# Theory of Minds: Managing mental state information in working memory is associated with the dorsomedial subsystem of the default network and social integration

Meghan L. Meyer & Eleanor Collier
Dartmouth College
Department of Psychological and Brain Sciences

Corresponding Author:

Meghan L. Meyer
Dartmouth College
HB 6207 Moore Hall
Hanover, NH 03755
Phone: 603-646-3181
Meghan.L.Meyer@Dartmouth.Edu

**Abstract**

We often interact with multiple people at a time and consider their various points-of-view to facilitate smooth social interaction. Yet, no prior research systematically examines how our brains track multiple mental states at once, nor whether skill in this domain links to navigating real-world social interactions. To fill this gap, we developed a novel social working memory paradigm in which participants manage two- or four-people's mental states in working memory, as well as control trials in which they alphabetize two- or four-people's names in working memory. In Study 1, we found that the dorsomedial subsystem of the default network (dorsomedial prefrontal cortex and tempoparietal junction) increases activity with more mental states managed in working memory. In contrast, these regions decrease activity with more non-mental state information (the number of names alphabetized) managed in working memory. In Study 2, only individual differences in managing mental states in working memory, specifically on trials that posed the greatest mental state load to working memory, correlated with social network integration. Collectively, these findings add further support to the hypothesis that social working memory relies on partially distinct brain systems and may be a key ingredient to success in a social world.

**Key Words:** Theory of mind, social cognition, default network, working memory

Imagine the following scenarios: a business meeting in which employees and bosses keep track of multiple colleagues' intentions; an interracial interaction in which individuals manage their stereotypes and anxieties during social inference; a gossip session in which friends tailor what is said based on one another's knowledge of the situation. These scenarios are tied by a common theme: people actively manage a great deal of social information in mind to ensure smooth social interactions. How do we pull off such complex social information processing on the fly?

The maintenance and manipulation of high-level social information (e.g., personality traits, mental states, interpersonal relationships) has been characterized as social working memory (SWM) and there is evidence that this process relies on partially distinct brain systems from those supporting non-social forms of working memory. Specifically, regions of the brain's default network, which includes medial prefrontal cortex (MPFC), dorsomedial prefrontal cortex (dMPFC), precuneus/posterior cingulate (PC/PCC), tempoparietal junction (TPJ), temporal poles (TPs) and the hippocampal formation, systematically increase activation when participants maintain and manipulate high-level social information in their minds (i.e., information about friends' personalities; Meyer, Spunt, Berkman, Taylor, & Lieberman, 2012; Meyer, Taylor, & Lieberman, 2015). In contrast, this same network is well-known to systematically disengage during non-social forms of working memory (Anticevic, Repovs, Shulman, & Barch, 2010; McKiernan, Kaufman, Kucera-Thompson, & Binder, 2003). In fact, the default network even shows this differential relationship when participants process the same stimuli either socially or non-socially in working memory. In one study, participants completed working memory trials in which they were shown two, three, or four of their own friends' names and next either ranked them along trait dimensions (e.g., who is the funniest?) or alphabetized them (e.g., whose name is first alphabetically?). Within the same participants, the default network parametrically increased activation as a function of the number of friends' traits considered in working memory even though it parametrically decreased activation as a function of the number of friends' names alphabetized in working memory (Meyer

et al., 2015). Thus, the default network appears to be differentially associated with high-level social and non-social information processing in working memory.

While these findings are provocative, they generate many more questions than answers. A particularly important question to answer is whether the results generalize to multiple types of high-level social information processing managed in working memory. To date, only one paradigm has been developed to study high-level social information processing in working memory. As noted above, in this paradigm participants rank their friends 'from most-to-least' along trait dimensions during a delay period and subsequently answer true/false questions about their ranking. While such trait processing surely 'counts' as high-level social information processing, if the default network is differentially associated with SWM, similar findings should emerge for other forms of high-level social information managed in working memory.

Additionally, it has been suggested that SWM capacity may be an important ingredient for social integration: just as non-social working memory skills relate to academic success (Conway, Conwan, Bunting, Therriault, & Minkoff, 2002; Rohde & Thompson, 2007; St. Clair-Thompson & Gathercole, 2006), SWM skills may relate to interpersonal success (Meyer et al., 2012; Meyer et al., 2015; Krol, Meyer, Lieberman, & Bartz, 2018). Indeed, individuals who can manage more high-level social information in mind should make the most accurate predictions about others and critically, feeling accurately understood by peers promotes friendship with them (Cross, Bacon, & Morris, 2000; Reis & Shaver, 1988). Yet, pinpointing whether SWM accuracy relates to social integration has been challenging, in part, because responses to the previously used trait ranking paradigm are subjective. When a participant indicates that one of their friends is funnier than the other on a given SWM trial, it is unclear whether or not they answered accurately. What is needed is a SWM paradigm in which task accuracy can be computed objectively, so that individual differences in SWM capacity can be linked to differences in social integration.

The goal of the present research was to create a new SWM paradigm to fill these gaps. In this paradigm, participants first watch a video montage, which introduces a social network of

characters with different interpersonal relationships (i.e., friends, lovers, competitors, and enemies). Participants next complete a SWM task in which they are asked to determine how a given character would feel, based on other characters' feelings. Critically, the number of characters shown varies from trial-to-trial, with half of the trials showing two characters and the other half showing four characters. This allows for the isolation of neural activity and task performance as a function of SWM load, above and beyond mental state inference per se. Because interpersonal relationships create known emotional contingencies between people (e.g., people are sad to learn their friends are unhappy but delight to hear their enemies are unhappy (Cikara, Botvinick, & Fiske, 2011)), we were able to measure SWM task accuracy objectively. The paradigm also included non-SWM trials that were matched on a number of features (e.g., task difficulty), but that did not induce mental state inference. With this paradigm, we were able to test whether considering multiple mental states in working memory differentially engages the default network (Study 1) and whether SWM task accuracy uniquely predicts social integration (Study 2).

## Study 1

### Participants

Thirty-seven participants (13 male, 23 female, 1 other; Mean age=29.32 years, SD=11.32; 65% Caucasian, 19% Asian, 5% Hispanic or Latino, 11% other) completed Study 1 for course credit or monetary payment ($20/hour). Participants were eligible to participate if they did not have any metal in their body, were not claustrophobic, and were right-handed. The sample size was estimated to provide 99% power to detect a within-subject working memory type x load level interaction. Participants provided informed consent in accordance with the Dartmouth College Institutional Review Board (IRB).

### Procedure

*Social Working Memory Paradigm.* Directly before fMRI scanning, participants first watched a montage introducing them to characters from the television show *Orange is the New Black*. A professional television editor developed the montage to help ensure that it effectively

communicated the interpersonal relationships between the characters in an engaging format. Five characters were introduced and relationships between the characters included friends, enemies, lovers, and competitors. Before running the study, a sample of pilot subjects watched and described the relationships between the characters (N=19). All participants accurately described the relationships, suggesting that the relationship dynamics are well conveyed by the montage.

In Study 1, after watching the montage, participants indicated whether or not they had previously seen the television show *Orange is the New Black*. This allowed us to subsequently examine whether task performance varies as a function of familiarity with the characters in the montage. Participants next received instructions regarding how to complete the SWM and non-SWM trials and completed six practice trials (2 per condition) on a laptop so that they were familiar with the task prior to scanning.

While undergoing fMRI scanning, participants completed the primary task (Figure 1), which employs a 2 (working memory type: SWM vs. non-SWM) x 2 (working memory load: two vs. four) experimental design. Participants completed 36 randomly presented trials (9 per condition). For two-load SWM trials, participants see two characters on the screen (4 seconds). The mental state of one character is provided, indicated by a thumbs-up (positive mental state) or thumbs-down (negative mental state). The mental state of the other 'target' character is not provided. Next, participants are instructed to consider how the target character would feel, based on the other character's given mental state and relationship with the target character during a delay period (4 seconds). For example, if the two characters are friends and the character with the shown mental state is feeling positive, then the participant would reason that the other character, who likes their friend, would feel positive. After making the rating on a button box, the trial advanced to jittered fixation for a randomly generated duration between the range of 1.55-4.47 seconds (M=2.94 seconds).

Four-load SWM trials have a similar format. The key difference is that for these trials participants are shown *three* characters with mental states and one target character with no

mental state. During the subsequent delay period, participants reason how the target character would feel, based on all of the other characters' feelings. Thus, two-load and four-load trials are identical, except that the four-load trials afford greater SWM load. To further help determine task accuracy objectively, participants were instructed to reason about the mental states 'from left to right.' That is, in the example shown in Figure 1B, participants would reason about Piper's mental state in the order of Larry's influence on her mental state (she would be unhappy if Larry, her fiance, was unhappy (thus a rating of 1), her happiness would be taken up a notch if she find out that her enemy Tiffany was happy (now a rating of 2), and up another notch if she found out Claire, her girlfriend, is happy (now a rating of 3, the final, correct answer for this trial).

Non-SWM trials have the same format, except here participants alphabetize the characters' names over a delay period, also based on a (non-social) relationship. Participants see a set of characters and an anchor sign is shown under one character (4 secs). Participants are told that the anchor indicates that for a given trial, the alphabet line should start with the first letter of the character's name with the anchor. For example, if the anchor appeared under the character named Piper, 'P' should be treated as the first letter of the alphabet and 'U' should be treated as the last letter. During the delay period, participants alphabetize the characters' names based on this contingency. To ensure that the answer could vary for two-load trials, participants were shown the character with the anchor for half of the probe-responses and the target character for the other half of the probe-responses. As is the case in the SWM trials, the four-character trials were identical to the two-character trials, but simply afforded a greater working memory load. In their probe response, participants indicated the alphabetical position, based on the contingency, of the shown character.

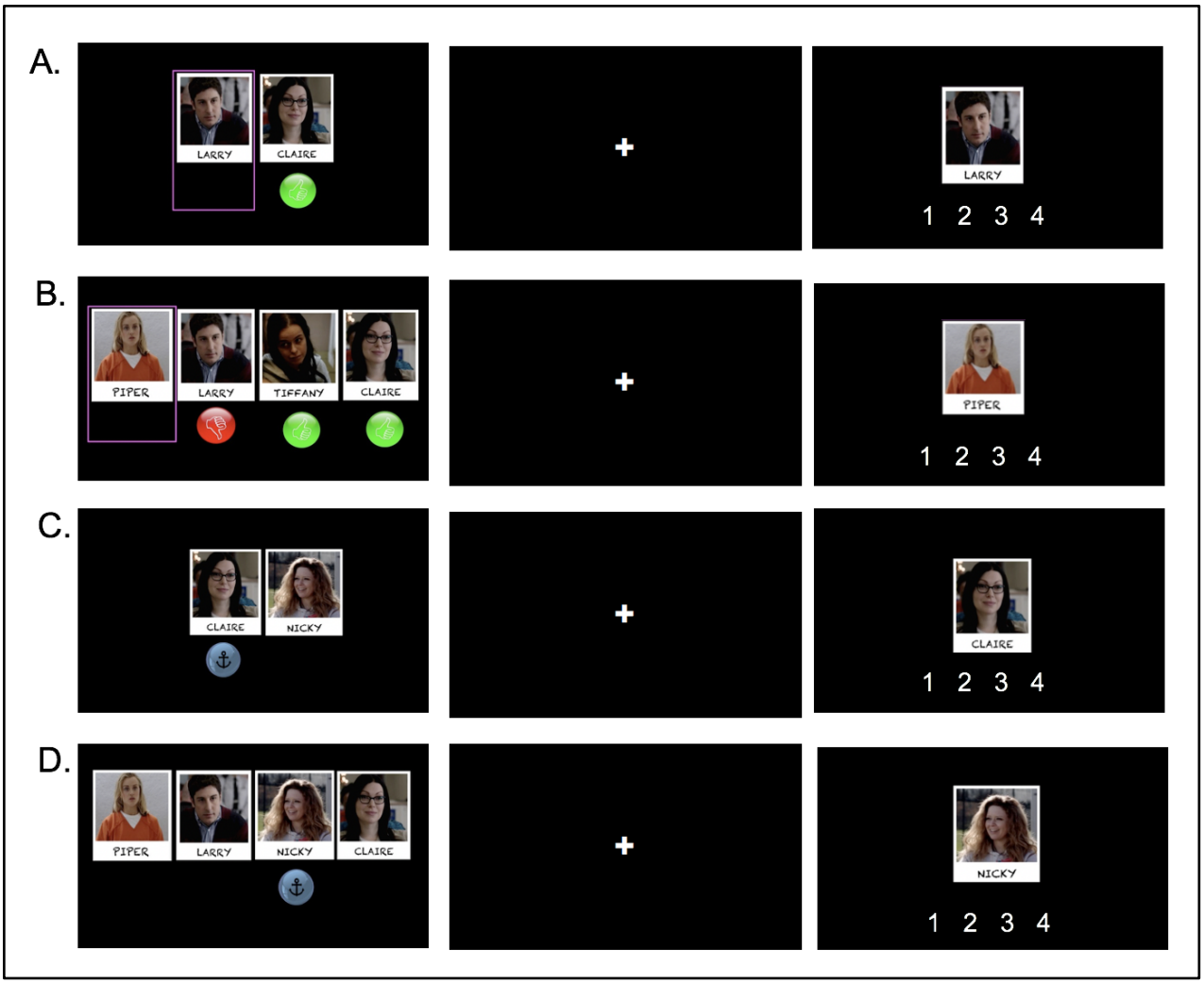


*Figure 1. Social Working Memory (SWM) Paradigm. For SWM trials (Panels A-B), participants determine how the target character (pink box) would feel, based on the other characters' feelings. For non-SWM trials (Panels C-D), participants alphabetize the characters' names, based on whose name anchors the alphabet line. Panels A and C show two-load working memory trials and Panels B and D show four-load working memory trials.*

*fMRI Scanning.* Scanning was conducted with a Siemens Trio 3T Prisma. Participants first completed a T2-weighted structural scan acquired coplanar with the functional images and with the following parameters: TR=2300 ms, TE=2.32 ms, 0.9 mm slice thickness, FOV=24 cm, matrix=256 × 256, flip angle=8°. During this time, participants watched the *Orange is the New*

*Black* montage for a second time to refresh their memory of the relationships between characters. The task was completed during two functional runs using an EPI gradient-echo sequence with the following parameters: TR=1000 ms, TE=30 ms, 2.5 mm slice thickness, FOV=24 cm, matrix=96 × 96 and flip angle=59°.

*fMRI Data Analysis.* Functional brain imaging data was first reoriented in SPM8 (Wellcome Department of Cognitive Neurology, Institute for Neurology, London, United Kingdom) and skull-stripped with Brain Extraction Tool (BET; Smith, 2002) in FSL. All subsequent fMRI analyses were performed with SPM8. Preprocessing steps included spatial realignment to correct for head motion, normalization into a standard stereotactic space as defined by the Montreal Neurological Institute, and spatial smoothing using an 8-mm Gaussian kernel, full width at half-maximum.

For each subject, neural activity was modeled for each condition using a general linear model. Because we observed differences in reaction time across our SWM and non-SWM conditions (see Results section below), we also ran models that included, for each condition, RT as a parametric modulator. These results are shown in Supplementary Figure 1. These models were run to ensure that clusters that varied between content (SWM vs. non-SWM) and load level (two-load vs. four-load) were not conflated with differences in RT across these conditions. Participants' first level models included six motion regressors for each of the motion parameters from image realignment. Three contrasts were computed for each subject. The primary contrast of interest was the interaction testing for greater activity associated with four (vs. two) SWM trials relative to four (vs. two) non-SWM trials. For completion, we also tested the main effect of working memory content (SWM vs. non-SWM) and the main effect of load level (two characters vs. four characters). Each subject's first-level models were then taken to a second, group-level analysis to test for clusters of activity associated with each contrast. Whole-brain results were thresholded at $p<.001$, cluster extent corrected for family wise error rate (FWE), which corresponded with the following cluster extent for each contrast: 131 voxels (content x load level interaction), 163 voxels (main effect of working memory content), and 161 voxels (main effect of load level).

**Results**

*Task Performance.* Task accuracy was defined as the percentage of correctly answered trials for each condition. There was no significant interaction between working memory content (SWM vs. non-SWM) and load level (two-load vs. four-load; $F(1,36)=.05$, $p=.820$, np2=.001; Figure 2A), suggesting that task performance was equivalent between the SWM and non-SWM trials. Consistent with this, there was no main effect of working memory content (SWM vs. non-SWM; $F(1,36)=2.97$, $p=.094$, np2=.08). In contrast, there was a significant main effect of load ($F(1,36)=81.90$, $p<.0001$, np2=.70) such that four-load trials were more challenging (M=55%, SD=24%) than two-load trials (M=82%, SD=17%). Task accuracy did not statistically differ between participants who had previously seen *Orange is the New Black* (N=15) relative to participants who had not seen the show (N=22; *p*'s across load level comparisons>.177, Cohen's d's<.42; See *Supplementary Table 1*). Collectively, these results suggest that task accuracy was equivalent for the SWM and non-SWM trials, that the four-load manipulation successfully taxes SWM and non-SWM, and that prior exposure to the television show *Orange is the New Black* does not moderate task performance.

Reaction time (RT) was defined as the speed (in seconds) with which participants made correct responses. Although participants' accuracy was equivalent on the SWM and non-SWM trials, RT varied across conditions. There was a significant content (SWM vs. non-SWM) x load level (two-load vs. four-load) interaction ($F(1,36)=31.21$, $p<.0001$, np2=.48), such that the decrement in speed for the four-load vs. two-load trials was larger for non-SWM trials (two-load M=.98, SD=.26; four-load M=1.74, SD=.36) than SWM trials (two-load M=.88, SD=.32; four-load M=1.11, SD=.65). There was also a main effect of content (SWM vs. non-SWM, $F(1,36)=28.56$, $p<.0001$, np2=.46), such that participants responded more quickly to SWM trials (M=1.00, SD=.47) than non-SWM trials (M=1.34, SD=.24). As is the case for task accuracy, there was also a main effect of load level (two-load vs. four-load, $F(1,36)=131.32$, p<.0001, np2=.79) such that participants responded more quickly to two-load (M=.93, SD=.25) than four-load trials (M=1.40,

SD=.41). RT did not statistically differ between subjects who had previously seen the television show *Orange is the New Black* compared to those who had not (*p*'s across load level comparisons>.481; Cohen's d's<.26).

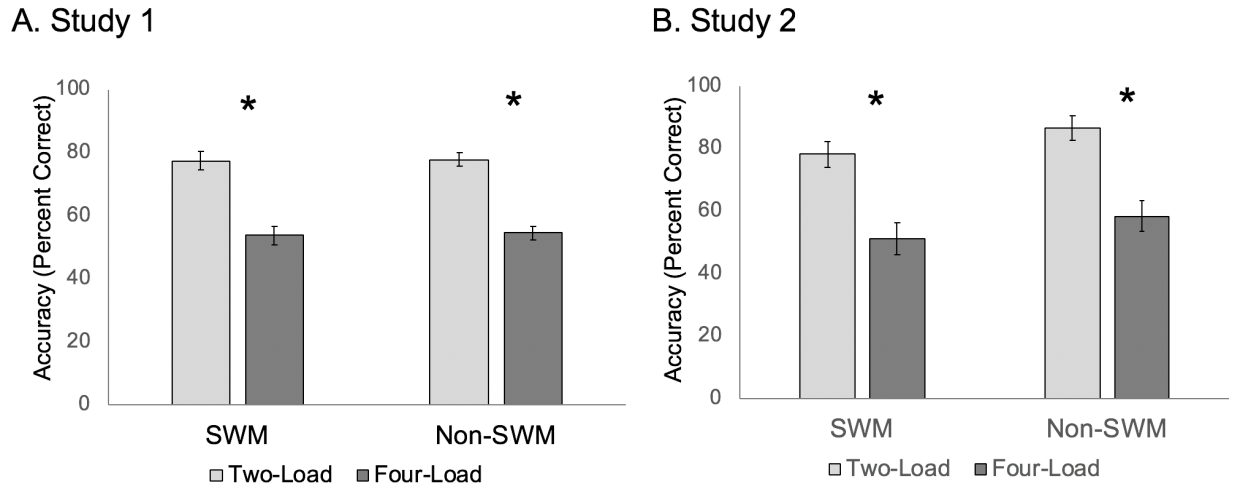


*Figure 2. Task accuracy in Study 1 (Panel A) and Study 2 (Panel B). In both studies, accuracy was matched for the SWM and non-SWM trials as a function of load level. Additionally, four-load trials were significantly more challenging than two-load trials.*

*fMRI Results.* The primary goal of the neuroimaging analyses was to test whether default network regions are differentially associated with SWM and non-SWM load. To test this, we used an interaction contrast to search for clusters of activity more strongly associated with four (vs. two) SWM trials relative to four (vs. two) non-SWM trials. This analysis revealed clusters in the dorsomedial subsystem of the default network, specifically DMPFC and TPJ bilaterally (Figure 3A, Supplementary Table 2). Parameter estimates from these clusters confirmed the hypothesized direction of the interaction: the clusters showed increased activity in response to SWM load and decreased activity in response to non-SWM load (Figure 3B). This is the first demonstration that these regions are sensitive to mental state demands to working memory. The

reverse contrast showed clusters of activation in supplementary motor area, caudate nucleus, posterior parietal lobe, and precentral and postcentral gyri (Supplementary Table 2).

Collapsing across load level, we found that multiple regions in the default network (MPFC, VMPFC, left PC/PCC, lTPJ, and bilateral TP) showed greater activity in response to the SWM vs. non-SWM trials (Supplementary Figure 1, Supplementary Table 2). These results further support the idea that the default network is associated with SWM, and mental state inference more generally. Collapsing across working memory content, and replicating past working memory research, clusters in DLPFC and bilateral inferior parietal lobe showed greater activity in response to four-load vs. two-load trials (Supplementary Figure 1, Supplementary Table 2). For the interested reader, clusters that emerged in the comparison of non-SWM vs. SWM included lateral prefrontal cortex, posterior parietal lobe, anterior insula and middle occipital lobe (Supplementary Figure 1, Supplementary Table 2). Only the MPFC showed greater activity to two-load vs. four-load trials (Supplementary Figure 1, Supplementary Table 2). Follow-up analyses showed that all clusters that differentiated SWM and non-SWM trials remained significant when analyses included, for each condition, RT as a parametric modulator (Supplementary Figure 2), suggesting observed differences in neural activity cannot be explained by differences in RT between SWM and non-SWM trials.

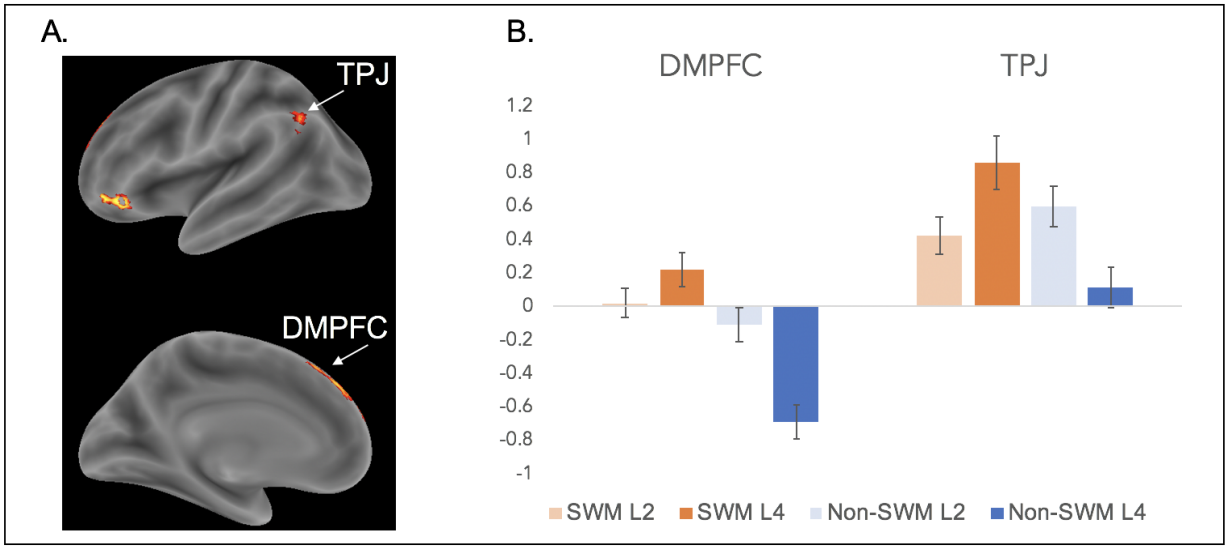


*Figure 3. Neuroimaging results for the interaction contrast testing for activity associated with four-load (vs. two-load) SWM trials relative to four-load (vs. two-load) non-SWM trials. Panel A shows clusters in the dorsomedial subsystem of the default network (DMPFC, TPJ). Panel B shows parameter estimates from the DMPFC and TPJ for each condition.*

The goal of Study 1 was to test, for the first time, whether default network regions increase activity as a function of the number of mental states managed in working memory. DMPFC and TPJ default network regions systematically increased activation as a function of the number of mental states managed in working memory, despite showing decreased activity as a function of the number of names alphabetized in working memory. Critically, SWM and non-SWM trials were matched on task accuracy and these neural patterns persisted when controlling for differences in RT across conditions. Collectively, these results suggest that engaging default network regions during SWM generalizes to managing mental states in working memory. Because our task measures SWM performance objectively and given past suggestions that SWM skill may relate to social success (Meyer et al., 2012; Meyer et al., 2015; Krol et al., 2018), our next goal addressed in Study 2 was to assess whether individual differences in SWM preferentially relate to individual differences in social integration.

## Study 2

### Participants

Eighty-eight participants (Males=48, Females=49; average age=23.53 years, SD=4.49,66% Caucasian, 16% Asian, 9% Black or African American, 7% Hispanic or Latino, 2% Other) completed Study 2. Given research showing that social network size decreases during the years following young adulthood (English & Carstensen, 2014), to avoid age confounds in our results we constrained recruitment to participants between the ages of 18-29. Sample size was determined by a power analysis in G*Power 3.1, which showed that at least 85 participants were needed to allow for 80% power to detect a medium effect size when examining the correlation between task performance and social network size. Participants completed the study for course credit or monetary payment ($6). Two participants did not complete the social integration questionnaire, leaving a final sample of 86 for correlation analysis.

### Procedure

*Social Working Memory Paradigm.* Participants completed the SWM paradigm developed for Study 1. Participants watched the *Orange is the New Black* montage and indicated whether or not they had seen the television show in the past. Next, they received task instructions, completed practice trials, and finally performed the experimental task consisting of 36 randomly presented trials (9 per condition).

*Social Integration.* To assess social integration, we used a questionnaire developed by Stiller & Dunbar (2007) which measures the number of social network members interacted with over the past seven days. In this measure, participants are asked to list all of the people with whom they had personal contact or communication within the past 7 days, excluding i) work colleagues seen only in a work environment (unless they considered them genuine friends), ii) contacts with professionals for appointments (such as doctors), and iii) other casual acquaintances (e.g., brief encounters in a shop).

### Results

*Task Performance.* We replicated the task accuracy findings observed in Study 1. There was neither a significant content (SWM vs. non-SWM) x load level (two-load vs. four-load) interaction ($F(1, 87)=1.67$, $p=.200$, np2=.02), nor a significant main effect of content (SWM vs. non-SWM; $F(1, 87)=1.36$, p=.248, np2=.02; Figure 2B). These findings again suggest that the SWM and non-SWM trials are matched on accuracy. There was a significant main effect of load level (one-load vs. three-load; $F(1,87)=139.32$, $p<.0001$, np2=.61) such that four-load trials (M=75%, SD= 19%) were more challenging than two-load trials (M=54%, SD=20%). As in Study 1, task accuracy did not statistically differ between participants who had previously seen *Orange is the New Black* (N=32) relative to participants who had not (N=39; *p*'s>.407, Cohen's d's<.19; See *Supplementary Table 1*). Collectively, these results again suggest that difficulty level (in terms of accuracy) was equivalent for the SWM and non-SWM trials, that the four-load manipulation successfully taxes working memory resources, and that previous exposure to the television show does not moderate task performance.

We also replicated the RT results observed in Study 1. There was a significant content (SWM vs. non-SWM) x load level (two-load vs. four-load) interaction ($F(1,84)=48.28$, $p<.0001$, np2=.37), such that the decrement in speed for the four-load vs. two-load trials was larger for non-SWM trials (two-load M=1.04, SD=.41; four load M=1.67, SD=.37) than SWM trials (two-load M=.73, SD=.29; four load M=.92, SD=.39). There was a main effect of content (SWM vs. non-SWM, $F(1, 84)=255.47$, $p<.0001$, np2=.75), such that participants responded more quickly to SWM (M=.82, SD=.29) than non-SWM trials (M=1.35, SD=.26). As is the case for task accuracy, there was also a main effect of load level (two-load vs. four-load, $F(1, 84)=185.48$, $p<.0001$, np2=.69) such that participants responded more quickly to two-load (M=.89, SD=.24) vs. four-load trials (M=1.30, SD=.29). RT did not statistically differ between subjects who had previously seen Orange is the New Black compared to those who had not (*p*'s across load level comparisons>.105; Cohen's d's<.40).

*Relationship Between SWM and Social Integration*. Interaction with social network members significantly correlated with SWM accuracy on the two-load trials ($r$=.21, $p$=.045) and four-load trials ($r$=.31, $p$=.003, Figure 4A). A regression analysis demonstrated that the relationship between social network member interaction and four-load SWM accuracy remained significant when controlling (i.e., adding as a covariate) two-load SWM accuracy (SWM L4 $\boldsymbol{\beta}$=.30, $t$=2.23, $p$=.028; SWM L2 $\boldsymbol{\beta}$=.03, $t$=.21, $p$=.834). Thus, SWM capacity, above and beyond mental state inference in general, may be key to navigating social network interactions. In contrast, interaction with social network members did not significantly correlate with non-SWM accuracy on the two-load ($r$=.05, $p$=.643) or four-load trials ($r$=.17, $p$=.108).

Reaction time analyses paralleled those observed for task accuracy. Individuals who interact with the most social network members are also faster to make a correct response to SWM four-load trials ($r$=-.23, $p$=.032, Figure 4B) and marginally SWM two-load trials ($r$=-.19, $p$=.073). In contrast, interaction with social network members was not significantly related to reaction time on non-SWM four-load trials ($r$=-.09, $p$=.393) nor two-load trials ($r$=-.17, $p$=.107).

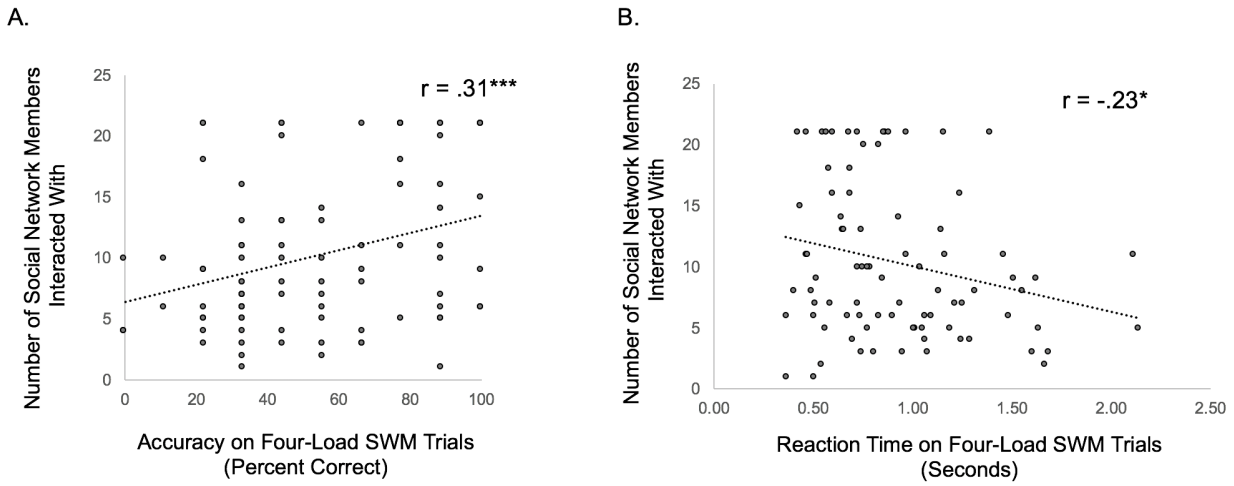


*Figure 4. Relationship between social working memory (SWM) task performance on four-load trials and the number of social network members interacted with over a seven-day period. Panel*

*A shows results for task accuracy and Panel B shows results for reaction time (RT). *** indicates p<.005 and * indicates p<.05.*

## Discussion

How do we keep track of multiple people's mental states at a time? Here we found that DMPFC and TPJ, key nodes of the dorsomedial subsystem of the default network (Andrews-Hanna, Reidler, Sepulcre, Poulin, & Buckner, 2010; Yeo et al., 2011), increase activity as a function of the number of mental states considered in working memory. In contrast, these regions decreased activity as a function of non-mental state information (the number of names alphabetized) considered in working memory. Such findings are consistent with the suggestion that the default network is differentially associated with SWM and non-SWM. Moreover, only individual differences in SWM task performance, specifically on trials that posed the greatest SWM load, correlated with social integration. Thus, SWM capacity may be a key ingredient to navigating everyday social life.

Although the default network comprises several brain regions, here we found that only the DMPFC and TPJ increased activity with the number of mental states managed in working memory. These findings are consistent with past research on mental state inference, which consistently implicate TPJ and DMPFC in assessing others' minds (Dodell-Feder, Koster-Hale, Bedny, & Saxe, 2011; Gallagher, Happe, Brunswick, Fletcher, Frith, & Frith, 2000; Schurz, Radua, Aichhorn, Richlan, & Perner, 2014). Relevant to understanding the role of different default network regions in SWM, graph analytic methods show that the default network is composed of three subsystems: the core subsystem comprising MPFC and PC/PCC, the medial temporal lobe subsystem comprising hippocampal formation and ventromedial MPFC, and the dorsomedial subsystem comprising DMPFC and TPJ, as well as TPs (Andrews-Hanna et al., 2010). Thus, it is possible that different default subsystems are preferentially associated with maintaining and manipulating different types of social information in working memory.

For example, MPFC and PC/PCC are consistently implicated in self-referential processing, as well as thinking about close others, such as friends and family (Denny, Kober, Wager, & Ochsner, 2012; Krienen, Tu, & Buckner, 2010). Here we observed that MPFC reduced activation in response to the higher social and non-social load trials (i.e., the comparison of two-load vs. four-load trials, collapsed across SWM and non-SWM trials). In contrast, previous SWM research found that MPFC increased activity as a function of the number of friends' traits considered in working memory (Meyer et al., 2012; 2015). Our findings suggest that MPFC disengages in response to working memory load, unless it is needed for the working memory process at hand, as is the case when managing social information about *close others*. Interestingly, no research has investigated the brain basis of managing self-concept information in working memory, despite social psychological theory suggesting this process is important for self-regulation and fulfilling interpersonal relationships (Markus & Wurf, 1986). Future research should parameterize working memory load across self, close other, and stranger targets to further clarify the role of MPFC, and the default network subsystems more generally, in SWM.

The observation that the dorsomedial subsystem preferential supports mental states in working memory provides new insight into the functional properties of this system and its role in theory-of-mind (TOM), or the capacity to represent others' mental states. The classic paradigm used to assess TOM is the false-belief test (Wimmer & Perner, 1983; Dufour et al., 2013). In this paradigm, participants first observe person A witness person B place an item in one of two boxes. Person A leaves the room and Person B moves the item to the other box. Person A returns to the room and the participant is asked which box Person A believes the item is in. Although not traditionally framed as a working memory task, determining a correct answer to false belief tasks requires working memory; participants must maintain Person A's belief over time to answer correctly. Thus, an alternative interpretation of TPJ and DMPFC activity in response to false-belief tasks is that these regions support the momentary maintenance of others' mental states in working memory. One impediment to this interpretation from past work is that traditional false-

belief tasks do not parameterize working memory load, making it difficult to disentangle whether these regions are important for mental state inference generally versus maintaining mental state inferences in working memory specifically. Because our two-load and four-load SWM trials required mental state inference, but only the four-load trials increased mental state load in working memory, our results provide helpful insight: TPJ and DMPFC increase activation as a function of the number of mental state demands to working memory, above and beyond mental state inference per se. Future work may build on this finding, assessing whether activity in these regions during false belief tasks reflects a (social) working memory process.

Our paradigm is also useful for assessing TOM capacity in adults, who traditionally perform at ceiling on most false-belief tasks (Brunet, Sarfati, Hardy-Bayle, & Decety, 2000; Fletcher et al., 1995; Walter, Ciaramidaro, Enrici, Pia, & Bara, 2004). TOM research focuses heavily on when children acquire the capacity to infer mental states and how these mechanisms are altered in populations who struggle to understand others, such as individuals with an autism spectrum disorder (ASD). However, far less research examines the upper bounds of TOM expertise, perhaps in part because there are little-to-no paradigms that effectively create variability in healthy adult performance. In contrast, participants in Studies 1 and 2 did not perform at ceiling on our tasks, and individual differences in task performance related to real-world social integration. Thus, our paradigm may be an efficient way to begin to unpack individual differences in adult TOM and its link to social prowess.

Our paradigm is also the first to assess the brain basis of a critical skill for navigating social networks: reasoning about how interpersonal relationships create emotional contingencies between people. Imagine a work meeting, in which your close friend and fellow colleague offers a suggestion to the group. Some of the other colleagues strongly agree with the idea and others strongly disagree. Knowing how to respond to your friend requires more than just tallying who did and did not like the suggestion. It further requires appreciating the relationships between your friend and the other colleagues. Dissent from a competitor and dissent from another friend will

impact your friend's feelings quite differently (Cikara et al., 2011). Our primary goal here was to develop a new SWM paradigm, and the emotional contingency framework created an efficient way to do so. The next step is to more fully unpack how the social brain represents the emotional contingencies between social network members. Representational similarity analysis may prove particularly fruitful in this regard. For example, this approach could tease apart whether DMPFC and TPJ represent the most salient emotional contingencies in response to a given event (e.g., if something bad happens to a person, their closest friend and greatest enemy's reactions may be similarly represented as salient in these brain regions). Alternatively, these regions may cluster emotional contingencies by valence (e.g., mental states of friends and family may cluster together and, separately, the mental states of enemies and competitors may cluster together). Our approach paves the way for future work to examine these and related questions surrounding how the brain represents emotional contingencies between people.

The behavioral results from our new paradigm also surpasses the limitations of past paradigms used to assess whether SWM skills relate to social network integration. Past work has shown that the ability to manage people's personality traits in working memory (Krol et al., 2018) and maintain information about characters' mental states from a story (Stiller et al., 2007) correlate with the number of social network members interacted with over the course of a week. The proposed interpretation of these results is that skill in maintaining and manipulating people's internal characteristics in working memory may help us cultivate social networks. However, two limitations impede this interpretation. The first study used the SWM task in which participants rank a set of friends along trait dimensions 'from most-to-least' in working memory and determine accuracy by comparing participants' answers on a given SWM trial to their ratings of their friends' traits provided two weeks earlier. Thus, task accuracy is computed with subjective assessments of friends' traits, and variables such as whether a new event changed a participant's perception of their friend(s) confound the results. The finding that the ability to remember characters' mental states from a story correlates with social network integration confounds other variables with task

performance, most notably reading comprehension. The paradigm developed here surpasses these limitations, as task accuracy is computed objectively and does not rely on reading comprehension (all stimuli are presented visually). Moreover, we found that task accuracy specifically on the four-load SWM trials, above and beyond task accuracy on the two-load SWM trials, uniquely predicts the number of social network members interacted with over a week. This finding further suggests that the ability to maintain and manipulate mental states in working memory, above and beyond mental state inference per se, is associated with navigating larger social networks. Future work may reveal the causal arrow of this relationship, assessing whether SWM skills prospectively predict social network size, whether cultivating larger social networks builds SWM capacity, or both.

Developing this SWM paradigm is also critical to advancing understanding of SWM mechanisms in clinical disorders, many of which are characterized by deficits in social cognition, working memory, or both. For example, individuals with ASD struggle to understand other people's mental states, yet some individuals with high-functioning ASD have spared non-SWM skills (Bennetto, Pennington, & Rogers, 1996; Ozonoff & Strayer, 2001). In contrast, individuals with social anxiety show heightened sensitivity to social information, but deficits in non-SWM (Amir & Bomeyea, 2011). Finally, schizophrenia (Couture, Penn, Roberts, 2006; Goldman-Rakic, 1994; Pickup & Frith, 2001), post-traumatic stress disorder (PTSD; Plana, Lavoie Battaglia, & Achim, 2014; Vasterling et al., 2002), and attention deficit disorder (ADD; Gropper & Tannock, 2009; Martinussen, Hayden, Hogg-Johnson, & Tannock, 2005; Ukermann et al., 2010) are all associated with both social cognitive difficulties and compromised working memory skills. Given these patterns, SWM may be important to examine in these populations. However, the previous SWM trait ranking paradigm requires that participants have at least ten close friends for stimulus generation. Individuals with mental health conditions often struggle to make social connections (Beeney et al., 2019; Girard et al., 2017), making it difficult to implement the previous SWM task

in these populations. The new paradigm presented here does not require that a participant have a full social network and thus may be better suited to study SWM skills in clinical populations.

In summary, we examined, for the first time, how people manage mental state inferences in working memory and whether skill in this domain relates to real-world social integration. Study 1 showed that regions of the dorsomedial subsystem of the default network increase in activity as a function of the number of mental states considered in working memory, despite decreasing as a function of the number of non-mental state information considered in working memory. Study 2 showed that skill in managing more mental states in working memory preferentially correlated with social integration. These findings add further support to the idea that SWM relies on partially distinct brain mechanisms and may be critical to interpersonal success.

**Acknowledgments:** We would like to thank Joseph Mastromonaco for creating the *Orange is the New Black* montage and Matthew Lieberman for his suggestions on the experimental design. We would also like to thank Thuy-Vy Nguyen and Remy Borinsky for their assistance with data collection and analysis for Study 2.

**References**

Amir, N., & Bomyea, J. (2011) Working memory capacity in generalized social phobia. Journal of Abnormal Psychology, 120, 504–509.

Andrews-Hanna, J. R., Reidler, J. S., Sepulcre, J., Poulin, R., & Buckner, R. L. (2010). Functional-anatomic fractionation of the brain's default network. Neuron, 65, 550-562.

Anticevic, A., Repovs, G., Shulman, G. L., & Barch, D. M. (2010). When less is more: TPJ and default network deactivation during encoding predicts working memory performance. NeuroImage, 49(3), 2638-2648.

Bennetto L, Pennington BF, Rogers SJ (1996) Intact and impaired memory functions in autism. Child Development, 67, 1816–1835.

Beeney, J. E., Lazarus, S. A., Hallquist, M. N., Stepp, S. D., Wright, A. G. C., Scot, L. N, … Pilkonis, P. A. (2019). Detecting the presence of a personality disorder using interpersonal and self-dysfunction. Journal of Personality Disorders, 33(2), 229-248.

Brunet, E., Sarfati, Y., Hardy-Bayle, M. C., & Decety, J. (2000). A PET investigation of the attribution of intentions with a nonverbal task. NeuroImage, 11, 157–166.

Cikara, M., Botvinick, M. M., & Fiske, S. T. (2011). Us versus them: Social identity shapes neural responses to intergroup competition and harm. Psychological Science, 22, 306-313.

Conway, A. R. A., Conwan, N., Bunting, M. F., Therriault, D. J., & Minkoff, S. R. B. (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. Intelligence, 30, 163-183.

Couture, S. M., Penn, D. L., & Roberts, D. L. (2006) The functional significance of social cognition in schizophrenia: A review. Schizophrenia Bulletin, 32(Suppl 1), S44–S63.

Cross, S.E., Bacon, P.L., Morris, M.L. (2000). The relational-interdependent self-construal and relationships. Journal of Personality and Social Psychology, 78(4), 791.

Denny, B. T., Kober, H., Wager, T. D., & Ochsner, K. N. (2012). A meta-analysis of functional neuroimaging studies of self- and other judgments reveals a spatial gradient for mentalizing in medial prefrontal cortex. Journal of Cognitive Neuroscience, 24(8), 1742-1752.

Dodell-Feder, D., Koster-Hale, J., Bedny, M., & Saxe, R. (2011). fMRI item analysis in a theory of mind task. NeuroImage, 55(2), 705-711.

Dufour, N., Redcay, E., Young, L., Mavros, P. L., Moran, J. M., Triantafyllou, C., … Saxe, R. (2013). Similar brain activation during false belief tasks in a large sample of adults with and without autism. PlosOne, doi: 10.1371/journal.pone.0075468

English, T., & Carstensen, L. L. (2014). Selective narrowing of social networks across adulthood is associated with improved emotional experience in daily life. International Journal of Behavioral Development, 38(2), 1-8.

Fletcher, P. C., Happ´, T., Frith, U., Baker, S. C., Dolan, R. J., Frackowiak, R. S. J., & Frith, C. D. (1995). Other minds in the brain: A functional imaging study of theory of mind in story comprehension. Cognition, 57, 109–128.

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

Girard, J. M., Wright, A. G. C., Beeney, J. E., Lazarus, S. A., Scott, L., Stepp, S., & Pilkonis, P. A. (2017). Interpersonal problems across levels of psychopathology hierarchy. Comparative Psychiatry, 79, 53-69.

Goldman-Rakic, P. S. (1994) Working memory dysfunction in schizophrenia. Journal of Neuropsychiatry and Clinical Neurosciences, 6, 348–357.

Gropper, R. J., & Tannock, R. (2009). A pilot study of working memory and academic achievement in college students with ADHD. Journal of Attention Disorders, 12(6), 574-581.

Krienen, F. M., Tu, P.-C., & Buckner, R. I. (2010). Clan Mentality: Evidence that the medial prefrontal cortex responds to close others. The Journal of Neuroscience, 30(41), 13906-13915.

Krol, S. A., Meyer, M. L., Lieberman, M. D., & Bartz, J. A. (2018). Social working memory predicts social network size in humans. Adaptive Human Behavior and Physiology. DOI: 10.1007/s40750-018-0100-9.

Markus, H., & Wurf, E. (1986). The dynamic self-concept: A social psychological perspective. Annual Review of Psychology, 38, 299-337.

Martinussen, R., Hayden, J. D. C., Hogg-Johnson, S., Tannock, R. (2005). A meta-analysis of working memory impairments in children with attention-deficit/hyperactivity disorder. Journal of the American Academy of Child & Adolescent Psychiatry, 44(4), 377-384.

McKiernan, K. A., Kaufman, J. N., Kucera-Thompson, J., & Binder, J. R. (2003). A parametric manipulation of factors affecting task-induced deactivation in functional neuroimaging. Journal of Cognitive Neuroscience, 15(3), 394-408.

Meyer, M. L., Spunt, R. P., Berkman, E. T., Taylor, S. E., & Lieberman, M. D. (2012). Evidence for social working memory from a parametric functional MRI study. Proceedings of the National Academy of Sciences, 109(6), 1883-1888.

Meyer, M. L., Taylor, S. E., & Lieberman, M. D. (2015). Social working memory and its distinctive link to social cognitive ability: an fMRI study. Social Cognitive and Affective Neuroscience, 10(10), 1338-1347.

Ozonoff S, Strayer DL (2001) Further evidence of intact working memory in autism. J Autism Dev Disord 31:257–263.

Pickup, G. J., Frith, C. D. (2001) Theory of mind impairments in schizophrenia: Symptomatology, severity and specificity. Psychological Medicine, 31, 207–220.

Plana, I., Lavoie, M.-A., Battaglia, M., & Achim, A. M. (2014). A meta-analysis and scoping review of social cognition performance in social phobia, posttraumatic stress disorder and other anxiety disorders. Journal of Anxiety Disorders, 28, 169-177.

Reis, H.T., Shaver, P. (1988). Intimacy as an interpersonal process. In: Duck, S., editor. Handbook of Personal Relationships. Chichester, UK: Wiley, pp. 367–389.

Rohde, T. E. & Thompson, L. A. (2007). Predicting academic achievement with cognitive ability. Intelligence, 35(1), 83-92.

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

Smith, S. M. (2002) Fast robust automated brain extraction. Human Brain Mapping, 17, 143-155.

St. Clair-Thompson, H. L., & Gathercole, S. E. (2006). Executive functions and achievements in school: Shifting, updating, inhibition, and working memory. The Quarterly Journal of Experimental Psychology, 59(4), 745-759.

Stiller, J., & Dunbar, R. I. M. (2007). Perspective-taking and memory capacity predict social network size. Social Networks, 29(1), 93–104.

Uekermann, J., Kraemer, M. Abdel-Hamid, M., Schimmelmann, B. G., Hebebrand, J., Daum, J., Wiltfang, B. K. (2010). Social cognition in attention-deficit hyperactivity disorder (ADHD). Neuroscience & Biobehavioral Reviews, 34(5), 734-743.

Vasterling, J. J., Duke, L. M., Brailey, K., Constans, J. I., Allain, J., Albert, N., & Sutker, P. B. (2002). Attention, learning, and memory performances and intellectual resources in Vietnam veterans: PTSD and no disorder comparisons, Neuropsychology, 16(1), 5-14.

Walter, H., Ciaramidaro, A., Enrici, I., Pia, L., & Bara, B. G. (2004). Understanding intentions in social interaction: The role of the anterior paracingulate cortex. Journal of Cognitive Neuroscience, 16, 1854–1863.

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

Yeo, B. T. T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M.,... Buckner, R. L. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. Journal of Neurophysiology, 106, 1125–1165. http://dx.doi.org/10.1152/ jn.00338.2011