Behavior Research Methods (2025) 57:2
https://doi.org/10.3758/s13428-024-02519-9

ORIGINAL MANUSCRIPT

Establishing construct validity for dynamic measures of behavior using naturalistic study designs

Roberto C. French[1] · **Daniel P. Kennedy**[1] · **Anne C. Krendl**[1]

Accepted: 14 October 2024

Abstract
There has been a recent surge of naturalistic methodology to assess complex topics in psychology and neuroscience. Such methods are lauded for their increased ecological validity, aiming to bridge a gap between highly controlled experimental design and purely observational studies. However, these measures present challenges in establishing construct validity. One domain in which this has emerged is research on theory of mind: the ability to infer others' thoughts and emotions. Traditional measures utilize rigid methodology which suffer from ceiling effects and may fail to fully capture how individuals engage theory of mind in everyday interactions. In the present study, we validate and test a novel approach utilizing a naturalistic task to assess theory of mind. Participants watched a mockumentary-style show while using a joystick to provide continuous, real-time theory of mind judgments. A baseline sample's ratings were used to establish a "ground truth" for the judgments. Ratings from separate young and older adult samples were compared against the ground truth to create similarity scores. This similarity score was compared against two independent tasks to assess construct validity: an explicit judgment performance-based paradigm, and a neuroimaging paradigm assessing response to a static measure of theory of mind. The similarity metric did not have ceiling effects and was significantly positively related to both the performance-based and neural measures. It also replicated age effects that other theory of mind measures demonstrate. Together, our multimodal approach provided convergent evidence that dynamic measures of behavior can yield robust and rigorous assessments of complex psychological processes.

Keywords Theory of mind · Naturalistic task · Dynamic and static stimuli · Neuroimaging

Introduction

In recent years, many domains of psychological research have begun to shift toward using naturalistic stimuli in order to better capture the dynamic and multimodal psychological processes that unfold in everyday life (Aliko et al., 2020; Dawel et al., 2021; Hamilton & Huth, 2020; Nastase et al., 2020; Serre et al., 2015; Sonkusare et al., 2019; Zhaoyang et al., 2018). However, these measures have limitations, notably the complexity of capturing desired behaviors in dynamic tasks and concerns about the potential construct validity of these measures (Krendl, Hugenberg, & Kennedy, 2023; Risko et al., 2012; Serre et al., 2015; Yeung, Apperly, & Devine, 2023). Additionally, these measures often capture explicit behaviors using static outcomes, such as self-report (Linas et al., 2016) and discrete decisions (Quesque & Rossetti, 2020; Yaremych & Persky, 2019), which may lose important nuances offered by the dynamic approach. One domain in which this approach has been particularly limiting is explicit theory of mind. Indeed, many current measures of explicit theory of mind, including those using naturalistic stimuli, yield ceiling effects (Krendl, Hugenberg, & Kennedy, 2023; Quesque & Rossetti, 2020; Yeung, Apperly, & Devine, 2023), which may limit their interpretability and utility. Moreover, an important constraint with the small number of dynamic theory of mind measures (Grainger et al., 2019; Krendl, Hugenberg, & Kennedy, 2023; Krendl et al., 2023a, 2023b) is that it can be challenging to establish their construct validity due to the complexity of the stimuli. Given the amorphous and multidimensional aspects of naturalistic and dynamic stimuli, assessing the underlying meaning of the responses is a difficult task.

✉ Roberto C. French
robfren@iu.edu

[1] Department of Psychological and Brain Sciences, Indiana University, 1101 E. 10th Street, Bloomington, IN 47405, USA

Published online: 04 December 2024

To best assess convergent validity of a novel measure, it requires similar previously validated construct measures to be related (Yeung, Apperly, & Devine, 2023), which can prove problematic for dynamic and naturalistic assessments of theory of mind. An alternate approach Yeung and colleagues describe is a "known-group validity" whereby known clinical groups that differ in the measure can be used to demonstrate the novel measure displays similar difference, i.e. older adults perform worse on certain theory of mind tasks than young adults. Such a difference may help provide "known-group validity." The current study sought to address these limitations and assess such aspects of validity and establish the rigor and utility of a novel and dynamic measure of explicit theory of mind.

Theory of mind, the ability to infer others' thoughts and emotions (Frith & Frith, 2005), has been widely studied by social psychologists, clinical psychologists, cognitive scientists, aging researchers, and developmental scientists (Brüne et al., 2007; Demichelis et al., 2020; Henry et al., 2013; Peterson et al., 2009). Deficits in theory of mind have been commonly observed in clinical populations whose hallmark feature include impaired social comprehension, such as autism spectrum disorder and schizophrenia; (Bora et al., 2009; Peterson et al., 2009). Healthy and pathological aging have also both been commonly associated with theory of mind deficits, potentially due to declining neural systems that support theory of mind (Demichelis et al., 2020; Henry et al., 2013). As theory of mind is a conceptually complex construct (Apperly, 2012), some traditional measures have struggled to capture known theory of mind deficits (e.g., Scheeren et al., 2013). Thus, recent work has shifted toward using naturalistic paradigms based on the assumption that dynamic stimuli may be more accurate because they capture the multimodal complexity of real-world social interactions (e.g., Byom & Mutlu, 2013; Dziobek et al., 2006; Grainger et al., 2019; Johansson Nolaker et al., 2018). Indeed, a recent study with older adults found that their performance on a dynamic theory of mind task predicted the structure and composition of their real-world social relationships (Krendl, Hugenberg, & Kennedy, 2023).

A limitation of standard measures of explicit theory of mind is that they often yield ceiling effects (Bora et al., 2009; Chung et al., 2014; Krendl et al., 2023a, 2023b; Turner & Felisberti, 2017), with neurotypical adult populations commonly performing with almost 100% accuracy. A recent review assessing theory of mind measures in adult populations (Yeung, Apperly, & Devine, 2023) found that approximately half of the measures assessed suffered from ceiling effects. Ceiling effects present a number of problems: First, they generally cause the sample distribution to deviate from a normal distribution, an implicit assumption in many of the most common statistical analyses (Garson, 2012), forcing the usage of nonparametric analyses, transformations, or a reasonable motive as to why parametric tests should be used in the face of assumption violations. Nonparametric analyses offer less statistical power than their counterparts, requiring higher sample sizes and, due to the nature of the analyses, show whether there is a difference, but not the magnitude of that difference (Altman & Bland, 2009). Secondly, because many populations do not show ceiling effects on explicit theory of mind judgments (e.g., individuals with autism spectrum disorder or schizophrenia, or older adults) (Bora et al., 2009; Chung et al., 2014; Henry et al., 2013; Turner & Felisberti, 2017), unlike their typical comparison group (neurotypical young adults), between-group comparisons may be confounded by these inherent task limitations. Thus, identifying tasks that capture dynamic explicit theory of mind judgments has two important benefits: first, it would capture important nuances in how individuals initiate and modify their theory of mind judgments; second, it would limit ceiling effects which obscure individual differences at the upper end of performance.

A challenge in using naturalistic designs to explicit theory of mind accuracy is that performance is generally assessed by having participants making discrete judgments based on the dynamic information they have observed (Quesque & Rossetti, 2020). By constraining participants to specific responses to measure accuracy, these measures may fail to capture important nuances in how perceivers form and adjust their theory of mind judgments. Simply put, in real-world interactions, theory of mind accuracy may be updated and refined over the course of the interaction as individuals become more familiar with their interaction partner. Indeed, prior work has shown that individuals rely more on autobiographical memory when engaging theory of mind with familiar (versus unfamiliar) others (Rabin & Rosenbaum, 2012), suggesting that prior information may influence theory of mind judgments. However, there are currently a number of measures for capturing dynamic explicit theory of mind, though the construct validity of these measures has been shown to be hard to capture (Grainger et al., 2019; Krendl, Hugenberg, & Kennedy, 2023; Krendl et al., 2023a, 2023b). Recent work by Yeung, Apperly and Devine (2023) aims to address various theory of mind measures and on what fronts they demonstrate convergent validity with other similar measures and within clinical and nonclinical groups. We investigate methods of establishing such validity here.

Currently, one of the closest proxies to measuring explicit theory of mind without ceiling effects is through neuroimaging, notably functional magnetic resonance imaging (fMRI). Because fMRI experiments measure changes in the blood-oxygen-level-dependent (BOLD) signal in the brain in response to specific stimuli (Raichle, 1998), the unit of measurement in these studies is relatively arbitrary but is not prone to ceiling effects (Logothetis, 2002). Moreover, the BOLD signal has high variability (Garrett et al., 2010),

and has been widely studied in the context of theory of mind (Schurz et al., 2014, 2021). fMRI has also been widely used to compare explicit theory of mind performance across populations, including young and older adults (Hughes et al., 2019, 2020; Moran et al., 2012). Thus, the neural regions underlying theory of mind have been well characterized (Schurz et al., 2014, 2021), which provides a strong scaffolding for the current work. BOLD differences in such brain regions have been often used to characterize difference in theory of mind ability between groups, but within-group BOLD variability and its relation to theory of mind performance has also been reported in prior work (Kanske et al., 2015; Udochi et al., 2022). Kanske and colleagues (2015) report that performance in their EmpaTOM task related positively to neural activity in a network of regions associated with theory of mind. Similarly, Udochi and colleagues (2022) find that during a theory of mind task, neural activity in regions of the default and frontoparietal network positively predicted social cognitive ability. In previous work (Cassidy et al., 2021), we found that the activity within the medial prefrontal cortex during a person perception task was positively related to performance in an out-of-scanner Reading the Mind in the Eyes task (Baron-Cohen et al., 2001). However, as neuroimaging research is time-intensive and costly, it requires heavy investment to perform theory of mind research, inevitably leading to smaller sample sizes, while purely behavioral methods allow for greater recruitment capabilities.

In the current study, we thus examined the construct validity, rigor, and utility of a novel dynamic measure of explicit theory of mind—using joystick responses to collect participants' real-time theory of mind assessments during a movie-watching task. An important benefit of the measure is that it is intuitive, flexible, and relatively impervious to ceiling effects. Moreover, it is well suited to measure a wide range of psychological processes, making it a promising tool for measuring dynamic behavior. Our key goals were to ensure that the measure primarily did not have ceiling effects and yielded a normal distribution of responses, replicated well-established group differences in theory of mind, and had construct validity (e.g., was associated with performance on standard theory of mind tasks). To do this, we asked participants to complete continuous joystick ratings of awkwardness while watching a mockumentary-style television show.

Recognition and understanding of awkwardness in social interaction has been used in previous theory of mind assessments (Heavey et al., 2000; Pantelis et al., 2015). Conceptually, social awkwardness has some overlap with social faux pas because it involves understanding and interpreting violations of a social norm pas (Baron-Cohen et al., 1999; Stone et al., 1998). However, identifying social awkwardness can be complex and engages multiple aspects of theory of mind, including belief inferences, emotion recognition, and social gaffes detection (for similar conceptualization, see Heavey et al., 2000; Pantelis et al., 2015). Thus, in contrast to traditional measures of theory of mind (e.g., reading the mind in the eyes; the false belief task) (Baron-Cohen et al., 2001; Saxe & Kanwisher, 2003; Zaitchik, 1990) that only assess unitary aspects of theory of mind (e.g., understanding emotions or inferring beliefs), social awkwardness judgments require individuals to continuously integrate multiple types of theory of mind judgments. As such, they may better capture the complexity of everyday interactions than unimodal measures. Consistent with this reasoning, prior work has found that social awkwardness judgments are sensitive in detecting everyday theory of mind failures (Gedek et al., 2018; Heavey et al., 2000), and engage activation in a network of brain regions associated with theory of mind (Pantelis et al., 2015). Together, these findings suggest that awkwardness judgments may be a relatively comprehensive and sensitive measure of theory of mind.

Prior work has utilized consensus-based approaches to quantify innately subjective social psychological stimuli (Moran et al., 2004). We leveraged a similar approach to determine individual accuracy on our task. We then compared participants' relative accuracy on this novel task to their performance on a more traditional explicit theory of mind measure. Specifically, participants saw a different episode of the same mockumentary show, but explicit theory of mind accuracy was measured through questions that engaged theory of mind, similar to traditional explicit theory of mind approaches (Quesque & Rossetti, 2020). By using similar stimuli for both tasks, we could then compare accuracy in their joystick ratings to their performance on the traditional question-and-answer measure. We predicted that the joystick ratings would not show ceiling effects and be normally distributed, unlike the more traditional question-and-answer assessment of explicit theory of mind (Hypothesis 1A). In Hypothesis 1B, we predicted that performance across the two tasks would be related, demonstrating construct validity.

Second, we predicted that the joystick ratings would show similar construct validity and greater heterogeneity for groups that have been widely shown to have theory of mind deficits. We focused specifically on older adults, as extensive work has shown that they underperform on standard theory of mind tasks relative to young adults (e.g., Henry et al., 2013), and recent work has extended these age deficits to dynamic theory of mind tasks (Grainger et al., 2019; Krendl, Hugenberg, & Kennedy, 2023; Krendl et al., 2023a, 2023b). Hypothesis 2 thus predicted that older adults' joystick ratings would be less accurate (relative to a baseline comparison) than young adults (replicating well-established age deficits), but would still relate to performance on a standard, explicit theory of mind task, suggesting it has construct validity across samples.

Finally, we compared young adults' accuracy in their continuous joystick ratings to their performance on independent, well-validated measure of explicit theory of mind that did not have ceiling effects. Specifically, young adults completed a standard explicit theory of mind task, the false belief task (Saxe & Kanwisher, 2003; Zaitchik, 1990), while undergoing fMRI, and then completed the joystick ratings task outside of the scanner. As similar research designs have done (Pantelis et al., 2015), we then examined whether their joystick ratings accuracy was associated with the magnitude of activation in brain regions that have been widely implicated in theory of mind: the right temporo-parietal junction (rTPJ), posterior cingulate cortex (PCC), and medial prefrontal cortex (mPFC) (Schurz et al., 2014, 2021). We predicted that accuracy in their joystick ratings would be positively associated with the magnitude of their neural response to an independent theory of mind task (Hypothesis 3). Such a finding would demonstrate construct validity by verifying that their joystick rating accuracy was associated with neural activity associated with an independent, but well validated, measure of theory of mind. By leveraging a multimodal approach and replicating our findings across multiple samples, the current study is poised to provide convergent evidence that dynamic measures of behavior can yield robust and rigorous assessments of complex psychological processes.

Methods

Demographics

There were three groups of participants in the current study: a baseline sample, a young adult test sample, and an older adult test sample. All three groups completed a novel, dynamic theory of mind rating task. The young adult and older adult test samples also completed more standard assessments of theory of mind, and the young adult test sample further completed an fMRI study (to test Hypothesis 3). Participants completed additional measures not reported here as part of a larger study, but the order of the variables of interest in the current study was consistent across samples.

A preliminary power analysis was performed in G*Power (Faul et al., 2007), using a small effect size ($f^2 = 0.15$), three predictors, and $a = 0.05$, for a regression analysis, revealing that a target of $N = 77$ was required to achieve 80% power. Given a three-predictor regression model was the most complex statistical analysis performed, it was used here as a conservative estimate. The baseline sample was used to establish a consensus rating for the novel, dynamic task (discussed further below). This group consisted of 110 young adult ($M_{Age} = 18.7$ years, $SD = 0.95$) undergraduates at Indiana University who participated in exchange for partial course credit. The young adult test sample consisted of 114 different young adults ($M_{Age} = 21.9$ years, $SD = 4.0$) who were also undergraduates at Indiana University; they received monetary compensation for participating. Finally, the older adult test group consisted of 101 individuals over the age of 65 who were recruited from the Bloomington, Indiana community ($M_{Age} = 73.6$ years, $SD = 6.48$); they received monetary compensation. Older adults passed a six-item screener to ensure they were cognitively normal (Callahan et al. 2002). The studies were all approved by the Institutional Review Board at Indiana University. See Table 1 for full demographic information and description of the three samples' protocols.

Table 1 Full demographic information for the baseline sample, young adult sample, and older adult sample. Protocols for each sample are also briefly described here

		Baseline	Young adult test	Older adult test
N		110	114	101
Age (years)	Range	18–22	18–35	61–91
	Mean (*SD*)	18.7 (0.94)	21.9 (4.0)	73.6 (6.48)
Gender	Male (%)	34 (30.9%)	40 (35.1%)	49 (48%)
	Female (%)	75 (68.2%)	70 (61.4%)	53 (52%)
	Non-binary (%)	1 (0.9%)	4 (3.5%)	0 (0%)
Race	Asian (%)	17 (15.5%)	23 (20.2%)	2 (2%)
	African American (%)	5 (4.5%)	1 (0.9%)	0 (0%)
	White (%)	82 (74.5%)	80 (70.2%)	99 (98%)
	More than one (%)	4 (3.6%)	6 (5.3%)	0 (0%)
	Unknown (%)	2 (1.8%)	4 (3.5%)	0 (0%)
Protocol		**Novel, dynamic task**	**Novel, dynamic task, Static-response task, False belief task (fMRI)**	**Novel, dynamic task, Static-response task,**

Measures

Continuous theory of mind rating task

For the novel, dynamic theory of mind rating task, participants watched Season 2, Episode 2 ("Souvenir Shop") and Season 1, Episode 2 ("Petting Zoo") of the US mockumentary-style television show *Nathan for You*®.[1] The videos were each approximately 8 min long. While watching each task, the participants continuously rated the show's awkwardness using a Logitech Extreme 3D Pro joystick. Participants were instructed to move the joystick forward proportionally to how awkward they thought the video was throughout the entire viewing. This yielded a continuous, dynamic, and individualized rating of how awkward each moment in the video was perceived to be. We asked participants to rate awkwardness as a proxy measurement for theory of mind, similar to previous work (Pantelis et al., 2015). Recognition of socially awkward moments is a core feature of theory of mind (Baron-Cohen et al., 1999; Stone et al., 1998). Because the joystick is highly sensitive to individual movements and captures of range of measurements in real time, it provided an ideal continuous measure. MATLAB (version R2020a) utilizing the Psychophysics Toolbox Version 3 (Brainard & Vision, 1997; Kleiner, Brainard, & Pelli, 2007) was used to display the videos and record the movement of the joystick. Ratings were sampled at approximately 30 Hz (which mirrored the frame rate of the video presentation). The final number of samples across both videos were minimally variable between subjects ($M = 28{,}976.06$ samples, $SD = 20.01$).

Prior to watching the two videos, participants were presented with a calibration task in which they viewed a matrix of grayscale tiles that slightly varied in luminance from each other. The matrix of tiles gradually cycled through an increase and decrease of luminance over the course of 2 min in a predetermined manner. Participants were instructed to move the joystick forward in accordance with the overall brightness of the full matrix of tiles to calibrate themselves to the movement and range of the joystick. Participants were allowed to take as many attempts at the practice task as they needed to acclimate to the joystick.

Static response theory of mind task

To measure theory of mind performance on a more standard task, participants viewed and responded to questions about a different *Nathan for You* episode ("The Antique Shop"; Season 3, Episode 3). The task is described in full in Krendl, Hugenberg, & Kennedy, 2023. Briefly, the episode was cut to follow the key plotline about increasing sales for an antique shop. The approximate 6.5 min of the episode was segmented into 18 short clips (15–45 s, $M = 23$ s, $SD = 4$). Clips were truncated to roughly correspond to a scene or situation that would maximally engage theory of mind. For example, one clip focused on the set up for Nathan's plan to improve sales for the shop, whereas the following clip focused on the reveal of his plan and the shop owner's reaction. This approach allowed us to determine whether viewers could anticipate the plan in the first clip then evaluate the owner's beliefs about the plan in the second.

Clips were shown in sequential order, but between each clip, participants responded to 2–6 questions that related to classic facets of theory of mind: belief inference, detecting deception, understanding emotion, inferring motivation, and detecting social faux pas. For each question, participants were shown one correct answer and two foils. Response time was unlimited. There were a total of 63 questions, which included several control questions unrelated to theory of mind. Performance was measured as a proportion correct for the theory of mind questions only. It is described here as a "static response" task because, although the paradigm is dynamic, the behavioral outcome consists of static responses. An important benefit of this approach is that the theory of mind judgments being compared ("dynamic" and "static") are based on the same stimulus, but simply measured in different ways.

fMRI false belief task

To assess the robustness of the continuous theory of mind rating task (e.g., the joystick ratings), we next evaluated whether performance on this task was associated with the level of neural activity engaged during an independent, traditional theory of mind task: the false belief task (Saxe & Kanwisher, 2003; Zaitchik, 1990). The false belief task is a common theory of mind assessment that has been widely used in neuroimaging research on theory of mind (Hughes et al., 2019; Schurz et al., 2014, 2021). The task consists of 24 short vignettes describing a situation that details a character's beliefs (theory of mind condition), i.e., "When Lisa left Jacob, he was deep asleep on the beach. A few minutes later, a wave woke him. Seeing Lisa was gone, Jacob decided to go swimming," followed by a true or false question that required participants to infer the mental state of one of the characters: "Lisa now believes that Jacob is sleeping." In addition to the 12 theory of mind vignettes, there were also 12 control vignettes.

[1] Videos were shown in a random counterbalanced order for the baseline sample but were consistently shown in the order of "Petting Zoo" followed by "Souvenir Shop" for both test samples.

Control vignettes described a physical representation of a situation, i.e., "When the picture was taken of the house, it was one story tall. Since then, the renovators added an additional story and a garage," followed by a true or false question that measured understanding: "In the picture, the house is two stories tall and has a garage." The vignettes were presented across two functional runs (presented in counterbalanced order), each lasting approximately 5.5 min. Six unique false belief and six unique control vignettes were presented within a single run in pseudorandomized order. For each trial, the vignette was presented for 10 s. This was followed by an interstimulus interval of 0–6 s, then a true or false question which remained on the screen for 6 s. Participants indicated their response via keypress during this window. The vignettes were separated by an intertrial interval of 4–10 s. Participants completed a practice trial before the task began to ensure they could read and respond in the allotted timeframes. Four participants did not complete the fMRI portion of the study and so were excluded from the resulting fMRI analyses ($N = 110$).

Neuroimaging was performed with a 20-channel head/neck coil on a Siemens 3.0 T Prisma MRI scanner at the Indiana University Imaging Research Facility in Bloomington, Indiana. The task stimuli were presented using a projector illuminating a screen that was visible to participants through a mirror attached to the head coil. The false belief task was presented using E-Prime 3 through a Dell laptop running Windows 10. Anatomical scans were acquired with a high-resolution three-dimensional (3D) magnetization-prepared rapid gradient-echo sequence (sagittal rotation; 160 slices, echo time [TE] = 2.7 ms, repetition time [TR] = 1800 ms, inversion time [TI] = 900 ms, flip angle = 9°, 1.0 mm isotropic voxels; with no fat suppression). Functional scans were collected using an echo-planar image (EPI) sequence sensitive to blood oxygen level-dependent contrast (T2*; 54 slices with 2.2 mm thickness and no gap, TE = 30 ms, TR = 2000 ms, flip angle = 70°, 2.2 mm isotropic voxels, field of view [FOV] = 242 × 211.2 mm, in-plane matrix size = 110 × 96, A/P phase encoding direction). Slices were collected in an interleaved order (multi-band acceleration factor = 2). Preprocessing and analyses of functional data were conducted in SPM12 (Wellcome Trust Centre for Neuroimaging, London, UK). Images were realigned to correct for motion, normalized to the Montreal Neurological Institute (MNI) template, and smoothed using an 8 mm full width at half maximum (FWHM) isotropic Gaussian kernel.

We used a general linear model incorporating the two vignette types (theory of mind, control) and covariates of no interest (a session mean, a linear trend, and six movement parameters derived from realignment corrections). We computed parameter estimates (β) and t-contrast images (containing weighted parameter estimates) for each comparison at each voxel and for each participant. Regions of interest (ROI) were acquired from a Neurosynth meta-analytic search of the term "mentalizing" which included 151 studies (Yarkoni et al., 2011). Key benefits to using Neurosynth to define the locations of the ROIs associated with mentalizing is that it is more robust than identifying the regions from a single task, and it improves the overall rigor of this analysis by identifying the regions in an independent manner.

We identified clusters of interest from Neurosynth within the PCC, rTPJ, dorsomedial prefrontal cortex (dmPFC), and ventromedial prefrontal cortex (vmPFC) (Fig. 3) by isolating the peak z-score voxel location for each of the four regions: PCC (−2, −56, 36; $z = 10.54$), rTPJ (58, −56, 26; $z = 7.21$), dmPFC (6, 56, 20; $z = 11.6$), vmPFC (−4, 48, −18; $z = 10.06$). Then, for each participant, we extracted the mean activation in each cluster when participants completed the false belief task. For each ROI analysis, we extracted the average parameter estimates from a 6-mm sphere surrounding the peak coordinate of interest using the false belief condition versus baseline (fixation) contrast, and control condition versus baseline (fixation) contrast. For simplicity, we created a difference score for theory of mind minus control, so that positive values indicated greater activation in the theory of mind versus control condition. The resulting values were then input into independent regression models with the awkwardness similarity metric predicting the neural activity to determine whether they were related. Additional regressions were performed using the reaction time on correct responses to see how they related to the neural activity for each brain region.

Establishing consensus ratings and similarity metrics

An important challenge in using dynamic measures of behavior in response to naturalistic stimuli is that these measures do not necessarily have correct or incorrect responses, making it difficult to determine whether respondents' ratings are "accurate." Consensus approaches in which independent participants act as baseline can be useful in movie-watching paradigms to record and quantify subjective stimuli. For example, one study used an independent participant sample to identify humorous moments of a video by generating a consensus laugh track (Moran et al., 2004). As such, Hypothesis 1 and 2 measured the extent to which the young and older adult samples were similar to the consensus ratings generated from the baseline sample as the dynamic measure of explicit theory of mind. Though we anticipated that there would be some variability in the baseline samples' ratings, we expected that they would generally agree

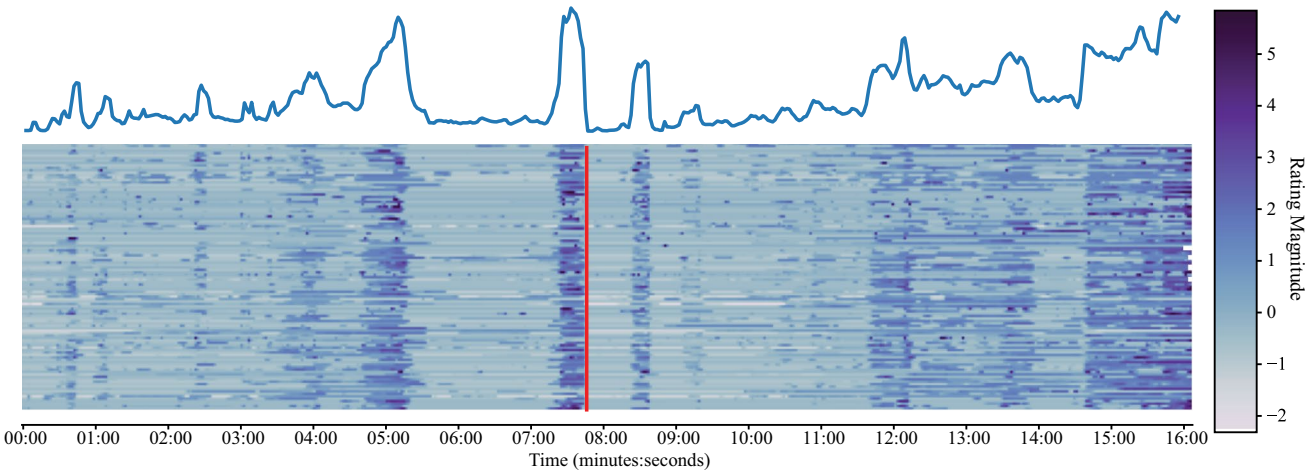

Fig. 1 Average awkwardness rating of the baseline sample plotted as continuous line (above); individual rating of all baseline subjects plotted below. Red line depicts separation between two episodes

on when moments of high or low awkwardness emerged in the dynamic video tasks, which would reflect some underlying consensus.[2]

We first created the consensus rating using the baseline sample's joystick ratings. To do this, for each participant we concatenated the awkwardness ratings from the novel, dynamic task to form a continuous rating time series that spanned the length of both videos. This rating timeseries was *z*-scored within individual to ensure all individual ratings were mapped on a similar scale. As seen in Fig. 1, the ratings were consistent across much of the time series, with the same time frames eliciting similar peaks and troughs across the individual ratings. Simply put, respondents generally agreed on when moments of relatively high or low awkwardness emerged in the video, though the magnitude to which those moments were seen as being awkward varied across participants. We tested the between-subject reliability of the baseline sample ratings by performing a leave-one-out cross-validation between each individual and the mean of the remaining individuals in the baseline sample, which revealed an average correlation of $r = 0.58$, suggesting that respondents had similar response patterns but some variability in how they engaged in the task. We then averaged the 110 baseline sample ratings together to capture the overall consensus of the sample to use as the comparison point for Hypothesis 1B.

[2] As utilizing a solely young adult consensus as a baseline comparison may introduce bias when comparing against age groups, an independent secondary older adult sample ($N = 110$, mean age = 74.5 years, $SD = 6.8$ years) was collected to act as an older adult consensus baseline. All tests were additionally performed against this secondary baseline, and all results remain consistent regardless of baseline, demonstrating robustness of findings.

Importantly, as previously described, the video order was counterbalanced for the baseline subjects only. To assess whether the counterbalance video order influenced rating, the mean time series for the counterbalanced condition was compared with the alternate counterbalance condition, resulting in a very high correlation ($r = 0.953$), demonstrating the video order counterbalance had a minimal effect on rating. Due to this result, we opted to leave the concatenated time series ratings in the order consistent with the video orders shown to the two test samples: "Petting Zoo" followed by "Souvenir Shop."

We used the baseline sample's average consensus as an anchor point against which all the test subjects were compared. This was done by calculating an *awkwardness similarity metric* between the baseline consensus and each participant within the young adult and older adult test samples. To calculate this awkwardness similarity metric, we created individual concatenated and standardized rating time series across both videos for each participant in the test samples in the same way as was done for the baseline sample. Each individual test participant's resulting time series was then compared against the baseline consensus via Pearson's correlation. Here, a higher correlation reflects a greater similarity between baseline and test sample participant. This correlation was Fisher *r*-to-z-transformed to conform the correlations to a normal distribution for future use in statistical models (Fisher, 1915), finally resulting in a value, hereon referred to as the "awkwardness similarity metric," for each participant in the young and older adult test samples. A split-half correlation was performed to assess the reliability of the measure within subject, when corrected via the Spearman–Brown formula (Brown, 1910; Spearman, 1910) to better reflect the full length. This resulted in a correlation of 0.703 after correction.

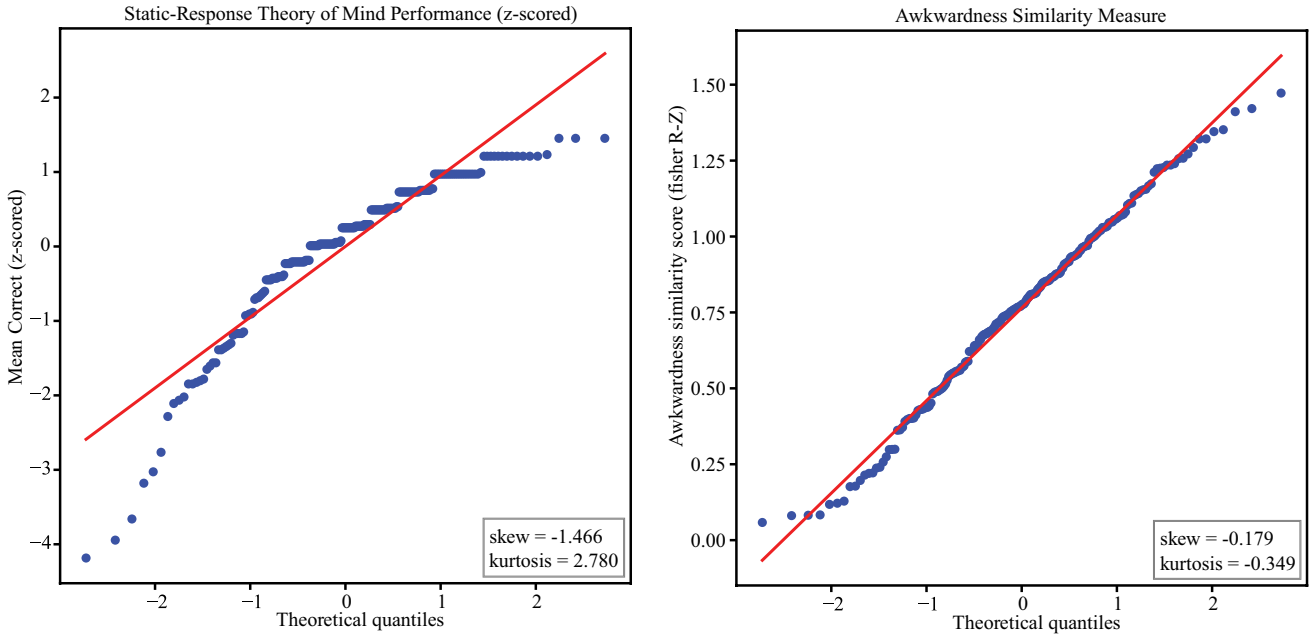

Fig. 2 Probability plots demonstrating theoretical normal distribution plotted as a red line and the true distribution plotted as the blue scatter of the static-response theory of mind task performance (left) and the awkwardness similarity metric (right)

Results

Hypothesis 1A: The measure is robust against ceiling effects, i.e., normally distributed; 1B: Young adults' awkwardness similarity metric relates to performance on the static response theory of mind task

Theory of mind performance on the static-response task was measured as proportion correct for the theory of mind questions. Consistent with prior work (Krendl, Hugenberg, & Kennedy, 2023), accuracy of the young adult sample on this task was high ($M_{score} = 90.3\%$; $SD = 6.8\%$). Also consistent with prior work (Bora et al., 2009; Chung et al., 2014; Turner & Felisberti, 2017), performance was significantly non-normally distributed as determined by a Shapiro–Wilk test, (Shapiro & Wilk, 1965) $W(214) = 0.89$, $p < 0.001$, and demonstrated high negative skew (− 1.399). However, supporting Hypothesis 1A, performance on the awkwardness similarity metric was normally distributed, $W(214) = 0.99$, $p = 0.16$, with minimal skew (skew = − 0.182). The lack of negative skew alongside the normal distribution demonstrates there were no ceiling effects. Probability plots demonstrating real and theoretical distributions of both measures can be viewed in Fig. 2. To test Hypothesis 1B, we constructed an ordinary least-squares regression model to assess the relationship between the awkwardness similarity metric for the young adult test sample and their performance on the static-response theory of mind task. The awkwardness similarity metric significantly predicted performance on the static response theory of mind task, $F(1, 112) = 6.339$, $p = 0.013$; $R^2 = 0.054$, in the positive direction ($\beta = 0.065$). The result thus provides support that the awkwardness similarity metric captures theory of mind.

Hypothesis 2: Young adults have a higher awkwardness similarity metric than older adults

Consistent with prior work (Krendl et al., 2023a, 2023b), older adults performed significantly worse on the static-response theory of mind task ($M = 84.3\%$, $SD = 9.3$) than young adults ($M = 90.3\%$ $SD = 6.8$; $t(214) = 5.4$, $p < 0.001$). The awkwardness similarity metric replicated this pattern of group differences in the same direction, with older adults scoring significantly lower ($M = 0.617$, $SD = 0.296$) than young adults; ($M = 0.897$, $SD = 0.242$; $t(214) = 7.6$, $p < 0.001$); see Fig. 3 for regression plot.

We next assessed the effects of age in predicting the static-response theory of mind task performance above that of the similarity metric by using a hierarchical regression. Here, the awkwardness similarity predicting the performance on the static response task was the first step, the second step added the inclusion of an age term (dummy coded as young adults = 1 and older adults = 0), and finally an age-by-similarity metric interaction term was additionally entered in the third step. The first step of the model was significant, $F(1, 213) = 27.11$, $p < 0.001$, accounting for

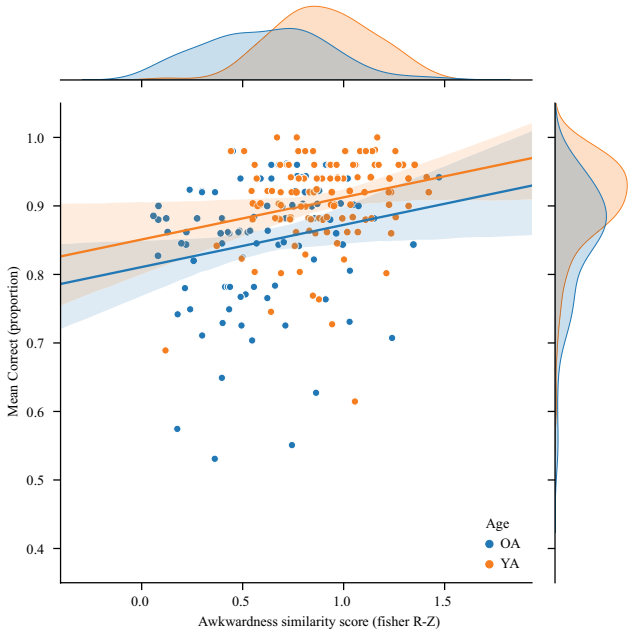

Fig. 3 Scatter plot with regression fit lines demonstrating relationship between awkwardness similarity metric and mean performance on the static-response theory of mind task, relationship plotted across both older and young adult groups. Marginal plots show distributions of dependent and independent variables across groups

11.3% of the overall variance. The second step accounted for 15.8% of the overall variance and was a significant model improvement over the first step ($\Delta F = 11.39$, $\Delta R^2 = 0.045$, $p < 0.001$). The addition of the interaction term in the second step did not significantly improve model fit ($\Delta F = 1.0 \times 10^{-6}$, $\Delta R^2 = 4.8 \times 10^{-9}$, $p = 0.99$), suggesting that the main effect was consistent across age groups. See Table 2 for full model results.

Hypothesis 3: Neural activation associated with independent theory of mind task is related to the awkwardness similarity metric

Hypotheses 1 and 2 explored the construct validity of the awkwardness similarity metric by comparing it to performance on a related task with static, rather than continuous, responses. In Hypothesis 3, we evaluated the robustness of the similarity metric by comparing it to an independent, but well-validated, measure of the theory of mind: the false belief task. Moreover, we examined whether neural activation in four brain regions (defined a priori) that have been widely implicated in theory of mind was positively associated with the similarity metric. In addition to demonstrating the robustness of the similarity metric by comparing it to an independent task, a benefit of using neural activation as one of the variables of interest is that it does not have the same ceiling effects as many explicit measures of theory of mind, including the static-response task.

Our analyses focused on four brain regions that have been commonly implicated in theory of mind (Schurz et al., 2014, 2021): the right temporoparietal junction, the dorsomedial prefrontal cortex, ventromedial prefrontal cortex, and the posterior cingulate cortex (Fig. 4). We defined peaks for each of these regions independently using Neurosynth to ensure rigor and robustness (see Methods), and further confirmed that these regions were

Table 2 Full regression model results between awkwardness similarity metric and validation metrics across hypotheses 1–3

	Predictor	$R^2/\Delta R^2$	F/ΔF	β	*t*	*P*	
Hypothesis 1	Awk. Similarity	0.054	6.339	0.065	2.52	0.013	*
Hypothesis 2	Step 1	0.113	27.11			< 0.001	***
	Awk. Similarity			0.092	5.21	< 0.001	***
	Step 2	0.158/0.045	19.94/11.39			< 0.001	***
	Awk. Similarity			0.062	3.14	0.002	**
	Age			0.040	3.38	0.001	**
	Step 3	$0.158/4.8 \times 10^{-9}$	$13.23/1.0 \times 10^{-6}$			< 0.001	***†
	Awk. Similarity			0.062	2.37	0.019	*
	Age			0.040	1.22	0.224	
	Interaction			-4.3×10^{-5}	−0.001	0.999	
Hypothesis 3	Brain Region						
	PCC	0.048	5.43	2.33	2.33	0.022	*
	rTPJ	0.049	5.56	2.38	2.36	0.02	*
	vmPFC	0.059	6.71	1.59	2.61	0.011	*
	dmPFC	0.005	0.55	0.58	0.74	0.458	

In Hypothesis 2, Age is bivariate- dummy coded as YA = 1, OA = 0. In Hypothesis 3, significance testing underwent FDR correction for multiple comparisons, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, † nonsignificant model fit over previous step

engaged during the false belief task using the theory of mind vs. control contrast from this task. There was significantly greater activation for the theory of mind versus control task in all four regions (Fig. 4), which is consistent with prior work showing that these brain regions are involved in mentalizing (Schurz et al., 2014, 2021).

We tested Hypothesis 3 by conducting four separate regressions, one for each of the four brain regions of interest. Each regression examined whether the similarity metric predicted greater activation in that region during the theory of mind relative to control tasks. In three regions, the relationship was significantly positively related: PCC: $F(1, 108) = 5.43$, $p = 0.022$, $R^2 = 0.048$; rTPJ: $F(1, 108) = 5.56$, $p = 0.02$, $R^2 = 0.049$; vmPFC: $F(1, 108) = 6.79$, $p = 0.011$, $R^2 = 0.059$. However, in the dmPFC, there was no significant relationship: $F(1, 108) = 0.55$, $p = 0.458$, $R^2 = 0.005$. To correct for the multiple comparisons, false discovery rate (FDR) correction was performed. The significant findings of PCC, rTPJ, and vmPFC hold through FDR correction. Gignac and Szodorai (2016) provide an empirically driven approach in redefining effect size cutoffs based on a corpus of meta-analytically derived correlations. Based on their approach, a correlation of 0.19 is representative of the median, and is described as a medium effect size. Of our results presented, the lowest significant R^2 described was 0.048; given all but the hierarchical regressions performed for hypothesis 2 were single independent variable regression models, this mathematically would equate to a correlation of 0.219, which is above the medium effect size proposed by Gignac and Szodorai (2016). See Table 2 for all model results. Full pairwise intercorrelation between all variables of interest can be seen in Table 3.

Discussion

The goal of the current study was to assess a novel, dynamic measure of explicit theory of mind that would be well suited to naturalistic study designs. Across three studies using different tasks and populations, we found that the novel awkwardness similarity metric was robust against the ceiling effects that are common in such assessments and had construct validity (Hypothesis 1). This effect persisted for a sample that has been widely shown to be impaired on theory of mind tasks (older adults) and replicated prior work showing that older adults underperformed on explicit measures of theory of mind relative to young adults (e.g., Henry et al., 2013). Finally, performance on the similarity metric was also related to the magnitude of neural activity associated with completing an independent, well-validated measure of explicit theory of mind. Importantly, while all extracted neural activation from the four brain regions was highly intercorrelated (Table 3), it demonstrated a significant relationship with only the awkwardness similarity metric and not the static-response task, indicating that the awkwardness similarity score captures some additional complexity or variance that relates to the underlying neural activation that the static response task does not. Together, these results suggest that dynamic measures of explicit behavior are robust, variable, and do not have ceiling effects. Moreover, they may provide deeper insight into nuances in behavior that underlie complex psychological processes.

The similarity metric in the current study provides a novel way of measuring explicit theory of mind. Given the long form (approximately 15 min) nature of the video, important context relating to the social interactions between characters is revealed through the extent of the videos. Removing the temporal unfolding of the narrative may remove important social context that is naturally present in real-life social interaction. Similar long-form social interaction has been presented to older adults in prior work (Lecce et al., 2019). Lecce and colleagues utilized the Movie for the Assessment of Social Cognition (MASC) (Dziobek et al., 2006), finding that older adults performed worse on performance during the MASC compared to younger adults but finding no age difference in a static paradigm, highlighting a potential importance of a naturalistic, dynamic, and temporally unfurling paradigm. Notably, the MASC requires intermittent interruption, which may interrupt the flow of contextual social information, whereby our paradigm operates completely continuously on a moment-by-moment basis.

Participants in the baseline sample demonstrated consistency in which moments they viewed as being relatively high or low in awkwardness (see Fig. 1). Utilizing an independent consensus as a baseline helped preserve a level of nuance in the awkwardness judgments. A key aspect of the paradigm involves the moment-by-moment rating, and thus having a dynamic and continuous measure was an important benefit of the consensus approach. Subjective psychosocial stimuli have been seen to arrive at consensus in other work; Todorov (2008) found consensus ratings on face trustworthiness better predicted amygdala activation than individual ratings. Moran and colleagues (2004) found that consensus moments of humor predicted activity in the left inferior frontal and posterior temporal cortex. Utilizing an independent sample to create a consensus baseline can thus preserve the dynamics and nuance of the richly sampled stimuli.

The fact that theory of mind accuracy on two different explicit theory of mind tasks was related to the extent to which the participants awkwardness ratings were similar to the consensus ratings suggests that the similarity metric may capture important nuances in theory of mind judgments. These nuances are not fully explored in standard measures of

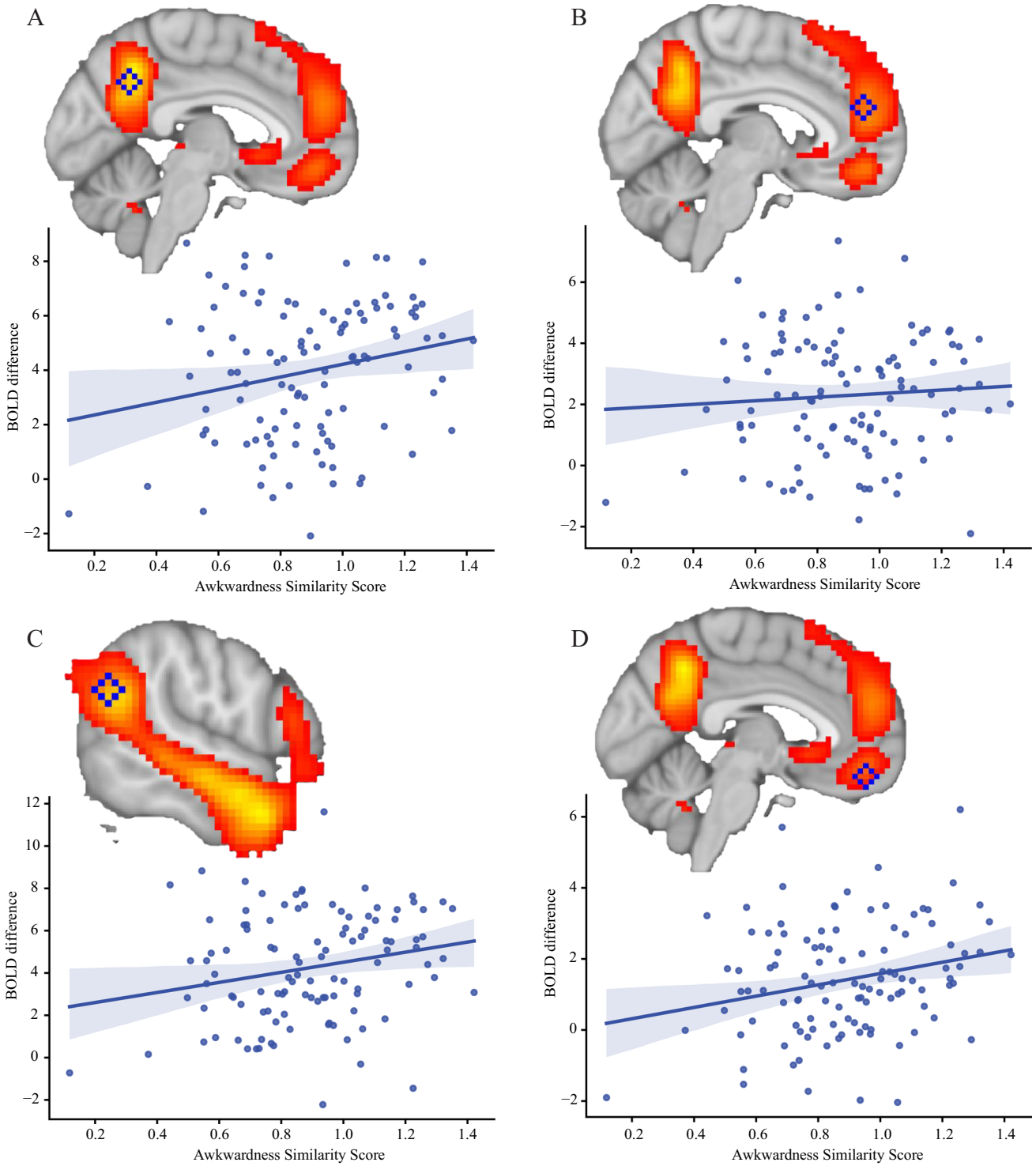

Fig. 4 False belief theory of mind vs. control contrast regions of task activation; blue highlights demonstrate the bounds of spheres used to extract BOLD activation within the a priori regions of interest, while orange depicts the significant activation in theory of mind trials > control trials in the false belief task, activation maps thresholded at FWE $p < 0.05$ with minimum 10 voxel cluster size correction. Extracted activation was compared against the awkwardness similarity metric as seen in the associated regression plots. **A** Posterior cingulate cortex, (−2, −56, 36). **B** Dorsomedial prefrontal cortex, (6, 56, 20). **C** Right temporoparietal junction, (58, −56, 26) **D** Ventromedial prefrontal cortex, (−4, 48, −18)

Table 3 Intercorrelation matrix of all tested variables of interest for young adult sample

	Static response	Awkwardness similarity	rTPJ	PCC	dmPFC	vmPFC
Static response	1	0.211*	0.165	0.166	0.114	0.084
Awkwardness similarity	-	1	0.221*	0.219*	0.071	0.243*
rTPJ	-	-	1	0.595**	0.649**	0.441**
PCC	-	-	-	1	0.605**	0.495**
dmPFC	-	-	-	-	1	0.43**
vmPFC	-	-	-	-	-	1

Brain region variables (rTPJ, PCC, dmPFC, vmPFC) are the theory of mind minus control contrasts localized within these a priori regions as described in methods. * $p < 0.05$, ** $p < 0.001$

explicit theory of mind and may provide important insights into how individuals engage and apply theory of mind in everyday life. Specifically, evaluations of theory of mind in clinical and older adult populations using standard static measures have yielded mixed results about the nature and magnitude of theory of mind deficits (Alkire et al., 2023; Grainger et al., 2019; Phillips et al., 2011). One reason for this might be that those measures only capture general theory of mind understanding, but do not capture dynamic processes that may be more sensitive to the complex underlying processes. Indeed, dynamic tasks have been shown to capture nuances in age deficits in theory of mind that static tasks do not (Cortes et al., 2021). Thus, since theory of mind is a complex construct (Apperly, 2012), measures that better capture this complexity might yield novel insights into potential group differences in theory of mind.

An additional benefit of the similarity metric is that it was relatively impervious to ceiling effects, suggesting that it may provide an alternative to traditional theory of mind assessments (e.g., the false belief task) that consistently show ceiling effects (Bora et al., 2009; Chung et al., 2014; Yeung, Apperly, & Devine, 2023). As ceiling effects may obscure group differences and limit the interpretability of results, tasks such as the similarity metric that minimize skew and are normally distributed may be more suitable for group comparisons in future work.

An important consideration in the current study is that the similarity metric was based on awkwardness judgments. We chose awkwardness judgments as they have been used as a proxy for theory of mind assessment (Heavey et al., 2000). Moreover, prior work has used a similar methodology to assess brain activation underlying theory of mind (Pantelis et al., 2015). Recognizing and understanding social awkwardness requires a complex application of theory of mind across both cognitive and affective dimensions (e.g., Heavey et al., 2000; Pantelis et al., 2015), and may better represent everyday ability (Gedek et al., 2018). However, our focus on awkwardness judgments may fail to capture all facets of theory of mind. Indeed, neuroimaging studies have shown that different types of theory of mind engage distinct neural mechanisms (Schurz et al., 2021), suggesting that single measures likely do not capture the full complexities of theory of mind (e.g., Apperly, 2012). An added benefit of the joystick approach is that it could be leveraged to capture multiple aspects of theory of mind (e.g., detecting deception). Future studies should explore the reliability of this approach and consider capturing other facets of theory of mind as these may assess other important nuances within naturalistic stimuli.

The fact that the similarity metric also observed age deficits, consistent with prior work (Demichelis et al., 2020; Henry et al., 2013), further bolsters these findings. Specifically, we found that older adults had lower awkwardness similarity metrics than young adults. They also underperformed on the static theory of mind task. Moreover, because there was no effect of the interaction between the metric and age group, it suggests that the metric performed similarly as a proxy for theory of mind across age groups. An important caveat to these findings is that older adults' similarity metrics were calculated by comparing their performance to young adults, which may have exacerbated the perceived age deficits. Please see footnote 2 in Methods for further exploration on this front.

Finally, the fact that the awkwardness similarity metric predicted the magnitude of neural response in several key brain regions associated with theory of mind on an independent task demonstrated the robustness of the current measure. BOLD variability and its relation to theory of mind ability has been shown to characterize between group performance difference in theory of mind ability but has also been demonstrated to be related to performance within a single group. Prior work has shown that neural activity in a constellation of regions associated with mentalizing is related positively to performance in theory of mind tasks both within the scanner (Kanske et al., 2015; Udochi et al., 2022) and demonstrating out-of-scanner ability in independent tasks (Cassidy et al., 2021). We would like to note, however, that though prior work has shown that increased neural activity

is associated with better theory of mind performance, the evidence for this relationship is relatively sparse within our paradigm, and future work should explore this. Additionally, we would like to note that the brain regions were selected through an agnostic meta-analytic tool, and the associated neural response was measured from an independent and well-validated theory of mind task (the false belief task), contributing to the overall rigor and robustness of these results. However, it is important to note that not all regions demonstrated a significant positive relationship with the metric. Specifically, the neural response in the dmPFC did not significantly predict the awkwardness similarity metric. There are two potential explanations for this. First, the role of the dmPFC in theory of mind is somewhat contentious; Otti et al. (2015) describe the mPFC as nonessential specifically to theory of mind processing and instead more domain-general and related to self-referential thinking. An alternative possibility is that activation in the discrete brain regions may reflect neural responses associated with only a specific aspect of theory of mind (e.g., understanding emotions, inferring beliefs). Indeed, recent meta-analyses have shown that different brain regions are engaged during different theory of mind tasks (Schurz et al., 2021), suggesting that focusing our metric on one dynamic aspect of theory of mind (e.g., detecting awkwardness) may not be reflected in all theory of mind brain regions. Related to this point, future work may expand beyond focusing on activation of singular brain regions to understand theory of mind, instead turning focus to networks of brain regions to better assess larger scale neural systems (e.g., Hughes et al., 2019).

There are several limitations in the current study that should be considered. Notably, while nearly all the models significantly supported our hypotheses that the awkwardness similarity measure predicted independent measures of theory of mind, the explained variance for these models was low. There are a few potential explanations for this. First, the low variance explained could have been due to limitations in the static response task. Specifically, the static task, as with many theory of mind tasks (Yeung, Apperly, & Devine, 2023), demonstrated ceiling effects, particularly on the questions that assessed cognitive theory of mind (e.g., inferring beliefs, understanding motivations, detecting deception). Ceiling effects ultimately reduce variance in the upper end of measurement and violate statistical assumptions of normality that parametric tests require. Second, prior work has shown modest to no relationships between different measures of theory of mind (Bottiroli et al., 2016; Warnell & Redcay, 2019). For example, Bottiroli and colleagues (2016) found a small within-subject correlation ($r = 0.19$) for performance on cognitive and affective theory of mind. Warnell and Redcay (2019) compared numerous traditional assessments of theory of mind in an adult population and found no significant correlation between any of the measures (all $r < 0.13$). One potential reason for the lack of relation between measures in these studies is that theory of mind may not be a unitary entity (Apperly, 2012; Schaafsma et al., 2015; Schurz et al., 2014, 2021), and each of these specific tasks taps into subdivisions that rely on different mechanisms. Fourth, the fact that our effects were replicated across three studies using multiple methods (behavioral, neuroimaging) also bolsters confidence in these findings. These seemingly small effects are above the median effect size found in a meta-analysis of research on psychological individual differences (Gignac & Szodorai, 2016). Finally, theory of mind is highly complex, and both the false belief task and static response task likely only assess slim facets of this larger construct. Consistent with this reasoning, social awkwardness judgments have been considered by some researchers to be more sensitive in detecting everyday theory of mind failures (Gedek et al., 2018; Heavey et al., 2000). The awkwardness similarity score may capture some overlapping, but distinct, aspects of theory of mind, thus resulting in the low variance explained. Future work could explore this possibility using different dynamic constructs (e.g., identifying deception rather than awkwardness) by relating these measures to real-world engagement of theory of mind, or perhaps assessing moment-by-moment changes in a neuroimaging paradigm as they relate to the momentary changes in behavior such as those performed in Masson and Isik (2021). Future expansion on such paradigms may help bridge socio-neurocognitive links. It should also be noted that the demographics of our samples are primarily White and predominantly female; previous work utilizing the same static-response task (Krendl, Hugenberg, & Kennedy, 2023) assessing theory of mind and a more diverse sample demonstrated race and gender differences when comparing in-person to online survey responses. Future studies assessing a more diverse population may help verify the robustness of our measure.

The current study demonstrates the utility and construct validity of a novel methodology that is designed to capture dynamic psychological processes during naturalistic tasks. The move to more naturalistic stimuli in psychology research is important as it provides important external context where strict experimental designs seek to reduce such context in lieu of control. Our measure demonstrates that such designs are feasible and even an improvement on aspects such as the removal of ceiling effects. This novel approach to a naturalistic task design demonstrates a method of capturing the complex psychological processes that are involved in the process we seek to better understand while promoting an experimental environment that better mirrors a real-world one.

Acknowledgements This work was supported by R01 AG075044 (PI: Krendl) from the National Institute on Aging and R01 MH110630 (PI:

Kennedy) from the National Institute of Mental Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We thank Lucas Hamilton, Amy Gourley, Samuel Rincón, Sarah Greenwell, Mia Freeman, and Carter Wittendorf for assistance with data collection.

Funding This work was supported by R01 AG075044 (PI: Krendl) from the National Institute on Aging and R01 MH110630 (PI: Kennedy) from the National Institute of Mental Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data availability Data and materials are available on the public OSF project: https://osf.io/5hkwr/

Code availability Code and methods are available on the author's GitHub: https://github.com/rcf004/similarity-rating

Declarations

Open Practices Statement Data are available at the author's OSF project: https://osf.io/5hkwr/. Code used for the experiments is available on the author's GitHub: https://github.com/rcf004/similarity-rating. None of the experiments were preregistered.

Ethics approval All experiments were approved by the Indiana University-Bloomington Institutional Review Board.

Consent to participate All participants completed an informed consent detailing the contents and aims of the study before participating.

Consent for publication All authors have consented to the final version of this manuscript.

Conflicts of interest None to report.

References

Aliko, S., Huang, J., Gheorghiu, F., Meliss, S., & Skipper, J. I. (2020). A naturalistic neuroimaging database for understanding the brain using ecological stimuli. *Scientific Data, 7*(1), 347

Alkire, D., McNaughton, K. A., Yarger, H. A., Shariq, D., & Redcay, E. (2023). Theory of mind in naturalistic conversations between autistic and typically developing children and adolescents. *Autism, 27*(2), 472–488

Altman, D. G., & Bland, J. M. (2009). Parametric v non-parametric methods for data analysis. *Bmj*, *338*

Apperly, I. A. (2012). What is "theory of mind"? Concepts, cognitive processes and individual differences. *Quarterly Journal of Experimental Psychology, 65*(5), 825–839

Baron-Cohen, S., O'riordan, M., Stone, V., Jones, R., & Plaisted, K. (1999). Recognition of faux pas by normally developing children and children with Asperger syndrome or high-functioning autism. *Journal of Autism and Developmental Disorders, 29*, 407–418

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the Mind in the Eyes" Test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *The Journal of Child Psychology and Psychiatry and Allied Disciplines, 42*(2), 241–251

Bora, E., Yucel, M., & Pantelis, C. (2009). Theory of mind impairment in schizophrenia: Meta-analysis. *Schizophrenia Research, 109*(1–3), 1–9

Bottiroli, S., Cavallini, E., Ceccato, I., Vecchi, T., & Lecce, S. (2016). Theory of Mind in aging: Comparing cognitive and affective components in the faux pas test. *Archives of Gerontology and Geriatrics, 62*, 152–162

Brainard, D. H., & Vision, S. (1997). The psychophysics toolbox. *Spatial Vision, 10*(4), 433–436

Brown, W. (1910). Some experimental results in the correlation of mental abilities 1. *British Journal of Psychology, 1904-1920*, *3*(3), 296–322

Brüne, M., Abdel-Hamid, M., Lehmkämper, C., & Sonntag, C. (2007). Mental state attribution, neurocognitive functioning, and psychopathology: What predicts poor social competence in schizophrenia best? *Schizophrenia Research, 92*(1–3), 151–159

Byom, L. J., & Mutlu, B. (2013). Theory of mind: Mechanisms, methods, and new directions. *Frontiers in Human Neuroscience, 7*, 413

Callahan, C. M., Unverzagt, F. W., Hui, S. L., Perkins, A. J., & Hendrie, H. C. (2002). Six-item screener to identify cognitive impairment among potential subjects for clinical research. *Medical Care, 40*(9), 771–781. https://doi.org/10.1097/00005650-200209000-00007

Cassidy, B. S., Hughes, C., & Krendl, A. C. (2021). Age differences in neural activity related to mentalizing during person perception. *Aging, Neuropsychology, and Cognition, 28*(1), 143–160

Chung, Y. S., Barch, D., & Strube, M. (2014). A meta-analysis of mentalizing impairments in adults with schizophrenia and autism spectrum disorder. *Schizophrenia Bulletin, 40*(3), 602–616

Cortes, D. S., Tornberg, C., Bänziger, T., Elfenbein, H. A., Fischer, H., & Laukka, P. (2021). Effects of aging on emotion recognition from dynamic multimodal expressions and vocalizations. *Scientific Reports, 11*(1), 2647

Dawel, A., Miller, E. J., Horsburgh, A., & Ford, P. (2021). A systematic survey of face stimuli used in psychological research 2000–2020. *Behavior Research Methods*, 1–13

Demichelis, O. P., Coundouris, S. P., Grainger, S. A., & Henry, J. D. (2020). Empathy and theory of mind in Alzheimer's disease: A meta-analysis. *Journal of the International Neuropsychological Society, 26*(10), 963–977

Dziobek, I., Fleck, S., Kalbe, E., Rogers, K., Hassenstab, J., Brand, M., Kessler, J., Woike, J. K., Wolf, O. T., & Convit, A. (2006). Introducing MASC: A movie for the assessment of social cognition. *Journal of Autism and Developmental Disorders, 36*, 623–636

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika, 10*(4), 507–521

Frith, C., & Frith, U. (2005). *Theory of Mind. Current Biology, 15*(17), R644–R645

Garrett, D. D., Kovacevic, N., McIntosh, A. R., & Grady, C. L. (2010). Blood oxygen level-dependent signal variability is more than just noise. *Journal of Neuroscience, 30*(14), 4914–4921

Garson, G. D. (2012). Testing statistical assumptions. In: Statistical associates publishing Asheboro, NC

Gedek, H. M., Pantelis, P. C., & Kennedy, D. P. (2018). The influence of presentation modality on the social comprehension of naturalistic scenes in adults with autism spectrum disorder. *Autism, 22*(2), 205–215

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences, 102*, 74–78

Grainger, S. A., Steinvik, H. R., Henry, J. D., & Phillips, L. H. (2019). The role of social attention in older adults' ability to interpret

naturalistic social scenes. *Quarterly Journal of Experimental Psychology, 72*(6), 1328–1343

Hamilton, L. S., & Huth, A. G. (2020). The revolution will not be controlled: Natural stimuli in speech neuroscience. *Language, Cognition and Neuroscience, 35*(5), 573–582

Heavey, L., Phillips, W., Baron-Cohen, S., & Rutter, M. (2000). The Awkward Moments Test: A naturalistic measure of social understanding in autism. *Journal of Autism and Developmental Disorders, 30*, 225–236

Henry, J. D., Phillips, L. H., Ruffman, T., & Bailey, P. E. (2013). A meta-analytic review of age differences in theory of mind. *Psychology and Aging, 28*(3), 826

Hughes, C., Cassidy, B. S., Faskowitz, J., Avena-Koenigsberger, A., Sporns, O., & Krendl, A. C. (2019). Age differences in specific neural connections within the Default Mode Network underlie theory of mind. *NeuroImage, 191*, 269–277

Hughes, C., Faskowitz, J., Cassidy, B. S., Sporns, O., & Krendl, A. C. (2020). Aging relates to a disproportionately weaker functional architecture of brain networks during rest and task states. *NeuroImage, 209*, 116521

Johansson Nolaker, E., Murray, K., Happé, F., & Charlton, R. A. (2018). Cognitive and affective associations with an ecologically valid test of theory of mind across the lifespan. *Neuropsychology, 32*(6), 754

Kanske, P., Böckler, A., Trautwein, F.-M., & Singer, T. (2015). Dissecting the social brain: Introducing the EmpaToM to reveal distinct neural networks and brain–behavior relations for empathy and Theory of Mind. *NeuroImage, 122*, 6–19

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox-3?

Krendl, A. C., Hugenberg, K., & Kennedy, D. P. (2023). Comparing data quality from an online and in-person lab sample on dynamic theory of mind tasks. *Behavior Research Methods*, 1–23

Krendl, A. C., Mannering, W., Jones, M. N., Hugenberg, K., & Kennedy, D. P. (2023b). Determining whether older adults use similar strategies to young adults in theory of mind tasks. *The Journals of Gerontology: Series B, 78*(6), 969–976

Lecce, S., Ceccato, I., & Cavallini, E. (2019). Investigating ToM in aging with the MASC: From accuracy to error type. *Aging, Neuropsychology, and Cognition, 26*(4), 541–557

Linas, B., Genz, A., Westergaard, R. P., Chang, L. W., Bollinger, R. C., Latkin, C., & Kirk, G. D. (2016). Ecological momentary assessment of illicit drug use compared to biological and self-reported methods. *JMIR mHealth and uHealth, 4*(1), e4470.

Logothetis, N. K. (2002). The neural basis of the blood–oxygen–level–dependent functional magnetic resonance imaging signal. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *357*(1424), 1003–1037

Masson, H. L., & Isik, L. (2021). Functional selectivity for social interaction perception in the human superior temporal sulcus during natural viewing. *NeuroImage, 245*, 118741

Moran, J. M., Jolly, E., & Mitchell, J. P. (2012). Social-cognitive deficits in normal aging. *Journal of Neuroscience, 32*(16), 5553–5561

Moran, J. M., Wig, G. S., Adams, R. B., Jr., Janata, P., & Kelley, W. M. (2004). Neural correlates of humor detection and appreciation. *NeuroImage, 21*(3), 1055–1060

Nastase, S. A., Goldstein, A., & Hasson, U. (2020). Keep it real: Rethinking the primacy of experimental control in cognitive neuroscience. *NeuroImage, 222*, 117254

Otti, A., Wohlschlaeger, A. M., & Noll-Hussong, M. (2015). Is the medial prefrontal cortex necessary for theory of mind? *PLoS ONE, 10*(8), e0135912

Pantelis, P. C., Byrge, L., Tyszka, J. M., Adolphs, R., & Kennedy, D. P. (2015). A specific hypoactivation of right temporo-parietal junction/posterior superior temporal sulcus in response to socially awkward situations in autism. *Social Cognitive and Affective Neuroscience, 10*(10), 1348–1356

Peterson, C. C., Garnett, M., Kelly, A., & Attwood, T. (2009). Everyday social and conversation applications of theory-of-mind understanding by children with autism-spectrum disorders or typical development. *European Child & Adolescent Psychiatry, 18*(2), 105–115

Phillips, L. H., Bull, R., Allen, R., Insch, P., Burr, K., & Ogg, W. (2011). Lifespan aging and belief reasoning: Influences of executive function and social cue decoding. *Cognition, 120*(2), 236–247

Quesque, F., & Rossetti, Y. (2020). What do theory-of-mind tasks actually measure? Theory and practice. *Perspectives on Psychological Science, 15*(2), 384–396

Rabin, J. S., & Rosenbaum, R. S. (2012). Familiarity modulates the functional relationship between theory of mind and autobiographical memory. *NeuroImage, 62*(1), 520–529

Raichle, M. E. (1998). Behind the scenes of functional brain imaging: A historical and physiological perspective. *Proceedings of the National Academy of Sciences, 95*(3), 765–772

Risko, E. F., Laidlaw, K. E., Freeth, M., Foulsham, T., & Kingstone, A. (2012). Social attention with real versus reel stimuli: Toward an empirical approach to concerns about ecological validity. *Frontiers in Human Neuroscience, 6*, 143

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. https://doi.org/10.1016/s1053-8119(03)00230-1

Schaafsma, S. M., Pfaff, D. W., Spunt, R. P., & Adolphs, R. (2015). Deconstructing and reconstructing theory of mind. *Trends in Cognitive Sciences, 19*(2), 65–72

Scheeren, A. M., de Rosnay, M., Koot, H. M., & Begeer, S. (2013). Rethinking theory of mind in high-functioning autism spectrum disorder. *Journal of Child Psychology and Psychiatry, 54*(6), 628–635

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neuroscience & Biobehavioral Reviews, 42*, 9–34

Schurz, M., Radua, J., Tholen, M. G., Maliske, L., Margulies, D. S., Mars, R. B., Sallet, J., & Kanske, P. (2021). Toward a hierarchical model of social cognition: A neuroimaging meta-analysis and integrative review of empathy and theory of mind. *Psychological Bulletin, 147*(3), 293

Serre, F., Fatseas, M., Swendsen, J., & Auriacombe, M. (2015). Ecological momentary assessment in the investigation of craving and substance use in daily life: A systematic review. *Drug and Alcohol Dependence, 148*, 1–20

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika, 52*(3/4), 591–611

Sonkusare, S., Breakspear, M., & Guo, C. (2019). Naturalistic stimuli in neuroscience: Critically acclaimed. *Trends in Cognitive Sciences, 23*(8), 699–714

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*(3), 271

Stone, V. E., Baron-Cohen, S., & Knight, R. T. (1998). Frontal lobe contributions to theory of mind. *Journal of Cognitive Neuroscience, 10*(5), 640–656

Todorov, A. (2008). Evaluating faces on trustworthiness: An extension of systems for recognition of emotions signaling approach/avoidance behaviors. *Annals of the New York Academy of Sciences, 1124*(1), 208–224

Turner, R., & Felisberti, F. M. (2017). Measuring mindreading: A review of behavioral approaches to testing cognitive and affective mental state attribution in neurologically typical adults. *Frontiers in Psychology, 8*, 47

Udochi, A. L., Blain, S. D., Sassenberg, T. A., Burton, P. C., Medrano, L., & DeYoung, C. G. (2022). Activation of the default network during a theory of mind task predicts individual differences in agreeableness and social cognitive ability. *Cognitive, Affective, & Behavioral Neuroscience*, 1–20

Warnell, K. R., & Redcay, E. (2019). Minimal coherence among varied theory of mind measures in childhood and adulthood. *Cognition, 191*, 103997

Yaremych, Haley E.., & Persky, Susan. (2019). Tracing physical behavior in virtual reality: A narrative review of applications to social psychology. *Journal of Experimental Social Psychology, 85*, 103845. https://doi.org/10.1016/j.jesp.2019.103845

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

Yeung, E. K. L., Apperly, I. A., & Devine, R. T. (2023). Measures of individual differences in adult theory of mind: A systematic review. *Neuroscience & Biobehavioral Reviews*, 105481

Zaitchik, D. (1990). When representations conflict with reality: The preschooler's problem with false beliefs and "false" photographs. *Cognition, 35*(1), 41–68

Zhaoyang, R., Sliwinski, M. J., Martire, L. M., & Smyth, J. M. (2018). Age differences in adults' daily social interactions: An ecological momentary assessment study. *Psychology and Aging, 33*(4), 607