. We would hope that our study catalyzes similar efforts, not just for evaluating extant methods, but developing and validating new ones. The result will be a description of ToM that is as rich as the role it plays in human sociality.

Acknowledgments

The Authors would like to acknowledge Mike Tyszka, Tim Armstrong, and the Caltech Brain Imaging Center for help with the neuroimaging; the Caltech Conte Center for Social Decision-Making (P50MH094258) for funding support; and two anonymous Reviewers for their comments.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.neuroimage.2014.05.023.

References

Apperly, I., 2012. What is "theory of mind"? Concepts, cognitive processes and individual differences. Q. J. Exp. Psychol. 65 (5), 825–839. http://dx.doi.org/10.1080/17470218.2012.676055.
Ashburner, J., 2007. A fast diffeomorphic image registration algorithm. Neuroimage 38 (1), 95–113. http://dx.doi.org/10.1016/j.neuroimage.2007.07.007.
Brainard, D.H., 1997. The psychophysics toolbox. Spat. Vis. 10 (4), 433–436. http://dx.doi.org/10.1163/156856897X00357.
Campbell, D.T., 1960. Recommendations for apa test standards regarding construct, trait, or discriminant validity. Am. Psychol. 15 (8), 546.
Campbell, D.T., Fiske, D.W., 1959. Convergent and discriminant validation by the multitrait–multimethod matrix. Psychol. Bull. 56 (2), 81.
Carrington, S., Bailey, A., 2009. Are there theory of mind regions in the brain? A review of the neuroimaging literature. Hum. Brain Mapp. 30 (8), 2313–2335. http://dx.doi.org/10.1002/hbm.20671.
Dennett, D.C., 1989. The intentional stance. , The MIT Press, Cambridge.
Denny, B., Kober, H., Wager, T., Ochsner, K., 2012. A meta-analysis of functional neuroimaging studies of self- and other judgments reveals a spatial gradient for mentalizing

in medial prefrontal cortex. J. Cogn. Neurosci. 24 (8), 1742–1752. http://dx.doi.org/10.1162/jocn_a_00233.
Diedrichsen, J., Shadmehr, R., 2005. Detecting and adjusting for artifacts in fMRI time series data. Neuroimage 27 (3), 624–634. http://dx.doi.org/10.1016/j.neuroimage.2005.04.039.
Dodell-feder, D., Koster-hale, J., Bedny, M., Saxe, R., 2011. fMRI item analysis in a theory of mind task. Neuroimage 55 (2), 705–712. http://dx.doi.org/10.1016/j.neuroimage.2010.12.040.
Dufour, N., Redcay, E., Young, L., Mavros, P., Moran, J., Triantafyllou, C., Saxe, R., 2013. Similar brain activation during false belief tasks in a large sample of adults with and without autism. PLoS ONE 8 (9), e75468. http://dx.doi.org/10.1371/journal.pone.0075468.
Gobbini, M., Koralek, A., Bryan, R., Montgomery, K., Haxby, J., 2007. Two takes on the social brain: a comparison of theory of mind tasks. J. Cogn. Neurosci. 19 (11), 1803–1814. http://dx.doi.org/10.1162/jocn.2007.19.11.1803.
Gopnik, A., Wellman, H.M., 1992. Why the child's theory of mind really is a theory. Mind Lang. 7 (1-2), 145–171.
Grinband, J., Wager, T., Lindquist, M., Ferrera, V., Hirsch, J., 2008. Detection of time-varying signals in event-related fMRI designs. Neuroimage 43 (3), 509–520. http://dx.doi.org/10.1016/j.neuroimage.2008.07.065.
Heider, F., 1958. The Psychology of Interpersonal Relations. Wiley, New York.
Heider, F., Simmel, M., 1944. An experimental study of apparent behavior. Am. J. Psychol. 57 (2), 243. http://dx.doi.org/10.2307/1416950.
Jones, E.E., Davis, K.E., 1965. From acts to Dispositions the Attribution Process in Person Perception. In: Berkowitz, L. (Ed.), , Vol. 2. Elsevier, San Diego, pp. 219–266. http://dx.doi.org/10.1016/S0065-2601(08)60107-0.
Kelley, H.H., 1973. The processes of causal attribution. Am. Psychol. 28 (2), 107–128. http://dx.doi.org/10.1037/h0034225.
Kennedy, D., Adolphs, R., 2012. The social brain in psychiatric and neurological disorders. Trends Cogn. Sci. 16 (11), 559–572. http://dx.doi.org/10.1016/j.tics.2012.09.006.
Kriegeskorte, N., Mur, M., Bandettini, P., Mur, M., Bandettini, P., Kriegeskorte, N., 2008. Representational similarity analysis — connecting the branches of systems neuroscience. Front. Syst. Neurosci. 2 (4), 1–28.
Leslie, A., Friedman, O., German, T., 2004. Core mechanisms in "theory of mind". Trends Cogn. Sci. 8 (12), 528–533. http://dx.doi.org/10.1016/j.tics.2004.10.001.
Lieberman, M., 2010. Social Cognitive Neuroscience, In: Fiske, S.T., Gilbert, D.T., Lindzey, G. (Eds.), 5th ed. McGraw-Hill, New York, pp. 143–193.
Mar, R., 2011. The neural bases of social cognition and story comprehension. Annu. Rev. Psychol. 62, 103–134. http://dx.doi.org/10.1146/annurev-psych-120709-145406.
Nichols, T., Brett, M., Andersson, J., Wager, T., Poline, J., 2005. Valid conjunction inference with the minimum statistic. Neuroimage 25 (3), 653–660. http://dx.doi.org/10.1016/j.neuroimage.2004.12.005.
Premack, D., Woodruff, G., 1978. Does the chimpanzee have a theory of mind? Behav. Brain Sci. 1 (04), 515. http://dx.doi.org/10.1017/S0140525X00076512.
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. http://dx.doi.org/10.1016/S1053-8119(03)00230-1.
Saxe, R., Powell, L., 2006. It's the thought that counts: specific brain regions for one component of theory of mind. Psychol. Sci. 17 (8), 692–699. http://dx.doi.org/10.1111/j.1467-9280.2006.01768.x.
Saxe, R., Carey, S., Kanwisher, N., 2004. Understanding other minds: linking developmental psychology and functional neuroimaging. Annu. Rev. Psychol. 55, 87–124. http://dx.doi.org/10.1146/annurev.psych.55.090902.142044.
Schultz, R., Grelotti, D., Klin, A., Kleinman, J., Van Der Gaag, C., Marois, R., Skudlarski, P., 2003. The role of the fusiform face area in social cognition: implications for the pathobiology of autism. Philos. Trans. R. Soc. Lond. B Biol. Sci. 358 (1430), 415–427. http://dx.doi.org/10.1098/rstb.2002.1208.
Schurz, M., Radua, J., Aichhorn, M., Richlan, F., Perner, J., 2014. Fractionating theory of mind: a meta-analysis of functional brain imaging studies. Neurosci. Biobehav. Rev. 42C, 9–34. http://dx.doi.org/10.1016/j.neubiorev.2014.01.009.
Spunt, R., Lieberman, M., 2012a. An integrative model of the neural systems supporting the comprehension of observed emotional behavior. Neuroimage 59 (3), 3050–3059. http://dx.doi.org/10.1016/j.neuroimage.2011.10.005.
Spunt, R., Lieberman, M., 2012b. Dissociating modality-specific and supramodal neural systems for action understanding. J. Neurosci. 32 (10), 3575–3583. http://dx.doi.org/10.1523/JNEUROSCI.5715-11.2012.
Spunt, R., Lieberman, M., 2013. The busy social brain: evidence for automaticity and control in the neural systems supporting social cognition and action understanding. Psychol. Sci. 24 (1), 80–86. http://dx.doi.org/10.1177/0956797612450884.
Spunt, R., Falk, E., Lieberman, M., 2010. Dissociable neural systems support retrieval of how and why action knowledge. Psychol. Sci. 21 (11), 1593–1598. http://dx.doi.org/10.1177/0956797610386618.
Spunt, R., Satpute, A., Lieberman, M., 2011. Identifying the what, why, and how of an observed action: an fMRI study of mentalizing and mechanizing during action observation. J. Cogn. Neurosci. 23 (1), 63–74. http://dx.doi.org/10.1162/jocn.2010.21446.
Van Overwalle, F., Baetens, K., 2009. Understanding others' actions and goals by mirror and mentalizing systems: a meta-analysis. Neuroimage 48 (3), 564–584. http://dx.doi.org/10.1016/j.neuroimage.2009.06.009.
Van Overwalle, F., Baetens, K., Marien, P., Vandekerckhove, M., 2013. Social cognition and the cerebellum: a meta-analysis of over 350 fMRI studies. Neuroimage 1–53. http://dx.doi.org/10.1016/j.neuroimage.2013.09.033.
Wellman, H., Cross, D., Watson, J., 2001. Meta-analysis of theory-of-mind development: the truth about false belief. Child Dev. 72 (3), 655–684. http://dx.doi.org/10.1111/1467-8624.00304.
Yarkoni, T., Poldrack, R., Nichols, T., Van Essen, D., Wager, T., 2011. Large-scale automated synthesis of human functional neuroimaging data. Nat. Methods 8 (8), 665–670. http://dx.doi.org/10.1038/nmeth.1635.