Accepted at *Journal of Experimental Psychology: General* on 17th September 2019.

# Understanding How Minds Vary Relates to Skill in Inferring Mental States, Personality, and Intelligence.

Jane R. Conway[1,2*], Michel-Pierre Coll[3], Hélio Clemente Cuve[3], Sofia Koletsi[4], Nicholas Bronitt[4], Caroline Catmur[4] & Geoffrey Bird[2,3]

[1] Institute for Advanced Study in Toulouse.

[2] MRC Social, Genetic & Developmental Psychiatry Centre, Institute of Psychiatry, Psychology & Neuroscience, King's College London.

[3] Department of Experimental Psychology, University of Oxford.

[4] Department of Psychology, Institute of Psychiatry, Psychology & Neuroscience, King's College London.

*Correspondence: jane.conway@iast.fr and geoff.bird@psy.ox.ac.uk.

Word count: 8535.

**Abstract**

Using a 'theory of mind' allows us to explain and predict others' behaviour in terms of their mental states, yet individual differences in the accuracy of mental state inferences are not well understood. We hypothesised that the accuracy of mental state inferences can be explained by the ability to characterise the mind giving rise to the mental state. Under this proposal, individuals differentiate between minds by representing them in 'Mind-space' – a multidimensional space where dimensions reflect any characteristic of minds that allows them to be individuated. Individual differences in the representation of minds and the accuracy of mental state inferences are explained by one's model of how minds can vary (Mind-space), and ability to locate an individual mind within this space. We measured the accuracy of participants' model of the covariance between dimensions in Mind-space that represent personality traits, and found this was associated with the accuracy of mental state inference (Experiment 1). Mind-space accuracy also predicted the ability to locate others within Mind-space on dimensions of personality and intelligence (Experiment 2). Direct evidence for the representation of minds in mental state inference was obtained by showing that the location of others in Mind-space affects the probability of particular mental states being ascribed to them (Experiment 3). This latter effect extended to mental states dependent upon representation of trait covariation (Experiment 4). Results support the claim that mental state inference varies according to location in Mind-space, and therefore that adopting the Mind-space framework can explain some of the individual differences in theory of mind.

*Keywords:*

theory of mind; individual differences; personality; social cognition; Mind-space.

## Introduction

When trying to understand other people's behaviour, our explanations are greatly enriched by referring to their mental states, such as what they believe, know, desire or intend. This 'theory of mind' (ToM) ability is considered crucial in social interactions, from everyday relationships to political negotiations and criminal trials. The scientific study of ToM has spanned 40 years (Premack & Woodruff, 1978) and multiple disciplines, including developmental, socio-cognitive, clinical, and comparative psychology, artificial intelligence, and neuroscience (Gallagher & Frith, 2003; Happé, 1994; Heyes, 2015; Rabinowitz et al., 2018). However, a fundamental challenge in the ToM literature persists: what is it that makes some people better at inferring mental states than others (see Repacholi & Slaughter, 2003, for discussion)?

There are two main reasons why individual differences in ToM have been difficult to explicate. First, empirical measurement of unobservable mental states is difficult, necessitating that for most tasks the 'correct' and 'incorrect' mental state inferences are predetermined by the authors based on rationality and logic (Baron-Cohen, Leslie, & Frith, 1985) or by consensus (Dziobek et al., 2006). With such task designs, performance does not reflect the accuracy of mental state inference, but instead how rational, or how typical, mental state inferences are. Even when task performance has the potential to reflect the accuracy rather than rationality/typicality of the participant's mental state inference (e.g. the 'Beauty Contest', Nagel, 1995), results provide little insight into individual variance in the representational or inferential processes by which that inference was derived (Heyes, 2014). Second, due to these difficulties measuring the accuracy of mental state inferences, individual differences in performance on ToM tasks have typically been attributed to domain-

general abilities (Devine & Hughes, 2014; Milligan, Astington, & Dack, 2014; Sabbagh, Xu, Carlson, Moses, & Lee, 2006) rather than domain-specific processes or representational structures. Verbal skills, memory, or inhibitory control contribute to performance on ToM tasks that demand those abilities, but cannot explain variance unique to mental state inference.

Previous work describing improvements in ToM from early to late childhood and into adulthood has revealed continuing improvements in mental state inference (so-called 'advanced ToM', e.g. Osterhaus, Koerber & Sodian, 2016). This work details how, during development, individuals gradually incorporate additional sources of information into their mental state inferences, and therefore provides one framework within which to understand individual differences in ToM. For example, as social and emotional understanding becomes (1) increasingly more sophisticated, and (2) integrated into mental state inferences (e.g. Baron-Cohen, O'Riordan, Stone, Jones, & Plaisted, 1999; Burnett, Bird, Moll, Frith, & Blakemore, 2009), individual differences in either the degree of social/emotional understanding or its integration into mental state reasoning could explain individual differences in the accuracy of mental state inferences.

The work presented here is concerned with a second way in which individual differences in the accuracy of mental state inference can be understood: the representation of others' minds. Crucially, minds moderate the link between situational contexts and the mental states they evoke: two different target minds in the same situation may generate completely different mental states. The accuracy with which those target minds can be represented, therefore, is likely to contribute to

accuracy in inferring the target's mental states. Thus, the experiments reported here address how individual differences in mind representation may give rise to individual differences in the accuracy of mental state inference. The work is based on the hypothesis that a major source of naturalistic variance in the probability of others having particular mental states is variability in the people in one's environment. Mental states are the product of a specific individual mind, and therefore accurate representation of how minds vary likely affects the accuracy of any mental state inference (Conway et al., 2019).

Empirical work suggests that representation of minds, and the processes occurring within minds, are initially not explicitly integrated with mental state inference, but become so as children develop. For example, Ruffman (1996) found that until 7 years of age children often find it easier to attribute an incorrect false belief than a correct true belief, when attributing a true belief would require the child to understand the distinction between knowledge states in an individual's mind (i.e. they may be ignorant about X but know Y). Instead, young children applied a simple rule of the form "if a person didn't see something then they cannot know it". Thus, for children below 7 years of age, in at least some situations, mental state inference is determined by the situation an individual is in, not by a model of how minds, and the processes within minds, inform mental states.

Older children slowly begin to understand explicitly the link between minds and mental state inferences. This is most clearly demonstrated by the work on 'interpretive theory of mind', the understanding that two individuals can be exposed to exactly the same information and yet draw different conclusions. For example,

children above 7 years of age are able to understand that two individuals who are shown the same small portion of a picture can make different inferences about the picture as a whole (Lalonde & Chandler, 2002). Around the age of 10, children can understand that it is impossible to know which of two percepts will be formed by an unknown individual when they perceive an ambiguous figure which affords two distinct percepts (such as a visual illusion; Osterhaus et al., 2016).

With respect to an implicit understanding of the link between minds and mental states, a rudimentary understanding may be gained in childhood and is certainly present during adolescence and adulthood. For example, during stereotyping, individuals decide that minds of a certain type (e.g. those belonging to out-groups) are more likely to hold particular beliefs or to have certain intentions than minds of another type (e.g. those of the in-group). To illustrate, work on Fiske, Cuddy, Xu, & Glick's (2002) Stereotype Content Model has shown that the two dimensions characterising stereotype content (warmth and competence) are associated with changes in the frequency of inferred mental states. For example, the warmth dimension changes the inferred intentions of the stereotyped individual, such that groups associated with high warmth are expected to hold positive intentions towards the self, while those associated with low warmth are expected to hold negative intentions towards the self (see Fiske et al., 2002). While these mental states are broad and non-specific, they may be operationalised in very specific ways in particular contexts. For example, during a sales negotiation, a member of a group stereotyped as warm may be thought to favour fairness over profit, while a member of a group stereotyped as cold might be expected to favour profit over fairness. Even children of between 3-5 years of age show a rudimentary understanding of gender stereotypes,

and use them to determine what males and females are likely to desire (Aboud, 1988; Wellman & Liu, 2004). Thus, from relatively early in development, judgements of the probability of particular mental states are altered on the basis of the type of mind giving rise to them (although this link may not be explicitly represented until late childhood).

The preceding work demonstrates therefore, that at least by older childhood or adolescence, a target's mind is explicitly represented in order to infer the probability of particular mental states. The experiments reported here build on this work to test the hypothesis that individual differences in mind representation may explain individual differences in the accuracy of mental state inferences. Specifically, we hypothesised that minds may be represented as locations within a multidimensional space ('Mind-space') in which dimensions reflect any discriminable aspect of minds, such as their cognitive abilities (e.g. intelligence) and behavioural tendencies (e.g. personality traits; Conway et al., 2019). As such, Mind-space is similar to the idea of Face-space (Valentine, 1991; Valentine, Lewis, & Hills, 2016), which is theorised to be a multi-dimensional space where dimensions represent ways in which faces can be discriminated. Once formed, individual faces are thought to be represented as points within this multi-dimensional space. Mind-space may be thought of as analogous to Face-space. For example, target minds A and B may be represented in a 3-dimensional Mind-space with dimensions of working memory, extraversion, and conscientiousness, but each target is located at a different point within the space according to their characteristics. One benefit of representing minds within a multi-dimensional space is that covariance between dimensions can be more easily represented and utilised to make mental state inferences. Locating a mind within

Mind-space could permit accurate mental state inference because the target's mental states are, in part, dependent on their location in the space. For example, if I can accurately place targets A and B along the extraversion dimension, I could better predict their respective attitudes (i.e. mental states) towards attending a party. A person is therefore more likely to be accurate at inferring a target's mental states if: (1) the person represents the relevant dimensions and any covariance between dimensions; (2) they can accurately locate a mind in Mind-space based on samples of behaviour; (3) they use a target's location in Mind-space combined with situational factors when generating mental state inferences. (See Figure 1 for a full example.)

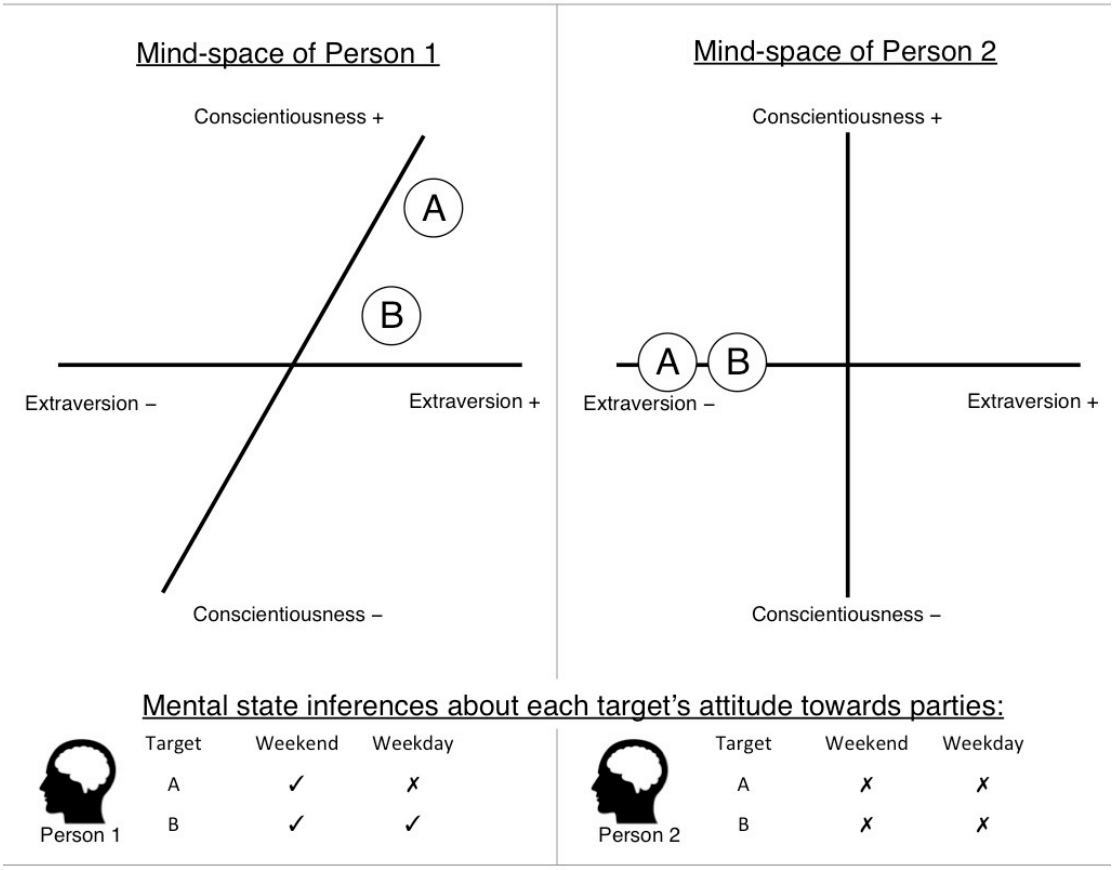


*Figure 1.* Schematic illustration of how the Mind-space framework can be used to explain individual differences in Theory of Mind (ToM). The Mind-space framework suggests that individual differences in ToM are due to: (1) The accuracy of the

representation of the dimensions within which minds vary and the relationship between these dimensions (i.e. Mind-space); (2) The ability to locate a target mind within Mind-space; (3) The ability to combine diagnostic information about the situation the target is in with the target's position in Mind-space to accurately infer their mental state; (4) The propensity to consider position in Mind-space before making a mental state inference (not illustrated). Person 1 and Person 2 are asked to estimate the attitude of two targets (A and B) towards parties on weekends and weekdays based on how extraverted they appear. Person 1 can accurately locate the targets on the extraversion dimension, but Person 2 cannot. Person 1's Mind-space accurately reflects the positive correlation between conscientiousness and extraversion whereas Person 2's does not. Due to Person 1's accurate representation of Mind-space, only Person 1 can infer the targets' degree of conscientiousness on the basis of their degree of extraversion. This enables Person 1 to infer that because Target A is more extravert than B, Target A is also more conscientious than B, and so Person 1 can predict that Target A will more likely have diverging attitudes to parties on the weekend vs. a weekday. Person 2 has no basis to predict differential attitudes to parties based on the day of the week, and this is furthered compounded by their failure to locate the targets accurately within their Mind-space. As a result, Person 1 makes more accurate mental state inferences than Person 2.

We aimed to measure the accuracy of the covariance between dimensions that represent personality traits in an individual's Mind-space. Personality is particularly apt for this first test of the Mind-space theory because factor analyses have established that traits can be represented using five (Goldberg, 1990) or six (Ashton & Lee, 2007) dimensions. Although each dimension is distinct there is some degree of

correlation between them, thus the existing personality literature provides ground truth values for the average covariance between traits in the population (or at least ground truth values for the population completing a particular personality test at a particular moment in history). The presence of covariation across a number of dimensions would be most efficiently represented in a multi-dimensional space such as Mind-space. We therefore developed the 'Personality Pairs Task' which asks participants to estimate the average correlations between traits on six personality dimensions (Ashton & Lee, 2009). These estimated correlations can then be compared to ground truth values from a similar population to determine the accuracy of an individual's Mind-space. If there exists a relationship between the representation of minds and the inference of mental states, we hypothesised that performance on a ToM task would be associated with Mind-space accuracy (Experiment 1).

In Experiment 2, we sought to test whether Mind-space accuracy predicts the ability to locate a target mind within Mind-space. Accordingly, participants in Experiment 2 completed the Personality Pairs Task and were asked to estimate the personality and intelligence of a number of targets on the basis of video-recorded 'thin-slices' of behaviour. Such thin-slices provide minimal experience of a target yet can result in surprisingly accurate predictions of their traits and abilities (Borkenau, Mauer, Riemann, Spinath, & Angleitner, 2004; Carney, Colvin, & Hall, 2007). Participants were asked to locate each target on personality and intelligence dimensions and their estimates were compared to ground truth values we collected for each target. If Mind-space accuracy predicts the ability to locate an individual within Mind-space, scores on the Personality Pairs Task should predict the accuracy of participants' target location estimates. The design of Experiment 2 also allowed us to

assess if similarity in personality between the participant and the target affects the accuracy of trait judgements. Higher accuracy for targets similar to the self may reflect an egocentric bias whereby participants anchor their judgements of the targets' traits on their own traits (Epley, Keysar, Van Boven, & Gilovich, 2004), and such egocentricity would result in more accurate judgements when the target is similar, but less accurate judgements for dissimilar targets. Under the Mind-space framework, providing one can accurately locate oneself within Mind-space, similarity effects would be due to increased experience of the mapping between one's position in Mind-space and behaviour across situations. This greater experience would enable a target's position in Mind-space to be derived from behaviour more accurately, and across a greater number of situations, if the target occupied a similar position as the self within Mind-space (Conway et al., 2019). Under either account, if similarity in personality between the participant and the target affects the accuracy of trait judgements, then we should observe higher accuracy on the thin-slice location task for targets that are similar to the participant compared to those who differ.

Even if results in accordance with the predictions of the Mind-space framework are observed in Experiments 1 and 2, it could be argued that they do not provide a direct test of the Mind-space framework itself. They are not designed to provide evidence that participants incorporate the position of a target mind within Mind-space when inferring the content of their mental states. Accordingly, in Experiment 3, we investigated how the position of targets in Mind-space, combined with situational information, affects the probability of particular mental states being inferred. This work builds on, but goes beyond, previous demonstrations that older children recognise that two minds may produce different mental states when exposed

to the same information (Lalonde & Chandler, 2002), or that different types of minds may be associated with different probabilities of generally positive or negative intentions towards the in-group (Fiske et al., 2002), by showing quantitatively the degree to which the probability of certain mental states is updated as target minds move through Mind-space, and as other minds move through the target's Mind-space.

Participants in Experiment 3 were presented with a series of vignettes based on the Sally-Anne False Belief Task (Baron-Cohen et al., 1985). In this task, Sally places a marble in her basket and leaves the scene; while she is away Anne takes the marble from Sally's basket and puts it in her own box. The critical test question asks: where will Sally look for the marble on her return? The ability to ascribe a false belief to Sally – that she will look for the marble in the location where she left it (her basket) rather than where it really is (Anne's box) – is considered a litmus test of theory of mind (Dennett, 1978; Wimmer & Perner, 1983). False belief tasks involving an unseen change-of-location have been used extensively to test the theory of mind ability of human infants (Baillargeon, Scott, & He, 2010), children (Kulke, Reiß, Krist, & Rakoczy, 2017), people with autism (Happé, 1994), non-human primates (Heyes, 2017), and artificial agents (Rabinowitz et al., 2018). However, these tasks do not take into account the representation of the particular minds of Sally and Anne; in the task they are merely anonymous protagonists (Conway et al., 2019). We presented participants with vignettes in which the Sally character varied across four levels of paranoia, and the Anne character across four levels of dishonesty. We predicted that the mental state attributed to Sally by the participant would vary as a function of where Sally was in the participant's Mind-space, and where the participant believed Anne to be in Sally's Mind-space; specifically that at higher levels of paranoia and

dishonesty, participants would be less likely to infer that Sally would look in her basket where she left her marble, and be more likely to infer that Anne has stolen the marble and hidden it in her own box. If this prediction is supported, it would provide direct evidence for the incorporation of position in Mind-space when inferring mental states.

Experiment 3 has the potential to show that a characteristic of the target mind is represented and used to inform mental state inferences for which it is relevant. It does not, however, have the potential to show that the target mind is represented within a multi-dimensional space. Experiment 4 therefore used the same basic design as Experiment 3, but tested the following prediction: that providing a participant with information about a target mind's location on certain 'source' dimensions should allow that target's mind to be located on other dimensions, to the extent that those other dimensions covary with the source dimension within that participant's Mind-space. Accordingly, Experiment 4 asked participants to complete the same false belief vignettes as in Experiment 3, for a number of Sally characters that varied on source dimensions which a validation study suggested to be associated with paranoia in the general population. If varying the position of the Sally character on the source dimensions changes the mental state attributed to her, and crucially if it does so to the degree that the participant believes each source trait covaries with paranoia, then this would provide stronger evidence for the idea that target minds are located within a multi-dimensional space, and that target location in Mind-space is used in mental state inference.

Collectively, the four experiments were designed to provide complementary tests of the Mind-space theory. As detailed above, Experiments 3 and 4 account for variability in the minds available for representation and how the location of a mind in Mind-space affects the probability of which mental state is attributed to that mind. Experiment 2 examines the ability to locate a specific mind in Mind-space and how this relates to Mind-space accuracy. First, in Experiment 1, we test for a relationship between the accuracy of Mind-space and the accuracy of mental state inferences. If the accuracy of mental state inference is indeed determined by the accuracy of Mind-space, then those individuals who have a more accurate representation of how minds vary, in this case operationalised as the covariance between personality dimensions, should also make more accurate mental state inferences.

## Experiment 1

### Method

**Participants.** Sixty adults volunteered to take part in this experiment in return for a small monetary sum or undergraduate research participation credits. Participants (48 female) were aged between 18 and 55 years old ($M$ = 23.62, $SD$ = 6.21). An *a priori* power calculation using the pwr package in R (Champely et al., 2018) indicated that for Cohen's $f^2$ = .15 and α= .05, a sample size of 58 would provide 80% power for the main hypothesis being tested (with two predictor variables). The local Research Ethics Committee approved the study.

**Measures.**

***Personality Pairs Task.*** The Personality Pairs Task (PPT) comprised 72 questions. Each question included a pair of items measuring traits on the HEXACO personality inventory (Ashton & Lee, 2009). The HEXACO-60 is a 60-item

questionnaire that captures six personality dimensions. Five of these are similar to those captured in five-factor personality models: Emotionality (E), similar to Neuroticism; Extraversion (X); Agreeableness (A); Conscientiousness (C); and Openness to Experience (O). Honesty-Humility (H) represents a sixth dimension not captured within the five-factor models (Ashton & Lee, 2007), and reflects traits of sincerity, fairness, greed-avoidance, and modesty. On each trial of the Personality Pairs Task, participants were asked to rate how likely, on average, is it that someone who has one trait would also have the other. For example: "On average, how likely is it that someone who *people think of as having a quick temper*, would also *make decisions based on the feeling of the moment rather than on careful thought*?" Participants responded using a sliding scale from 'Extremely Unlikely' (-100) to 'Neither Likely Nor Unlikely' (0), to 'Extremely Likely' (+100), and this response was divided by 100 to give a negative or positive estimated correlation coefficient. There were two pairs of traits presented for every combination of the six HEXACO personality dimensions. The actual inter-trait correlation values for the population were obtained from a sample ($N$ = 2,868) collected by Lee and Aston (Lee & Ashton, 2016). Participants' accuracy was computed by taking the absolute difference score between the population correlation and their estimated correlation between the traits, and calculating the mean difference score across the 72 trials. Smaller difference scores indicate higher accuracy at predicting the actual population correlation values, and therefore a more accurate Mind-space.

***Movie for the Assessment of Social Cognition (MASC)***. The MASC (Dziobek et al., 2006) is a naturalistic theory of mind task, which requires participants to watch a 15-minute video of four characters having dinner together. After each video segment, a multiple-choice question with four possible responses is asked.

There are 45 mental state questions and 21 control questions (Santiesteban, Banissy, Catmur, & Bird, 2015). The control questions do not require any mental state representation and account for non-mentalistic factors that may affect performance, e.g. memory, attention, verbal comprehension, or motivation. For the mental state questions, the multiple-choice options reflect four response types: no mental state inference; insufficient mental state inference; correct mental state inference; and excessive mental state inference. Participants' scores were computed as the percentage of correct responses on the mental state and control questions respectively; and for each of the three incorrect response types to mental state questions, the sum score of the number of errors was also computed (i.e. no mental state inference; insufficient mental state inference; and excessive mental state inference).

**Procedure.** Participants completed the study individually on a computer in a testing room in a single session of approximately one hour. The measures were presented in the following order: Personality Pairs Task [36 trials]; MASC; Personality Pairs Task [36 trials].

**Statistical Analyses.** Multiple regression models were performed using the *lm* function in R. To assess whether non-normality of residuals affected the models, robust regression models were also performed using the *boot* package in R (Canty & Ripley, 2017) to provide bootstrapped 95% confidence intervals of regression coefficients based on 2000 bootstrap samples. A close resemblance between the bootstrapped coefficients and the original coefficients indicated that non-normal distributions did not affect the model. The data for this study are available at https://doi.org/10.17605/OSF.IO/4K9HS.

**Results**

Descriptive statistics for all variables are presented in Table 1. To investigate whether Mind-space accuracy is associated with the accuracy of mental state inference after controlling for non-mentalistic reasoning ability, a multiple regression model was performed with PPT difference score as the outcome variable and percentage correct scores on the MASC mental state and control questions as the predictor variables (Table 2, Model 1.A: PPT mean difference score ~ MASC Mental State % correct + MASC Control % correct). The model explained a significant proportion of the variance in PPT scores, $R^2$ =0.13, $F$ (2, 57) = 4.20, $p$ = .02. As shown in Table 2 (Model 1.A), only performance on the MASC mental state questions significantly predicted accuracy on the PPT. Performance on the MASC control questions did not predict accuracy on the PPT. This suggests that those participants who performed better on a theory of mind task had a more accurate Mind-space, as indicated by lower difference scores on the PPT. That the relationship was observed for the mental state questions only, not the control questions, suggests that it is specific to theory of mind and not attributable to variance in other cognitive domains such as memory, attention, or verbal ability.

To further assess which type of theory of mind errors were associated with poorer Mind-space accuracy, a second multiple regression model was performed with PPT difference score as the outcome variable and error type sum scores on the MASC mental state questions as the predictor variables (Model 1.B: PPT mean difference score ~ MASC no mental state inference + MASC insufficient mental state inference + MASC excessive mental state inference). The model explained a significant proportion of the variance in PPT scores, $R^2$ = 0.22, $F$ (3, 56) = 5.17, $p$ = .003. Only

errors indicating no mental state inference significantly predicted performance on the PPT (Table 2, Model 1.B). Errors indicating insufficient or excessive mental state inference did not predict PPT performance. These results show that those who failed to make any mental state inference had a less accurate Mind-space, as indicated by higher difference scores on the PPT.

**Table 1**

*Descriptive Statistics for Experiment 1*

| Variable | *Mean* | *SD* | *Range* |
|---|---|---|---|
| PPT Difference Score | 0.37 | 0.13 | 0.15 – 0.70 |
| Mental State (MS) Qs % Correct | 77.55 | 11 | 40 – 93.33 |
| Control Qs % Correct | 90.79 | 6.84 | 71.43 – 100 |
| Errors: No MS Inference | 1.58 | 1.61 | 0 – 6 |
| Errors: Insufficient MS Inference | 3.58 | 3.14 | 0 – 17 |
| Errors: Excessive MS Inference | 4.93 | 2.58 | 0 – 11 |

*Note.* PPT = Personality Pairs Task. MS = Mental State. Qs = Questions.

**Table 2**

*Experiment 1 Regression Analyses: Predictors of Performance on the Personality Pairs Task*

| Predictor | *B* | *SE* | *95% CI* | *Bootstrap 95% CI* | *β* | *t* | *p* |
|---|---|---|---|---|---|---|---|
| Model 1.A | | | | | | | |
| Mental State Qs % Correct | -0.004 | 0.002 | [-0.007, -0.001] | [-0.007, -0.001] | -0.31 | -2.30 | .03* |
| Control Qs % Correct | -0.002 | 0.002 | [-0.007, 0.003] | [-0.007, 0.003] | -0.11 | -0.81 | .42 |
| Model 1.B | | | | | | | |
| Errors: No MS Inference | 0.033 | 0.010 | [0.013, 0.053] | [0.014, 0.053] | 0.41 | 3.27 | .002** |
| Errors: Insufficient MS Inference | -0.001 | 0.005 | [-0.011, 0.009] | [-0.011, 0.008] | -0.02 | -0.16 | .88 |
| Errors: Excessive MS Inference | 0.009 | 0.006 | [-0.002, 0.021] | [-0.005, 0.022] | 0.19 | 1.61 | .11 |

*Note.* Qs = Questions. MS = Mental State. * $p < .05$. ** $p < .01$.

**Discussion**

As predicted, Experiment 1 demonstrated that performance on a ToM task was associated with Mind-space accuracy as measured by the Personality Pairs Task. A relationship was observed both for overall ToM accuracy and for errors indicating a failure to infer any mental state. Building on previous evidence that adults represent others' minds when inferring mental states (e.g. Fiske et al., 2002), these results provide evidence for the relationship between the accuracy of mind representation and the accuracy of mental state inference.

In Experiment 2, we tested the following predictions: that those with a more accurate Mind-space would be better able to locate specific targets within Mind-space; and that similarity in personality to the target will affect the accuracy with which they do so (Conway et al., 2019). The accuracy of Mind-space was again measured using the Personality Pairs Task. The ability to locate individuals within Mind-space accurately was assessed using a thin-slice procedure in which participants watched short video-recordings of a number of targets reciting a simple sentence. They were asked to estimate the personality and intelligence of each target based on this 'thin-slice' of their behaviour, and participant estimates were compared to the target's actual personality and IQ scores as a measure of their accuracy. If results are as predicted, then participants who have a more accurate Mind-space as measured by the Personality Pairs Task should also be more accurate when locating individuals within Mind-space on the basis of thin-slices of their behaviour.

## Experiment 2

### Method

**Participants.** Sixty-eight adults that did not take part in Experiment 1 volunteered to take part in this experiment in return for a small monetary sum or undergraduate research participation credits. Participants (58 female) were aged between 18 and 57 years old ($M$ = 23.76, $SD$ = 7.52). An *a priori* power calculation indicated that for Cohen's $f^2$ = .15 and α= .05, a sample size of 66 would provide 80% power for the hypotheses being tested (with three predictor variables). The local Research Ethics Committee approved the study.

**Measures.**

***Behavioural samples of targets: thin-slice video stimuli.*** 'Thin-slices' of targets' behaviour were presented to participants via video stimuli. Ten males and ten females were recruited to feature as targets in the thin-slicing video stimuli. Each target was filmed from the chest up against a white background (See Supplemental Materials Video S.1, or https://doi.org/10.17605/OSF.IO/4K9HS) saying the phrase "Hi, I am a participant in this study and my ID number is xxxx". Each target was given a unique four-digit ID number to say. Video duration was between six and nine seconds (depending on the rate of the target's speech). Targets completed the self-report HEXACO-60 personality inventory, and the observer-report HEXACO-60 (Ashton & Lee, 2009) was completed by someone who knew them well. This procedure provided a mean self-reported score and observer-reported score for each target on each of the six dimensions on the HEXACO. The Matrix Reasoning and Vocabulary sub-scales of the Wechsler Abbreviated Scale of Intelligence 2nd edition

(Wechsler, 2011) were administered to targets, from which the target's Intelligence Quotient percentile rank was obtained.

***Ratings of behavioural samples of targets.*** For the personality ratings, participants were first given a description of the HEXACO personality inventory and the meaning of the six dimensions. They were provided with descriptions of all six dimensions and all statements one would agree and disagree with if one scored highly on each dimension. (Note that this task was performed after the participants completed the HEXACO in relation to their own personality and thus could not have affected their scores on this measure; see Procedure below for task order.) After the target's video was presented, participants were asked to rate that target's personality on each of the six dimensions on a sliding scale ranging from the 'lowest' to 'highest' possible score. These ratings provided a response between 1 and 5 that allowed for comparison with the target's mean on each dimension. Participant accuracy was computed by taking the absolute difference score on each dimension between (a) the target's self-reported mean and the participant's estimated mean, and (b) the target's observer-reported mean and the participant's estimated mean. Smaller difference scores indicate higher accuracy at predicting the target's personality.

For the intelligence ratings, as for personality, participants were first given instructions on how intelligence is defined and how to rate the target's intelligence compared to the general population where responses indicate the target's percentile rank (e.g. *On this scale, 'average' means that if you chose a group of 100 at random, half (50%) of them would be more intelligent and half (50%) of them would be less intelligent than the person you are rating; 'Top 25%' means that 75 people would be less intelligent than the person you are rating; 'Bottom 25%' means that 75 people*

*would be more intelligent than the person you are rating.).* After viewing the target's video, participants were asked to rate them on how intelligent they are compared to the general population on a scale from 0% to 100% with markers at 'Bottom 25%', 'Average', and 'Top 25%'. This allowed for comparison with the target's actual IQ percentile rank by taking the absolute difference score between the target's rank and the participant's estimate. As before, smaller difference scores indicate higher accuracy at predicting the target's IQ.

***Personality Pairs Task.*** As described in Experiment 1.

***Participant-Target similarity in personality.*** Participants completed the self-report HEXACO-60 personality inventory. Participants were asked to respond to statements on a 5-point Likert scale from 'Strongly Disagree' to 'Strongly Agree'. A mean score was computed for each of the six dimensions (minimum score = 1, maximum = 5). We then computed absolute difference scores between each participant and target by subtracting the participant's score for each of the six dimensions from the target's self-reported HEXACO scores. Smaller difference scores indicate more similarity between the participant and target.

**Procedure.** Participants completed the study individually on a computer in a testing room in a single session of approximately one hour. The measures were presented in the following order: Personality Pairs Task [72 trials]; Self-report HEXACO; Ratings of behavioural samples of targets from thin-slicing video stimuli [20 trials].

**Statistical analyses.** The statistical analyses were as for Experiment 1 with the addition of random effects to the linear models to take into account the variance across participants, targets and HEXACO personality dimensions. Analyses were performed using the *lmer* package (Bates et al., 2018). The data for this study are available at https://doi.org/10.17605/OSF.IO/4K9HS.

## Results

Descriptive statistics for all variables are presented in Table 3. To investigate whether those with a more accurate Mind-space were better able to locate specific targets within Mind-space, mixed models were performed. The outcome variable for Model 2.A was the difference between the target's self-reported score and the participant's estimate of it for each of the six HEXACO dimensions ('SRH difference score'). Model 2.B was similar except it used the target's observer-reported score ('ORH difference score'). Both models 2.A and 2.B had PPT difference score as the fixed effect, and participants (68) target (20) and personality dimensions (6) as random effects allowing for random intercepts. The outcome variable for Model 2.C was the difference between the target's IQ percentile and the participant's estimate of it ('IQ difference score'), with PPT difference score as the fixed effect and target (20) as the random effect. Additional information on the distribution of personality trait scores and their contribution to the accuracy of personality estimates is presented in Supplemental Materials (Fig S1 and Table S1).

**Table 3**

*Descriptive Statistics for Experiment 2*

| Variable | *Mean* | *SD* | *Range* |
|---|---|---|---|
| PPT Difference Score | 0.37 | 0.11 | 0.15 - 0.67 |
| SRH Difference Score | 0.83 | 0.62 | 0 - 3.60 |
| ORH Difference Score | 0.78 | 0.59 | 0 - 3.60 |
| IQ Difference Score | 20.58 | 14.05 | 0 - 71 |

*Note.* PPT = Personality Pairs Task. SRH = Self-report HEXACO. ORH = Observer-report HEXACO. IQ = Intelligence.

As shown in Table 4, performance on the PPT significantly predicted SRH difference scores (Model 2.A), ORH difference scores (Model 2.B), and IQ difference scores (Model 2.C). As hypothesised, those participants with a more accurate Mind-space, as indicated by lower difference scores on the PPT, were more accurate at estimating the target's self- and observer- reported scores on the HEXACO and the target's IQ percentile rank, thus supporting the prediction that they would more accurately locate targets in Mind-space based on a minimal sample of behaviour.

To investigate whether similarity in personality between the participant and the target was associated with the accuracy of trait judgements, we ran the same models as previously except now the fixed effect was the participant-target similarity score (Model 2.D: outcome variable = SRH; Model 2.E: outcome variable = ORH; Model 2.F: outcome variable = IQ). As shown in Table 5, degree of similarity significantly predicted SRH difference scores (Model 2.D) and ORH difference scores

(Model 2.E), but not IQ difference scores (Model 2.F). Participants who were more similar in personality to targets were more accurate at estimating the target's self-reported scores and observer-reported scores on the HEXACO personality measure, but personality similarity had no effect on estimates of the target's IQ.

**Table 4**

*Experiment 2: Regression Analyses*

| Predictor | Random Effects | Outcome | *B* | *SE* | *95% CI* | *Bootstrap 95% CI* | *t* | *p* |
|---|---|---|---|---|---|---|---|---|
| Model 2.A | | | | | | | | |
| PPT | Target; Personality Trait; Participant | SRH | 0.51 | 0.12 | [0.27, 0.75] | [0.26, 0.76] | 4.12 | <.001** |
| Model 2.B | | | | | | | | |
| PPT | Target; Personality Trait; Participant | ORH | 0.56 | 0.14 | [0.29, 0.83] | [0.28, 0.84] | 4.00 | <.001** |
| Model 2.C | | | | | | | | |
| PPT | Target | IQ | 5.80 | 2.70 | [0.52, 11.09] | [0.51, 11.03] | 2.15 | 0.03* |

*Note.* PPT = Personality Pairs Task. SRH = Self-report HEXACO. ORH = Observer-report HEXACO. IQ = Intelligence. For the random effects, there were 20 targets, six personality traits and 68 participants. * $p < .05$. ** $p < .001$.

**Table 5**

*Experiment 2: Regression Analyses*

| Predictor | Random Effects | Outcome | *B* | *SE* | *95% CI* | *Bootstrap 95% CI* | *t* | *p* |
|---|---|---|---|---|---|---|---|---|
| Model 2.D | | | | | | | | |
| Similarity | Target; Personality Trait; Participant | SRH | 0.17 | 0.01 | [0.15, 0.19] | [0.15, 0.19] | 16.34 | <.001** |
| Model 2.E | | | | | | | | |
| Similarity | Target; Personality Trait; Participant | ORH | 0.05 | 0.01 | [0.03, 0.07] | [0.03, 0.07] | 5.21 | <.001** |
| Model 2.F | | | | | | | | |
| Similarity | Target; Personality Trait; Participant | IQ | 0.15 | 0.19 | [-0.22, 0.52] | [-0.69, 0.25] | 0.81 | 0.42 |

*Note.* Similarity = Difference in personality between targets and participant. SRH = Self-report HEXACO. ORH = Observer-report HEXACO. IQ = Intelligence. For the random effects, there were 20 targets, six personality traits and 68 participants. ** $p < .001$.

**Discussion**

As predicted, Experiment 2 demonstrated that those with a more accurate Mind-space were better able to locate specific targets within Mind-space. Furthermore, similarity in personality to the target affected the accuracy of estimates of personality traits, but not IQ.

In Experiment 3, we sought quantitative evidence that the location of a target mind in Mind-space affects the probability of specific mental states being attributed to that target mind. Arguably, this has not been demonstrated in Experiments 1 and 2; for example, although Experiment 1 demonstrated an association between the accuracy of Mind-space and the accuracy of mental state inference (an association that was specific to mental state inference and therefore unlikely to be a product of domain-general individual differences in, for example, inferential ability or motivation), this association could be caused by individual differences in social-specific factors, such as social attention, which independently influence the accuracy of Mind-space and mental state inference, rather than the accuracy of Mind-space directly influencing the accuracy of mental state inference. Accordingly, Experiment 3 used a variant of the Sally-Anne task to vary the position of one character (Sally) within the participant's Mind-space, and the other character (Anne) within Sally's Mind-space. It was predicted that movement of a target mind along dimensions of Mind-space would alter the probability of specific mental states being attributed if they are dependent upon those dimensions given a specific situation.

The classic false belief unseen change-of-location task used in this experiment (the 'Sally-Anne' task) is a staple of ToM research (e.g. Baillargeon, Scott, & He,

2010; Kulke, Reiß, Krist, & Rakoczy, 2017; Happé, 1994; Rabinowitz et al., 2018). Experiment 3 modifies this simple task such that participants have to remember a personality feature for both characters and make a probabilistic judgement about one character's behaviour. Due to the additional working memory requirements introduced by the requirement to hold in mind the personality of the characters the use of a simple task was preferred, although the simplicity may limit the size of any effect observed.

## Experiment 3

### Method

**Participants.** Sixty-three adults volunteered to take part in this experiment in return for a small monetary sum or undergraduate research participation credits. Participants (51 female) were aged between 17 and 59 years old ($M$ = 25.08, $SD$ = 0.95). An *a priori* power calculation using G*Power (Faul, Erdfelder, Buchner, & Lang, 2009) indicated that for a medium effect size and $\alpha = .05$, a sample size of 24 would provide 80% power for the main hypotheses being tested (without covariates). The local Research Ethics Committee approved the study.

**Measures.**

***Mental State Stories.*** Thirty-two vignettes were presented to participants. Each vignette featured two characters and an unseen change-of-location as in the Sally-Anne False Belief task (Baron-Cohen et al., 1985). In each vignette: the 'Sally' character puts an object in a location; then leaves the scene during which time the 'Anne' character moves the object to a different location; 'Sally' later returns looking

for her object. There were four Sally characters (Emily, Ben, Amelia, George) and four Anne characters (Jessica, Oliver, Isabella, Jack). They are described as having been "*work colleagues for many years, so they all know one another very well*". Two vignettes were presented for every combination of Sally and Anne characters.

***Paranoia manipulation.*** The Sally characters were designed to vary across four levels of paranoia. Participants were told that these characters completed a questionnaire and were shown the questionnaire items and the characters' scores. The questionnaire items were three items taken from the Paranoia Scale (Fenigstein & Vanable, 1992), a 20-item measure of paranoia for use in non-clinical populations. The items were: *It is safer to trust no one; I tend to be on my guard with people who are somewhat more friendly than I expected; Some people have tried to steal my ideas and take credit for them*. Participants were told that the characters could score anywhere between 0 and 4 on each statement, and therefore between 0 and 12 in total, with higher scores indicating higher levels of agreement with the statements. Before each set of stories for each combination of Sally and Anne characters, participants were reminded of the items and the character's score. The four levels of paranoia corresponded to total scores of 0, 4, 8, and 12.

***Dishonesty manipulation.*** The Anne characters were manipulated across four levels of dishonesty using the same approach as for the Sally characters. The questionnaire items were three items taken from the Honesty-Humility dimension of the HEXACO personality inventory (Ashton & Lee, 2009). The items were: *If I knew I could never get caught, I would be willing to steal a million dollars; I'd be tempted to use counterfeit money, if I were sure I could get away with it; If I want something*

*from someone, I will laugh at that person's worst jokes.* The four levels of dishonesty corresponded to total scores of 0, 4, 8, and 12.

***Mental State Inference.*** After each mental state story**,** participants were asked to respond on a sliding scale with the extremes of the scale labelled with two response options. The options represented the two locations that in traditional unseen change-of-location tasks with binary measures reflect a false or true belief (i.e. respectively, where Sally knew the object to be last vs. where the object has been moved to by Anne). Participants were asked to move the slider so that it represents the probability that Sally will look in one of the two response locations. False and true belief options were counterbalanced across the right and left ends of the scale. Responses were coded so that a rating of 50 indicated neither location was more likely, ratings closer to 100 indicated greater probability of the false belief location, and ratings closer to 0 indicated greater probability of the true belief location.

***Manipulation check.*** After participants had completed all 32 mental state stories, they were shown the trials again with the Sally and Anne characters' scores and vignettes, but without the mental state inference response scale. Instead, they were asked to report, using a four-point Likert scale (from 'not at all' to 'highly'): How paranoid do you (the participant) think Sally is; How paranoid does Anne think Sally is; How honest do you (the participant) think Anne is; How honest does Sally think Anne is? This provided first and second-order inferences of the characters' traits.

***Self-report Measures.*** Participants also completed the full Paranoia scale (Fenigstein & Vanable, 1992); the Honesty-Humility subscale of the HEXACO (Ashton & Lee, 2009); the Autism Spectrum Quotient 10 (AQ10; Baron-Cohen, Wheelwright, Skinner, Martin, & Clubley, 2001), a measure of autistic traits (e.g. attention to detail or others' intentions); and the Perspective Taking Scale of the Interpersonal Reactivity Index (IRI PT; Davis, 1983), a measure of the tendency to consider another person's point of view.

**Procedure.** Participants completed the study individually on a computer in a testing room in a single session of approximately one hour. The measures were presented in the following order: Mental State Stories [32 trials]; Manipulation Check; AQ10; IRI PT; Paranoia Scale; Honesty-Humility HEXACO Scale.

**Statistical analyses.** The statistical analyses were conducted using a Repeated Measures Analysis of Variance in SPSS (v24, IBM, Armonk, NY, USA) with Paranoia (4 levels) and Dishonesty (4 levels) as the within-subject factors and the four self-report measures as covariates. The dependent variable was the probability rating on the mental state inference measure, which was the average rating of the two trials for each combination of the factor levels. Where assumptions of sphericity were violated, Greenhouse-Geisser corrected values are reported. Bonferroni corrections were used to adjust the alpha level when conducting post-hoc multiple comparisons. The data for this study are available at https://doi.org/10.17605/OSF.IO/4K9HS.

**Results**

Descriptive statistics for all variables are presented in Table 6. There were no significant effects of any of the covariates, and they were dropped from further models (note this did not affect the pattern of results). The lack of any effect of the covariates indicates that there was no relationship between participants' traits and the probability of their mental state inferences. There was a significant main effect of the Sally character's level of paranoia on the probability of the mental state inferred, $F$ (2.20, 136.11) = 57.96, $p < .001$, $\eta_p^2 = .48$. There was also a significant main effect of the Anne character's level of dishonesty on the probability of the mental state inferred, $F$ (3, 186) = 15.93, $p < .001$, $\eta_p^2 = .20$. These main effects were characterised by a significant negative linear trend indicating a reduction in the probability ratings of the Sally character looking in the location corresponding to a false belief, for both Paranoia, $F$ (1, 62) = 99.31, $p < .001$, $\eta_p^2 = .62$, and Dishonesty, $F$ (1, 62) = 32.12, $p < .001$, $\eta_p^2 = .34$ (full contrasts are shown in Table S2). The variables were not normally distributed and the robustness of ANOVA to departures of normality is debated (Glass, Peckham, & Sanders, 1972; Lix, Keselman, & Keselman, 1996), therefore two Robust Repeated Measures One-way ANOVA with 4 Factor Levels using 2000 bootstrap samples in the WRS2 package in R (Mair & Wilcox, 2018) were also carried out, and confirmed the results (Paranoia: $F = 57.06$, $F_{crit} = 2.95$, $p < .05$; Dishonesty: $F = 14.58$, $F_{crit} = 2.81$, $p < .05$; Post hoc comparisons shown in Table S3). The effects of paranoia and dishonesty on the probability of mental state inferences are shown in Figure 2.

There was a significant interaction effect between Sally's levels of paranoia and Anne's levels of dishonesty, $F$ (7.13, 441.79) = 8.82, $p < .001$, $\eta_p^2 = .12$. A simple

effects analysis showed that Sally's paranoia had an effect at all levels of Anne's dishonesty: Level 1: $V = 0.70$, $F\ (3, 60) = 47.27$, $p < .001$; Level 2: $V = 0.47$, $F\ (3, 60) = 17.86$, $p < .001$; Level 3: $V = 0.58$, $F\ (3, 60) = 27.62$, $p < .001$; Level 4: $V = 0.33$, $F\ (3, 60) = 9.64$, $p < .001$. Similarly, Anne's dishonesty had an effect at all levels of Sally's paranoia: Level 1: $V = 0.39$, $F\ (3, 60) = 12.58$, $p < .001$; Level 2: $V = 0.22$, $F\ (3, 60) = 5.65$, $p = .002$; Level 3: $V = 0.51$, $F\ (3, 60) = 21.07$, $p < .001$; Level 4: $V = 0.15$, $F\ (3, 60) = 3.56$, $p = .019$. Post hoc contrasts with corrections for multiple testing are shown in Table S3. The interaction was mainly driven by differences between levels 1 and 4 of Paranoia, with levels of Dishonesty having strongly different effects at level 1 of Paranoia but more similar effects at level 4.

The ratings of the characters' traits in the manipulation check are shown in Tables S4 and S5. Overall, they show that participants correctly inferred the characters' levels of paranoia or dishonesty from the information provided about their scores on the respective questionnaires.

**Table 6**

*Descriptive Statistics for Experiment 3*

| Variable | *Mean* | *SD* | *Range* |
|---|---|---|---|
| Mental State Probability: | | | |
| Paranoia Level 1 | 77.74 | 23.98 | 0 - 100 |
| Paranoia Level 2 | 72.22 | 21.46 | 0 - 100 |
| Paranoia Level 3 | 57.47 | 23.02 | 6.5 - 100 |
| Paranoia Level 4 | 48.94 | 26.04 | 0 - 100 |
| Dishonesty Level 1 | 69.41 | 27.50 | 0 - 100 |
| Dishonesty Level 2 | 65.76 | 24.11 | 0 - 100 |
| Dishonesty Level 3 | 61.33 | 24.93 | 0 - 100 |
| Dishonesty Level 4 | 59.87 | 27.49 | 0 - 100 |
| Honesty-Humility | 3.59 | 0.62 | 1.88 - 4.88 |
| Perspective Taking | 17.49 | 5.17 | 7 - 28 |
| Autism Quotient | 2.73 | 1.79 | 0 - 8 |
| Paranoia | 39.92 | 14.54 | 20 - 85 |

*Note.* Higher values on Mental State Probability indicate a higher probability of the false belief location. *SD* = Standard Deviation.

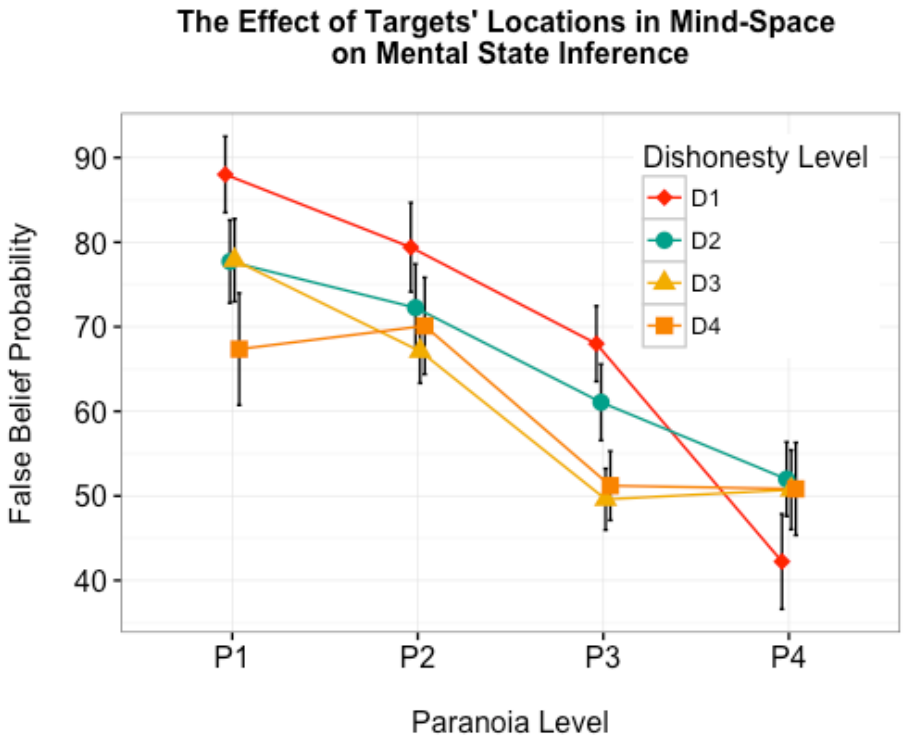


*Figure 2.* The effect of targets' locations in Mind-space on the probability of the mental state inferred. Note that higher values on the 'False Belief Probability' axis indicate higher probabilities of searching in the 'false belief' location, that is, where the Sally character left her object. Error bars show within-subject 95% confidence intervals around the means (Morey, 2008).

**Discussion**

The results of Experiment 3 are consistent with the idea that participants locate a target's mind within Mind-space before inferring the target's mental state, and that the location of the target mind within Mind-space is used to infer the probability of particular mental states. Specifically, the more paranoid that Sally was, and the more dishonest that Sally thought Anne was, the less likely participants were to predict that Sally would look in the location in which she left her object.

It is interesting to note that although the probability of ascribing a false belief to Sally decreased as paranoia and dishonesty increased, the probability ratings tended not to dip below 50%. This indicates that Sally was not likely to look in the false belief location, where she had left her object, but also not likely to look in the true belief location, where Anne had moved her object. This is most probably attributable to an aspect of the study design: although the stories mentioned only two locations as in the original task (Baron-Cohen et al., 1985), participants may have inferred that although Sally suspected her object had been moved, she did not know the exact location it had been moved to by Anne. Future studies may find increased true belief ratings by constraining the situational information further using pictorial stimuli rather than vignettes.

Although the task used was relatively simple, one can see large effects of changing the protagonists' position in Mind-space, and the position of the other character in the protagonist's Mind-space. Given that there is no objectively correct answer on this task, these results highlight the ambiguity in interpreting 'failures to represent the protagonist's false belief' in the standard version of the unseen change-of-location task without further interrogation of participants' reasoning. If the participant attributes paranoia/distrust to others in the absence of a cue to do so, they may respond in a manner which is typically interpreted as a failure to represent false belief (Happé & Frith, 1996).

While the results of Experiment 3 are consistent with one of the central tenets of the Mind-space theory - that the accuracy of mental state inference depends on the accuracy of characterising the target mind – Experiment 3 was not designed to show

that target minds are represented within a multi-dimensional space. Experiment 4 built upon the design of Experiment 3 in order to provide a more specific test of this aspect of the Mind-space theory. Accordingly, participants completed the same false belief vignettes task as used in Experiment 3 with a range of Sally characters. However, in Experiment 4, participants were given information about the Sally characters' scores on a range of traits (not including paranoia) which were selected on the basis of a validation study to covary with paranoia in the minds of a similar population to that which participants in Experiment 4 were drawn from. If participants represent minds within a multi-dimensional space in which covariances between dimensions are also represented, and use target locations within Mind-space to inform mental state inferences, then moving the Sally character on traits associated with paranoia should result in modified mental state inferences. Crucially, the size of the effect on mental state inference should vary for each participant as a function of the degree to which each trait is associated with paranoia within that participant's Mind-space.

## Experiment 4

### Methods

**Participants.** 55 participants (24 female) took part in an online task (built using the Gorilla Experiment Builder; Anwyl-Irvine, Massonnié, Flitton, Kirkham & Evershed, 2018) of approximately 20 minutes for monetary compensation. Participants were aged between 18 to 59 years old ($M$ = 31.35, $SD$ = 11.99), were residing in the UK, and reported English as their first language. Five participants were excluded prior to analysis after reporting mental health conditions in a screening questionnaire. The sample size for Study 4 was calculated *a priori* using simulations

(DeBruine & Barr, 2019; Brysbaert & Stevens, 2018) based on parameter estimates from Study 3. The results of these simulations indicate that with *N*=28 there is more than 80% power to detect an effect of magnitude similar to that observed in Experiment 3 with an alpha of .05. Twenty-eight was therefore set as the minimum sample size, but all participants volunteering to participate within the recruitment window were tested. The local Research Ethics Committee approved the study.

## Measures

### *Mental State Stories.*

The same 32 vignettes used in Experiment 3 were also used in this experiment.

### *Stimulus Validation Study*

A validation study using an analogous format to the Personality Pairs Task was devised in order to identify traits commonly associated with paranoia. In this study, 50 participants were asked to rate the association between 102 traits and paranoia using the same visual analogue scale as used in the Personality Pairs Task. The validation study was conducted online with participants resident in the UK who reported English as their first language. The results of this task were used to identify words which were commonly associated with paranoia (both negatively and positively) across participants (see Supplemental Materials Figure S2). Care was taken to ensure that the selected traits were not mere synonyms or antonyms of paranoia by cross-checking thesaurus entries (Thesaurus.com, Oxford English Thesaurus). In addition, words were excluded using *OpenMeaning (http://www.openmeaning.org/viz/)*, an online platform which allows for the

visualization of semantic spaces and provides a ranking of words of interest based on their semantic relatedness to a target word (in this case paranoia). None of the selected traits from the validation study appeared as one of the top 50 words semantically related to paranoia. Following this process, the final traits used in the experiment (known as 'source traits' hereafter) were *carefree*, *rational*, and *trusting*, which are negatively correlated with paranoia, and *superstitious*, *pessimistic* and *cautious,* which are positively correlated with paranoia.

***Paranoia Manipulation: Study 4***

As in Study 3, the 'Sally' characters were designed to vary across four levels of paranoia. However, in Study 4 paranoia was manipulated using the source traits which, on the basis of the validation study, were expected to result in Sally being placed at different positions along the paranoia dimension within Mind-space if covariation between traits is represented. Participants were told that the characters completed a questionnaire where they responded to a number of questions of the form: *"Please rate the degree to which you would describe yourself as:"* and then each of the six source traits was presented. Participants were told that the characters answered by choosing one of the following four options: *Not at All, A Little Bit, Somewhat,* and *Very Much*. At Paranoia Level 1, the Sally character responded '*Very Much*' to the three traits negatively correlated with paranoia, and '*Not at All'* to the three traits positively correlated with paranoia; at Level 2, the responses were '*A Little Bit'* to the positive traits and '*Somewhat'* to the negative traits; at Level 3, the responses were '*Somewhat'* to the positive traits and '*A Little Bit*'' to the negative traits; and at Level 4, the responses were '*Very Much'* to the positive traits and '*Not at All'* to the negative traits. These responses were designed to allow participants to infer

low paranoia at level 1 to high paranoia at level 4. Unlike Experiment 3, Study 4 did not include any dishonesty manipulation for the Anne character.

***Mental State Inference.*** Apart from the changes described above, the mental state inference task was the same as in Experiment 3.

***Explicit Paranoia and Association Ratings.*** After participants had completed all 32 mental state inference trials, they were shown each Sally character's questionnaire responses again and asked to report, using a four-point Likert scale (from 'not at all' to 'highly'): *"How paranoid do you think 'Sally' is?".* Following the paranoia ratings, participants were asked to estimate the association between paranoia and the six source traits used to manipulate Sally's paranoia using the same method as used in the Personality Pairs Task (see Table S7).

***Statistical Analyses.*** Statistical analysis was conducted using Linear Mixed Models implemented in the *lme4* package (Bates, Maechler, Bolker & Walker, 2014) in R. Experiment 4 is designed to test the predictions that: (1) participants locate minds within Mind-space based on information they are given about particular source traits; (2) they use that information to locate those minds on dimensions they believe to be correlated with the source traits; and (3) they use the location of minds within Mind-space to predict the probability of particular mental states. For these predictions to be supported, the data must show that each participant locates a particular Sally along the paranoia dimension according to the degree to which they believe the source traits are correlated with paranoia, and that this affects the mental states they attribute to that Sally character. Thus, a predicted relative paranoia score, for each participant

and each Sally, was derived by multiplying Sally's score on each source trait by the degree to which that participant thought that source trait was associated with paranoia (from the paranoia association ratings), and then summing across source traits. This final *Mean Predicted Relative Paranoia* (MPRP) score represents where the participant would locate each Sally on the paranoia dimension if Sally's scores on the source traits cause the participant to locate Sally on the paranoia dimension at a location in accordance with the participant's estimated correlation between the source traits and paranoia.

MPRP was included as a fixed effect to predict the False Belief Probability while controlling for trial and participant random intercepts (False Belief Probability ~ MPRP + (1 | trial) + (1 | participant). It was hypothesised that the higher the MPRP (i.e. the more paranoid Sally was thought to be), the less likely it would be for participants to attribute a false belief to Sally's character.

**Results**

Descriptive statistics for the estimated probability of the 'false belief' location as a function of Sally's scores on the source traits are presented in Table S6. As predicted, the model results show a significant effect of MPRP on the False Belief Probability attribution ($\beta$ = - 8.85, *95% CI* [-10.65, -7.03], $p < .001$, see Figure 3 and Table S8). Crucially, a model comparison including the MPRP model, a model with the Sally source traits (unweighted by their correlation with paranoia) as a fixed effect, and a null model, with all models carrying the same random effects structure, was also performed. The results indicated the MPRP model was significantly better than the null and the unweighted Sally source trait models ($\chi^2_{(1)} = 83.4$, $p < .001$, see

Table S9). Examination of the AIC and BIC values also showed that the MPRP model outperformed the Sally source traits model ($\Delta$AIC = 32, $\Delta$BIC = 42, where differences of 6 are generally considered to be non-negligible (Burnham & Anderson, 1998)). Thus, results suggest that participants (1) use their estimate of the correlation between the source traits and paranoia to estimate Sally's location on the paranoia dimension, and (2) use this information to inform their estimates of the probability of Sally's mental states.

As a manipulation check, we computed a slope that represents the change in explicit paranoia ratings across levels of Sally's scores on the source traits. This was achieved by calculating, for each participant, the mean explicit paranoia rating, and then mean-correcting each rating. Linear weights were then assigned for each level of Sally source traits and the weighted sum of the explicit paranoia ratings computed (all values for these computations are provided in the data file for this study in the OSF archive https://doi.org/10.17605/OSF.IO/4K9HS). These slope values represent the degree to which changing scores on the source traits (across Sally characters) produces changes in explicit paranoia ratings for each participant. When tested against zero using a one-sample t-test, the slopes were found to be significantly different from zero (indicating that changing the Sally character's scores on the source traits caused explicit paranoia ratings to change; $M = 8.27$, *95% CI* [7.52 – 9.01], $t(48) = 22.22$, $p < .001$. The same procedure was repeated on the MPRP data to derive slope values that reflect the degree to which paranoia ratings would change as a function of changing scores for the Sally character on the source traits, if participants based the paranoia ratings on their estimated correlations between source traits and paranoia. As expected, we found a significant positive correlation between

the explicit paranoia judgement slopes and the MPRP slopes ($r_{(47)} = .40$, $p = .005$). Thus, the degree to which participants adapted their explicit paranoia judgements as a function of Sally's scores on the source traits, corresponded with the MPRP calculated on the basis of participants' judgements of the correlation between the source traits and paranoia.

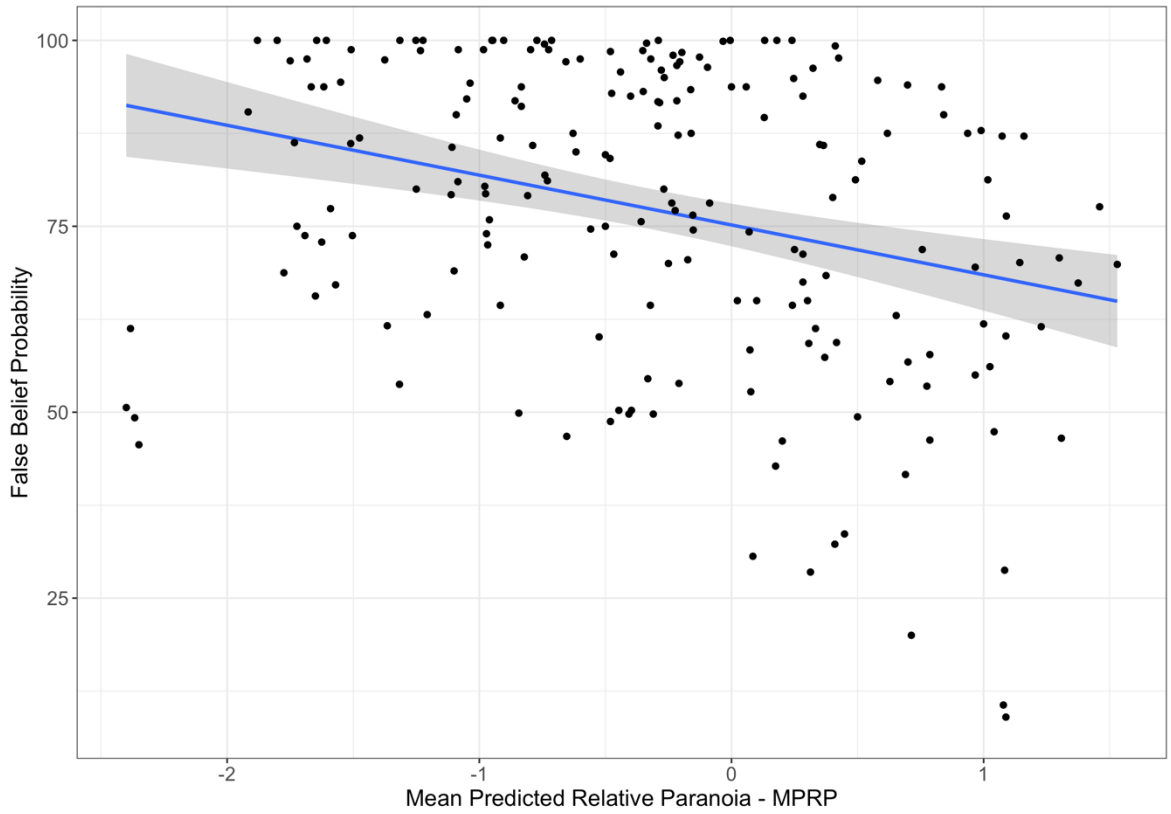


*Figure 3.* Effect of *Mean Predicted Relative Paranoia* (MPRP) score on 'False Belief' Attribution. Shaded area represents the 95% confidence interval. MPRP is calculated on the basis of the Sally character's scores on various traits and the degree to which each participant believes those traits to be associated with paranoia.

**Discussion**

Experiment 4 demonstrates that when provided with information about a target's mind that allows it to be located on a number of source dimensions, participants use that information to extrapolate the location of the target mind on

dimensions they believe covary with the source dimensions, and they do so in a manner which reflects the degree of estimated covariation. Furthermore, they use the estimated location on the new dimensions to make inferences about the target's mental states where relevant. This pattern of data is consistent with predictions from the Mind-space theory, and also with previous demonstrations that, for example, individuals are thought to have different mental states depending on their locations on dimensions of warmth and competence (Fiske et al., 2002).

## General Discussion

We sought to understand individual differences in theory of mind by testing a theory in which other minds are represented in a multidimensional space. Within this framework the position of a target mind within Mind-space is combined with information about the situation the target is in, in order to infer the probability of the target having particular mental states. Accordingly, individual differences in the accuracy of mental state inferences may be explained by factors including the accuracy of an individual's Mind-space (i.e. the degree to which their Mind-space accurately captures variance in other minds), and the ability to locate a target mind accurately within Mind-space. Experiment 1 demonstrated that variance in ToM ability (i.e. the accuracy of mental state inference) was associated with how accurately the covariance between personality dimensions was represented within Mind-space. Experiment 2 showed that the accuracy of Mind-space was associated with the ability to locate another person within Mind-space, on dimensions relating to personality traits and intelligence, based on a minimal sample of their behaviour. The results obtained in Experiment 3 support the prediction that the location of a target mind in Mind-space affects the probability of particular mental states being attributed

to that target given the situation they are in. Experiment 4 extended this result to show the dimensional nature of mind representation. Participants extrapolated from the location of a target's mind on source dimensions to estimate the target's location on novel dimensions of mind, and used this estimate to infer the probability of mental states.

The results of Experiment 1 demonstrate a relationship between understanding the structure of personality in the general population and the ability to make accurate mental state inferences about particular characters. In designing the MASC task, the authors ensured that each character had distinctive traits (e.g. outgoing vs. shy; Dziobek et al., 2006). Implicit in this task is the relationship between the characters' traits and the kind of mental states they generate, yet how traits and mental states relate to one another has rarely been addressed, particularly in adulthood. It should be acknowledged, however, that several trait theories of mind (person) representation exist, and some of these theories specify that traits may be associated with differential probabilities of particular mental states being inferred (for example the work on stereotyping by Fiske et al., 2002; for a full discussion of such theories and their relationship to Mind-space see Conway et al., 2019, 'Relationship to existing theories', p.805). Of particular relevance is the work of Tamir and Thornton (Tamir & Thornton, 2018; Tamir, Thornton, Contreras, & Mitchell, 2016), who argue that traits are represented in a 3-dimensional space, and that traits can be used to infer the probability of types of mental states (e.g. beliefs vs desires) and states of mind (e.g. fatigued vs invigorated), which can also be represented in a 3-dimensional space. Neuroimaging work has identified where in the brain traits and mental states may be represented: activation in the temporo-parietal junction tends to occur when

representing others' thoughts or beliefs when they differ from one's own (Saxe & Powell, 2006; Koster-Hale, Richardson, Velez, Asaba, Young & Saxe, 2017), whereas activation in the medial prefrontal cortex is thought to reflect representations of specific people and their enduring social traits (Hassabis et al., 2014; Mitchell, Cloutier, Banaji, & Macrae, 2006; Tamir et al, 2016; although see Cook, 2014). However, the demonstration that there is brain activation specific to mental states vs. traits does not provide a psychological account of how such information is used. The Mind-space framework attempts to provide a model to link representation of a particular mind and its qualities to inference of the mental states that this mind holds. The findings of Experiment 1 support the idea that the quality of mind representation may be a determinant of individual differences in theory of mind.

The results of Experiments 3 and 4 support the contention that mental state inference is a process in which the probability of a particular mental state in a given individual is inferred based on the learned probability of observing that mental state given the context and the individual's position in Mind-space (see Figure 1). Accordingly, in addition to the factors studied in the current experiments, the accuracy of mental state inferences is likely to be a product of two further factors: the accuracy with which position in Mind-space is mapped to the probability of particular mental states given a specific situation; and one's propensity to consider the position of the target mind in Mind-space before making a judgement as to the target's mental state. The finding that a less accurate Mind-space was associated with a lack of mental state inference (Experiment 1) may be especially relevant to this last factor. We speculated that an association between the accuracy of Mind-space and the ability to locate a target mind within Mind-space may be due to common effects of social

motivation, social attention, or social learning (Conway et al., 2019). Decreased social motivation in particular may explain why an individual may form inaccurate models of how minds vary, have a worse ability to locate minds within Mind-space, and be less likely to make mental state attributions.

With respect to the finding that the accuracy of Mind-space predicts the ability to locate others within Mind-space (Experiment 2), it is important to note that participants were not highly accurate in their estimates. This inaccuracy is likely attributable to the minimal exposure to the targets in the thin-slice videos. Predictive accuracy has been shown to improve when thin-slices are extended for some traits, for instance Carney, Colvin and Hall (2007) found good accuracy for judgements of extraversion, conscientiousness, and intelligence after 5 seconds, whereas longer exposure was required for neuroticism, openness to experience and agreeableness. Whether the accuracy of an individual's Mind-space predicts their ability to locate an individual within Mind-space after longer exposure, or predicts their ability to increase the accuracy with which they locate an individual after increased exposure, remains to be determined. It should also be acknowledged that these results may hold for only a small portion of Mind-space relating to personality. Personality represented a good initial test of the Mind-space theory as there is a wealth of data available on personality trait covariance, meaning that the accuracy of an individual's model of personality covariance can be established. However, whether these results would also be found for other aspects of Mind-space with little or no relation to personality (e.g. the factor structure of intelligence), also remains to be seen.

One possibility suggested by these data is that individuals may not have a unitary theory of mind ability, but rather that accuracy in the inference of mental states, and in locating another mind within Mind-space, may depend upon the particular mind to be modelled and its relationship to the kinds of minds one has previously encountered which have shaped one's Mind-space. This is supported by the finding that greater similarity between participants and targets resulted in more accurate trait judgements (Experiment 2), and that individuals use trait judgments when inferring mental states (Experiments 3 and 4). Therefore, individuals who are more typical of the population being represented (i.e. have average trait scores themselves) are more likely to make accurate inferences about the minds and mental states of others; both on average across inferences made for specific targets in the population, and for targets about whom nothing is known where the optimal strategy is to attribute average trait values to them.

Intriguingly, previous research on implicit personality theory indirectly supports the contention that those who have typical trait covariances across a number of dimensions make more accurate mental state inferences, but only if one accepts as true the hypothesis that the accuracy of mental state inference depends upon the accuracy of mind representation. Specifically, it has been demonstrated that an individual's model of personality is partly built upon their view of their own personality: if they have a causal model explaining the patterning of traits in their own personality (e.g. I am optimistic because I am intelligent and have always succeeded) they are likely to assume the same patterning of traits in the general population (i.e. that optimism is typically associated with intelligence; Critcher, & Dunning, 2009; Critcher, Rom, & Dunning, 2015). Individuals with trait covariance typical of the

population would therefore have a more accurate Mind-space if they based their population model on their own personality; and if accuracy of mind representation determines accuracy of mental state inference, they would make more accurate mental state inferences as a result.

The idea that one's theory of mind ability may depend on the target mind to be represented has interesting implications for atypical groups. Neurotypical participants may perform well on existing theory of mind tasks in which the 'correct' answers are derived by neurotypical consensus (e.g. Dziobek et al., 2006), as their own mind is similar to the average. Conversely, neurotypical participants may also have minds that are particularly easy to represent by the majority of the population. In contrast, those who have atypical minds may find it harder to represent the minds of neurotypical individuals, and in turn, be harder for neurotypical individuals to represent (Edey, Cook, Brewer, Johnson, Bird, & Press, 2016; Brewer et al., 2016). The same loss of accuracy is likely to occur when we need to represent the minds and mental states of out-groups (Sasson, Faso, Nugent, Lovell, Kennedy, & Grossman, 2017; Bruneau & Saxe, 2012).

Related suggestions have been made previously; for instance, Happé and Frith (1996) suggested that children diagnosed with Conduct Disorder may have a 'theory of nasty minds', that may be adaptive to aversive developmental environments and an accurate reflection, based on their prior experience, of how others think and behave. In their study of mental state inference, children with Conduct Disorder performed less well than typically developing children but better than those with Autism Spectrum Disorder, and showed a particular ability for mental state inference in

antisocial situations, such as bullying. Therefore, even in the absence of explicit information about others' traits, children with Conduct Disorder may ascribe more negative mental states than the typical population due to inaccurately locating others in Mind-space, and/or atypical mappings between locations in Mind-space and mental states.

In sum, these studies try to account for variance in the ability of humans to infer accurately the mental states of others. The empirical support for Mind-space presented here highlights the importance of modelling minds when considering individual differences in the representation and inference of others' mental states, personality, and intelligence.

## Author Contributions

J.R. Conway, C. Catmur, and G. Bird developed the study concept and design. Data collection and stimulus development was performed by J.R.C., H.C. Cuve, S. Koletsi, N. Bronitt, with additional assistance from M. Fagundez, and K. Overall. J.R.C. and H.C.C. (Expt. 4) performed the data analysis and interpretation under the supervision of M.P. Coll, C.C., and G.B. J.R.C. drafted the manuscript, J.R.C, H.C.C., M.P.C., C.C., and G.B. provided critical revisions, and all authors approved the final version of the manuscript for submission.

## Acknowledgments

This work was supported by an Economic and Social Research Council studentship [Ref: 1413340] awarded to J.R. Conway. J.R. Conway acknowledges IAST funding from the French National Research Agency (ANR) under the Investments for the Future (Investissements d'Avenir) program, grant ANR-17-EURE-0010.

## Context

The current paper is the first empirical test of a new theoretical framework advanced by the authors (Conway, Catmur, & Bird, 2019) that aims to explain individual differences in the accuracy of mental state inferences ('mentalizing' or 'theory of mind'). This paper reports four studies testing the predictions of a new mechanistic model of mentalizing – the 'Mind-space' model – which suggests that minds are represented within a multidimensional space, much as faces are thought to be represented within Face-space. This model recognizes that mental states are a product of, and dependent upon, the specific mind that gives rise to them. Under this model,

therefore, individual differences in mentalizing ability can be explained by individual differences in the ability to represent variance in minds, and in the ability to determine the characteristics of another's mind when attempting to infer their mental states. The Mind-space model presents a framework to understand variance in mentalizing ability, which has implications for the study of this ability in clinical groups (most notably Autism Spectrum Disorder), across childhood development, and its implementation in artificial agents.

**References**

Aboud, F. E. (1988). Children and prejudice. B. Blackwell.

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2019). Gorilla in our Midst: An online behavioral experiment builder. Behavior research methods, 1-20.

Ashton, M. C., & Lee, K. (2007). Empirical, Theoretical, and Practical Advantages of the HEXACO Model of Personality Structure. *Personality and Social Psychology Review*, *11*(2), 150–166. https://doi.org/10.1177/1088868306294907

Ashton, M. C., & Lee, K. (2009). The HEXACO-60: A short measure of the major dimensions of personality. *Journal of Personality Assessment, 91*(4), 340–345. https://doi.org/10.1080/00223890902935878

Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants. *Trends in Cognitive Sciences, 14*(3), 110–118. https://doi.org/10.1016/j.tics.2009.12.006

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R.H.B., Singmann, H., …Green, P. (2018). *Linear Mixed-Effects Models using 'Eigen' and S4.* Retrieved from https://github.com/lme4/lme4/ http://lme4.r-forge.r-project.org/

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a 'theory of mind'? *Cognition, 21*(21), 37–46. https://doi.org/https://doi.org/10.1016/0010-0277(85)90022-8

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(5), 407-418.

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The

Autism-spectrum Quotient (AQ): Evidence from Asperger Syndrome/High-Functioning Autism, Males and Females, Scientists and Mathematicians. *Journal of Autism and Developmental Disorders, 31*(1), 5-17.

Borkenau, P., Mauer, N., Riemann, R., Spinath, F. M., & Angleitner, A. (2004). Thin Slices of Behavior as Cues of Personality and Intelligence. *Journal of Personality and Social Psychology, 86*(4), 599–614. https://doi.org/10.1037/0022-3514.86.4.599

Brewer, R., Biotti, F., Catmur, C., Press, C., Happe, F., Cook, R., & Bird, G. (2016). Can Neurotypical Individuals Read Autistic Facial Expressions? Atypical Production of Emotional Facial Expressions in Autism Spectrum Disorders. *Autism Research, 9*(2), 262–271. https://doi.org/10.1002/aur.1508

Bruneau, E. G., & Saxe, R. (2012). The power of being heard: The benefits of 'perspective-giving' in the context of intergroup conflict. *Journal of Experimental Social Psychology, 48*(4), 855–866. https://doi.org/10.1016/j.jesp.2012.02.017

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. Journal of Cognition, 1(1).

Burnett, S., Bird, G., Moll, J., Frith, C., & Blakemore, S. J. (2009). Development during adolescence of the neural processing of social emotion. Journal of cognitive neuroscience, 21(9), 1736-1750.

Burnham, K. P., & Anderson, D. R. (1998). Practical use of the information-theoretic approach. In *Model Selection and Inference* (pp. 75-117). Springer, New York, NY.

Canty, A., & Ripley, B. (2017). *Bootstrap Functions.* Retrieved from https://cran.r-project.org/web/packages/boot/boot.pdf

Carney, D. R., Colvin, C. R., & Hall, J. A. (2007). A thin slice perspective on the accuracy of first impressions. *Journal of Research in Personality, 41*(5), 1054–1072. https://doi.org/10.1016/j.jrp.2007.01.004

Champely, S., Ekstrom, C., Dalgaard, P., Gill, J., Weibelzahl, S., Anandkumar, A., … Volcic, R., De Rosario, H. (2018). *Basic Functions for Power Analysis.* Retrieved from https://github.com/heliosdrm/pwr

Conway, J.R., Catmur, C., & Bird, G. (2019). Understanding Individual Differences in Theory of Mind via Representation of Minds, Not Mental States. *Psychonomic Bulletin and Review, https://doi.org/10.3758/s13423-018-1559-x*

Cook, J. L. (2014). Task-relevance dependent gradients in medial prefrontal and temporoparietal cortices suggest solutions to paradoxes concerning self/other control. *Neuroscience and Biobehavioral Reviews*, 42, 298–302. https://doi.org/10.1016/j.neubiorev.2014.02.007

Critcher, C. R., & Dunning, D. (2009). Egocentric pattern projection: How implicit personality theories recapitulate the geography of the self. Journal of personality and social psychology, 97(1), 1.

Critcher, C. R., Dunning, D., & Rom, S. C. (2015). Causal trait theories: A new form of person knowledge that explains egocentric pattern projection. Journal of personality and social psychology, 108(3), 400.

Davis, M. H. (1983). Measuring Individual Differences in Empathy: Evidence for a Multidimensional Approach. *Journal of Personality and Social Psychology, 44*(1), 113–126. https://doi.org/10.1037/0022-3514.44.1.113

DeBruine, L. M., & Barr, D. J. (2019, June 2). Understanding mixed effects models through data simulation. https://doi.org/10.31234/osf.io/xp5cy

Dennett, D. C. (1978). Beliefs about beliefs. *Behavioral and Brain Sciences, 4,* 568–570.

Devine, R. T., & Hughes, C. (2014). Relations between false belief understanding and executive function in early childhood: A meta-analysis. *Child Development, 85*(5), 1777–1794. https://doi.org/10.1111/cdev.12237

Dziobek, I., Fleck, S., Kalbe, E., Rogers, K., Hassenstab, J., Brand, M., … Convit, A. (2006). Introducing MASC: A movie for the assessment of social cognition. *Journal of Autism and Developmental Disorders, 36*(5), 623–636. https://doi.org/10.1007/s10803-006-0107-0

Edey, R., Cook, J., Brewer, R., Johnson, M. H., Bird, G., & Press, C. (2016). Interaction takes two: Typical adults exhibit mind-blindness towards those with autism spectrum disorder. *Journal of Abnormal Psychology, 125*(7), 879–885. https://doi.org/10.1037/abn0000199

Epley, N., Keysar, B., Van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of Personality and Social Psychology, 87*(3), 327–339. https://doi.org/10.1037/0022-3514.87.3.327

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior research methods, 41*(4), 1149-1160.

Fenigstein, A., & Vanable, P. A. (1992). Paranoia and Self-Consciousness. *Journal of Personality and Social Psychology, 62*(1), 129–138. https://doi.org/10.1037//0022-3514.62.1.129

Fiske, S. T., Cuddy, A. J., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. Journal of personality and social psychology, 82(6), 878.

Gallagher, H. L., & Frith, C. D. (2003). Functional imaging of "theory of mind." *Trends in Cognitive Sciences, 7*(2), 77–83. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12584026

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research, 42*(3), 237-288.

Goldberg, L. R. (1990). An Alternative 'Description of Personality': The Big Five Factor Structure. *Journal of Psychology and Social Psychology, 59*(6), 1216–1229. https://doi.org/10.1037//0022-3514.59.6.1216

Happé, F. G. (1994). An advanced test of theory of mind: understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of Autism and Developmental Disorders, 24*(2), 129–154. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8040158

Happé, F. G., & Frith, U. (1996). Theory of mind and social impairment in children with conduct disorder. *British Journal of Developmental Psychology, 14*, 385–398.

Hassabis, D., Spreng, R. N., Rusu, A. A., Robbins, C. A., Mar, R. A., & Schacter, D. L. (2014). Imagine all the people: How the brain creates and uses personality models to predict behavior. *Cerebral Cortex, 24*(8), 1979–1987.

Heyes, C. (2014). Submentalizing: I am not really reading your mind. Perspectives on Psychological Science, 9(2), 131-143.

Heyes, C. (2015). Animal mindreading: what's the problem? *Psychonomic Bulletin & Review, 22*(2), 313-327.

Heyes, C. (2017). Apes Submentalise. *Trends in Cognitive Sciences, 21*(1), 1–2.

Koster-Hale, J., Richardson, H., Velez, N., Asaba, M., Young, L., & Saxe, R. (2017) Mentalizing regions represent distributed, continuous, and abstract dimensions of others' beliefs. Neuroimage, 161, 9-18.

Kulke, L., Reiß, M., Krist, H., & Rakoczy, H. (2017). How robust are anticipatory looking measures of Theory of Mind? Replication attempts across the life span. *Cognitive Development*, (August), 0–1. https://doi.org/10.1016/j.cogdev.2017.09.001

Lalonde, C. E., & Chandler, M. J. (2002). Children's understanding of interpretation. New Ideas in Psychology, 20(2-3), 163-198.

Lee, K., & Ashton, M. C. (2016). Psychometric Properties of the HEXACO-100. *Assessment, 25*(5), 543-556. https://doi.org/10.1177/1073191116659134

Lix, L. M., Keselman, J. C., & Keselman, H. J. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test. *Review of Educational Research, 66*(4), 579-619.

Mair, P., & Wilcox, R. (2018). *WRS2: A Collection of Robust Statistical Methods.* Retrieved from: https://r-forge.r-project.org/projects/psychor/

Milligan, K., Astington, J. W., & Dack, L. A. (2014). Language and Theory of Mind : Meta-Analysis of the Relation Between Language Ability and False-belief Understanding. *Child Development, 78*(2), 622–646. https://doi.org/10.1111/j.1467-8624.2007.01018.x

Mitchell, J. P., Cloutier, J., Banaji, M. R., & Macrae, C. N. (2006). Medial prefrontal dissociations during processing of trait diagnostic and nondiagnostic person

information. *Social Cognitive and Affective Neuroscience, 1*(1), 49–55. https://doi.org/10.1093/scan/nsl007

Morey, R. D. (2008). Confidence Intervals from Normalized Data; A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology, 4*(2), 61-64.

Nagel, R. (1995). Unraveling in guessing games: An experimental study. The American Economic Review, 85(5), 1313-1326.

Osterhaus, C., Koerber, S., & Sodian, B. (2016). Scaling of advanced theory-of-mind tasks. *Child development*, *87*(6), 1971-1991.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences, 4,* 515–526.

Rabinowitz, N., Perbet, F., Song, H.F., Zhang, C., Eslami, S.M.A., & Botvinick, M. (2018). *Machine Theory of Mind.* Retrieved from arXiv:1802.07740v2

Repacholi, B. & Slaughter, V. (Eds.) (2003). Individual differences in theory of mind. Implications for typical and atypical development. New York: Psychology Press.

Ruffman, T. (1996). Do children understand the mind by means of simulation or a theory? Evidence from their understanding of inference. Mind & Language, 11(4), 388-414.

Sabbagh, M. A, Xu, F., Carlson, S. M., Moses, L. J., & Lee, K. (2006). The Development of Executive Functioning and Theory of Mind. *Psychological Science, 17*(1), 74–81. https://doi.org/10.1111/j.1467-9280.2005.01667.x

Santiesteban, I., Banissy, M. J., Catmur, C., & Bird, G. (2015). Functional Lateralization of Temporoparietal Junction: Imitation Inhibition, Visual Perspective Taking and Theory of Mind. *European Journal of Neuroscience, 42*(8), 2527-2533. https://doi.org/10.1093/biostatistics/manuscript-acf-v5

Sasson, N. J., Faso, D. J., Nugent, J., Lovell, S., Kennedy, D. P., & Grossman, R. B. (2017). Neurotypical Peers are Less Willing to Interact with Those with Autism based on Thin Slice Judgments. *Scientific Reports, 7*(October 2016), 1–10. https://doi.org/10.1038/srep40700

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

Tamir, D. I., & Thornton, M. A. (2018). Modeling the predictive social mind. Trends in Cognitive Sciences, 22(3), 201–212.

Tamir, D. I., Thornton, M. A., Contreras, J. M., & Mitchell, J. P. (2016). Neural evidence that three dimensions organize mental state representation: Rationality, social impact, and valence. Proceedings of the National Academy of Sciences, 113(1), 194–199.

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. The Quarterly Journal of Experimental Psychology Section A, 43(2), 161-204.

Valentine, T., Lewis, M. B., & Hills, P. J. (2016). Face-space: A unifying concept in face recognition research. The Quarterly Journal of Experimental Psychology, 69(10), 1996-2019.

Wechsler, D. (2011). *Wechsler Abbreviated Scale of Intelligence - Second Edition.* San Antonio, TX: NCS Pearson.

Wellman, H. M., & Liu, D. (2004). Scaling of theory of mind tasks. Child development, 75(2), 523-541.

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.